

IEEE Signal Processing Magazine

Volume 34 | Number 6 | November 2017

DEEP LEARNING FOR VISUAL UNDERSTANDING

Recent Advances:
Part 1

Performance Bounds for Parameter Estimation under Misspecified Models

The Future of Signal Processing

AI in Financial Markets

Mystery Curve

```

elif _operation == "MIRROR_Y":
    mirror_mod.use_x = False
    mirror_mod.use_y = True
    mirror_mod.use_z = False
elif _operation == "MIRROR_Z":
    mirror_mod.use_x = False
    mirror_mod.use_y = False
    mirror_mod.use_z = True

#selection at the end -add b
mirror_ob.select= 1
modifier_ob.select=1
bpy.context.scene.objects.active =
print("Selected" + str(modifier_ob)
    #mirror_ob.select = 1
time = bpy.context.scene.frame_current
bpy.data.objects[mirror_ob.name].

```

IEEE Signal Processing Society





ICIP 2018

<http://2018.ieeeicip.org>

IEEE INTERNATIONAL CONFERENCE ON IMAGE PROCESSING OCTOBER 7 - 10, 2018 * ATHENS, GREECE

Organizing Committee

General Chairs

Christophoros Nikou, University of Ioannina, Greece

Kostas Plataniotis, University of Toronto, Canada

Technical Program Chairs

Nikolaos Boulgouris, Brunel University London, UK

Lisimachos P. Kondi, University of Ioannina, Greece

Finance Chair

Aggelos Pikrakis, University of Piraeus, Greece

Plenary Chairs

John Apostolopoulos, Cisco Systems, USA

Athanassios Skodras, University of Patras, Greece

Tutorial Chairs

Christine Guillemot, INRIA, Rennes, France

Rafael Molina, University of Granada, Spain

Special Sessions Chairs

Guy Côté, Apple Inc., USA

Adriana Dumitras, Microsoft, USA

Awards Chair

Jean-Philippe Thiran, EPFL, Switzerland

Exhibits/Demo Chair

Adrian Bors, University of York, UK

Publication Chair

Jean-Luc Dugelay, EURECOM, France

Industry Liaison Chairs

Panos Nasiopoulos, University of British Columbia, Canada

Amir Said, Qualcomm, USA

Local Arrangement Chair

Stefanos Kollias, National Technical University of Athens, Greece

Publicity Chairs

Nikos Nikolaidis, Aristotle University of Thessaloniki, Greece

Konstantinos Papadis, Athens Information Technology, Greece

Students/Young Professionals Activities and Doctoral Consortium Chairs

Patrizio Campisi, Università degli Studi Roma Tre, Italy

Sotiris Tsaftaris, University of Edinburgh, UK

Nikolaos Thomos, University of Essex, UK

Registration Chair

Kostas Berberidis, University of Patras, Greece

IEEE Student Activities Chair

Kostas Karpouzis, National Technical University of Athens, Greece

Innovation Program Chairs

Jill Boyce, Intel Corporation

Haohong Wang, TCL Research America

International Liaisons

Panos Papamichalis, Southern Methodist University, USA

Xiao Ping Zhang, Ryerson University, Canada

Yong Man Ro, KAIST, Republic of Korea

Advisory Board

Aggelos Katsaggelos

Petros Maragos

Ioannis Pitas

Sergios Theodoridis



The 25th IEEE International Conference on Image Processing (ICIP) will be held in the Megaron Athens International Conference Centre, Athens, Greece, on October 7-10, 2018. ICIP is the world's largest and most comprehensive technical conference focused on image and video processing and computer vision. The conference will feature world-class speakers, tutorials, exhibits, and a vision technology showcase.

Topics of interest include, but are not limited to:

- Filtering, Transforms, Multi-Resolution Processing
- Restoration, Enhancement, Super-Resolution
- Computer Vision Algorithms and Technologies
- Compression, Transmission, Storage, Retrieval
- Multi-View, Stereoscopic, and 3D Processing
- Multi-Temporal and Spatio-Temporal Processing
- Biometrics, Forensics, and Content Protection
- Biological and Perceptual-based Processing
- Medical Image and Video Analysis
- Document and Synthetic Visual Processing
- Color and Multispectral Processing
- Scanning, Display, and Printing
- Applications to various fields
- Computational Imaging
- Video Processing and Analytics
- Visual Quality Assessment
- Deep learning for Images and Video
- Image and Video Analysis for the Web
- Image Processing for VR Systems
- Image Processing for Autonomous Vehicles

Paper Submission

Authors are invited to submit papers of not more than four pages for technical content including figures and references, with one optional page containing only references. Submission instructions, templates for the required paper format, and information on "no show" policy are available at 2018.ieeeicip.org.

Journal Paper Presentations

Authors of papers published in all IEEE Signal Processing Society fully owned journals as well as in IEEE Wireless Communication Letters will be given the opportunity to present their work at ICIP 2018, subject to space availability and approval by the Technical Program Chairs of IEEE ICIP 2018.

Innovation Program

Following the tradition that started in 2016, the ICIP 2018 Innovation Program Chairs will arrange an outstanding event with prominent speakers from the Industry.

Tutorials, Special Sessions, and Challenge Sessions Proposals

Tutorials will be held on October 7, 2018. Tutorial proposals must include title, outline, contact information, biography and selected publications for the presenter(s), and a description of the tutorial and material to be distributed to participants. For detailed submission guidelines, please refer to the tutorial proposals page. Special Sessions and Challenge Session Proposals must include a topical title, rationale, session outline, contact information, and a list of invited papers/participants. For detailed submission guidelines, please refer the ICIP 2018 website at 2018.ieeeicip.org.

Important Dates

Special Session Proposals:	November 15, 2017
Notification of Special Session Acceptance:	December 15, 2017
Tutorial Proposals:	December 15, 2017
Notification of Tutorial Acceptance:	January 15, 2018
Paper Submission:	February 7, 2018
Notification of Acceptance:	April 30, 2018
Camera-Ready Papers:	May 31, 2018

IEEE
Signal Processing Society

Contents

Volume 34 | Number 6 | November 2017

SPECIAL SECTION

DEEP LEARNING FOR VISUAL UNDERSTANDING: PART 1

- 24 FROM THE GUEST EDITORS**
Fatih Porikli, Shiguang Shan, Cees Snoek, Rahul Sukthankar, and Xiaogang Wang
- 26 DEEP REINFORCEMENT LEARNING**
Kai Arulkumaran, Marc Peter Deisenroth, Miles Brundage, and Anil Anthony Bharath
- 39 WEAKLY SUPERVISED LEARNING WITH DEEP CONVOLUTIONAL NEURAL NETWORKS FOR SEMANTIC SEGMENTATION**
Seunghoon Hong, Suha Kwak, and Bohyung Han
- 50 THE ROBUSTNESS OF DEEP NETWORKS**
Alhussein Fawzi, Seyed-Mohsen Moosavi-Dezfooli, and Pascal Frossard

©ISTOCKPHOTO.COM/MIKAEVY

PG. 142



ON THE COVER

This special issue of *IEEE Signal Processing Magazine* provides survey articles on the latest advances in deep learning for visual understanding. Its objective is to encourage a diverse audience of researchers and enthusiasts toward an effective participation in the solution of analogous problems in other signal processing fields by inseminating similar ideas.

COVER IMAGE: ©ISTOCKPHOTO.COM/MONSIU

- 63 VISUAL QUESTION ANSWERING**
Damien Teney, Qi Wu, and Anton van den Hengel
- 76 DEEP METRIC LEARNING FOR VISUAL UNDERSTANDING**
Jiwen Lu, Junlin Hu, and Jie Zhou
- 85 CONVOLUTIONAL NEURAL NETWORKS FOR INVERSE PROBLEMS IN IMAGING**
Michael T. McCann, Kyong Hwan Jin, and Michael Unser
- 96 DEEP MULTIMODAL LEARNING**
Dhanesh Ramachandram and Graham W. Taylor

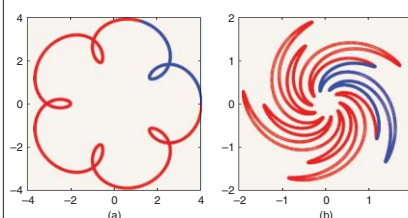
- 109 DEEP LEARNING FOR IMAGE-TO-TEXT GENERATION**
Xiaodong He and Li Deng

- 117 DEEP-LEARNING SYSTEMS FOR DOMAIN ADAPTATION IN COMPUTER VISION**
Hemant Venkateswara, Shayok Chakraborty, and Sethuraman Panchanathan

- 130 DEEP CONVOLUTIONAL NEURAL MODELS FOR PICTURE-QUALITY PREDICTION**
Jongyoo Kim, Hui Zeng, Deepti Ghadiyaram, Sanghoon Lee, Lei Zhang, and Alan C. Bovik

FEATURE

- 142 PERFORMANCE BOUNDS FOR PARAMETER ESTIMATION UNDER MISSPECIFIED MODELS**
Stefano Fortunati, Fulvio Gini, Maria S. Greco, and Christ D. Richmond



PG. 158

IEEE SIGNAL PROCESSING MAGAZINE (ISSN 1053-5888) (ISPREG) is published bimonthly by the Institute of Electrical and Electronics Engineers, Inc., 3 Park Avenue, 17th Floor, New York, NY 10016-5997 USA (+1 212 419 7900). Responsibility for the contents rests upon the authors and not the IEEE, the Society, or its members. Annual member subscriptions included in Society fee. Nonmember subscriptions available upon request. **Individual copies:** IEEE Members US\$20.00 (first copy only), nonmembers US\$241.00 per copy. Copyright and Reprint Permissions: Abstracting is permitted with credit to the source. Libraries are permitted to photocopy beyond the limits of U.S. Copyright Law for private use of patrons: 1) those post-1977 articles that carry a code at the bottom of the first page, provided the per-copy fee indicated in the code is paid through the Copyright Clearance Center, 222 Rosewood Drive, Danvers, MA 01923 USA; 2) pre-1978 articles without fee. Instructors are permitted to photocopy isolated articles for noncommercial classroom use without fee. **For all other copying, reprint, or republication permission,** write to IEEE Service Center, 445 Hoes Lane, Piscataway, NJ 08854 USA. Copyright © 2017 by the Institute of Electrical and Electronics Engineers, Inc. All rights reserved. Periodicals postage paid at New York, NY, and at additional mailing offices. **Postmaster:** Send address changes to IEEE Signal Processing Magazine, IEEE, 445 Hoes Lane, Piscataway, NJ 08854 USA. Canadian GST #125634188 **Printed in the U.S.A.**

Digital Object Identifier 10.1109/MSP.2017.2750304

COLUMNS

- 8 Panel and Forum**
Challenges and Open Problems in Signal Processing: Panel Discussion Summary from ICASSP 2017
Yonina C. Eldar, Alfred O. Hero III, Li Deng, Jeff Fessler, Jelena Kovačević, H. Vincent Poor, and Steve Young
- 14 Community Voices**
What Is the Future of Signal Processing?
Andres Kwasinski and Min Wu
- 17 Special Reports**
Medical Optical Imaging
John Edwards
- 21 Reader's Choice**
Top Downloads in IEEE Xplore
- 158 Lecture Notes**
The Mystery Curve: A Signal Processing Point of View
Soo-Chang Pei and Kuo-Wei Chang
- 164 Tips & Tricks**
Fast- and Low-Complexity atan2(a,b) Approximation
Vicente Torres, Javier Valls, and Richard Lyons
- 176 Perspectives**
To the Victor Go the Spoils: AI in Financial Markets
Xiao-Ping (Steven) Zhang

DEPARTMENTS

- 4 From the Editor**
Signals and Signal Processing: The Invisibles and the Everlastings
Min Wu
- 5 President's Message**
Signal Processing Is More than Its Beloved Name
Rabab Ward
- 170 Dates Ahead**



IEEE prohibits discrimination, harassment, and bullying.
For more information, visit
<http://www.ieee.org/web/aboutus/whatis/policies/p9-26.html>

IEEE Signal Processing Magazine

EDITOR-IN-CHIEF

Min Wu—University of Maryland, College Park
U.S.A.

AREA EDITORS

Feature Articles

Shugang Robert Cui—University of California,
Davis, U.S.A.

Special Issues

Douglas O'Shaughnessy—INRS, Canada

Columns and Forum

Kenneth Lam—Hong Kong Polytechnic University,
Hong Kong SAR of China

e-Newsletter

Ervin Sejdic—University of Pittsburgh, U.S.A.

Social Media and Outreach

Andres Kwasinski—Rochester Institute
of Technology, U.S.A.

EDITORIAL BOARD

Mrityunjoy Chakraborty—Indian Institute of
Technology, Kharagpur, India

George Christikos—Qualcomm, Inc.,
U.S.A.

Alfonso Farina—Leonardo S.p.A., Italy

Mounir Ghogho—University of Leeds,
U. K.

Lina Karam—Arizona State University, U.S.A.

C.-C. Jay Kuo—University of Southern California,
U.S.A.

Sven Lončarić—University of Zagreb, Croatia

Brian Lovell—University of Queensland, Australia

Jian Lu—Qihoo 360, China

Henrique (Rico) Malvar—Microsoft Research,
U.S.A.

Yi Ma—ShanghaiTech University, China

Stephen McLaughlin—Heriot-Watt University,
Scotland

Athina Petropulu—Rutgers University,
U.S.A.

Peter Ramadge—Princeton University,
U.S.A.

Shigeki Sagayama—Meiji University, Japan

Erchin Serpedin—Texas A&M University,
U.S.A.

Shihab Shamma—University of Maryland,
U.S.A.

Vahid Tarokh—Harvard University, U.S.A.

Wade Trappe—Rutgers University, U.S.A.

Gregory Wornell—Massachusetts Institute
of Technology, U.S.A.

Dapeng Wu—University of Florida, U.S.A.

ASSOCIATE EDITORS—COLUMNS AND FORUM

Ivan Bajic—Simon Fraser University, Canada

Rodrigo Capobianco Guido—
São Paulo State University, Brazil

Ching-Te Chiu—National Tsing Hua University,
Taiwan

Panayiotis (Panos) Georgiou—University of
Southern California, U.S.A.

Hana Godrich—Rutgers University, U.S.A.

Xiaodong He—Microsoft Research

Danilo Mandic—Imperial College, U.K.

Aleksandra Mojsilovic—
IBM T.J. Watson Research Center

Vishal Patel—Rutgers University, U.S.A.

Fatih Porikli—MERL
Shantanu Rane—PARC, U.S.A.

Saeid Sanei—University of Surrey, U.K.

Roberto Togneri—The University of
Western Australia

Alessandro Vinciarelli—IDIAP-EPFL

Azadeh Vosoughi—University of Central Florida

Stefan Winkler—UIUC/ADSC, Singapore

Changshui Zhang—Tsinghua University, China

ASSOCIATE EDITORS—e-NEWSLETTER

Csaba Benedek—Hungarian Academy
of Sciences, Hungary

Paolo Braca—NATO Science and Technology
Organization, Italy

Quan Ding—University of California,
San Francisco, U.S.A.

Pierluigi Failla—Compass Inc, New York,
U.S.A.

Marco Guerriero—General Electric Research,
U.S.A.

Yang Li—Harbin Institute of Technology, China

Yuhong Liu—Penn State University at Altoona,
U.S.A.

Andreas Merentitis—University of Athens,
Greece

Michael Muma—TU Darmstadt, Germany

Xiaorong Zhang—San Francisco State University,
U.S.A.

ASSOCIATE EDITOR—SOCIAL MEDIA/OUTREACH

Guijin Wang—Tsinghua University, China

IEEE SIGNAL PROCESSING SOCIETY

Rabab Ward—President

Ali Sayed—President-Elect

Carlo S. Regazzoni—Vice President,
Conferences

Nikos D. Sidiropoulos—Vice President,
Membership

Thrasyloulos (Thrasos) N. Pappas—
Vice President, Publications

Walter Kellerman—Vice President,
Technical Directions

IEEE SIGNAL PROCESSING SOCIETY STAFF

Rebecca Wollman—Publications Administrator

IEEE PERIODICALS MAGAZINES DEPARTMENT

Jessica Welsh, *Managing Editor*

Geraldine Krohn-Taylor,
Senior Managing Editor

Janet Dudar, *Senior Art Director*

Gail A. Schnitzer, *Associate Art Director*

Theresa L. Smith, *Production Coordinator*

Mark David, *Director, Business Development -
Media & Advertising*

Felicia Spagnoli, *Advertising Production Manager*


Dawn M. Melley, *Editorial Director*

Peter M. Tuohy, *Production Director*

Fran Zappulla, *Staff Director,
Publishing Operations*

Digital Object Identifier 10.1109/MSP.2017.2750305

SCOPE: IEEE Signal Processing Magazine publishes tutorial-style articles on signal processing research and applications as well as columns and forums on issues of interest. Its coverage ranges from fundamental principles to practical implementation, reflecting the multidimensional facets of interests and concerns of the community. Its mission is to bring up-to-date, emerging and active technical developments, issues, and events to the research, educational, and professional communities. It is also the main Society communication platform addressing important issues concerning all members.



ENGINEERING SUCCESS

At SNHU, we take education to new heights. The STEM programs at SNHU feature expert faculty, high-tech labs, and exciting internship opportunities. We take pride in providing our students with the tools and resources they need to analyze data, earn hands-on experience, and launch successful careers. SNHU currently offers:

**Aeronautical Engineering | Air Traffic Management | Aviation Management | Computer Science | Construction Management
Electrical and Computer Engineering | Mechanical Engineering**

New majors. New facilities. New ways to interpret the world. At SNHU, we'll help you earn the real-world experience employers want. It's all part of our ongoing commitment to seeing our students achieve their education and career goals. Come see for yourself! Visit the SNHU campus today.

Southern
New Hampshire
University

College of
Engineering, Technology,
and Aeronautics

snhu.edu | admission@snhu.edu | 603-645-9611

FROM THE EDITOR

Min Wu | Editor-in-Chief | minwu@umd.edu

Signals and Signal Processing: The Invisibles and the Everlastings

When you receive this issue of *IEEE Signal Processing Magazine*, a symposium, “The Future of Signal Processing,” was just held at the Massachusetts Institute of Technology (MIT). The symposium honored the career of Prof. Alan Oppenheim as one of the pioneers in signal processing research and education. Attendees from various organizations around the world discussed and shared insights of the profound roles that signal processing have played and envisioned the future trends of signal processing.

I delivered a talk with the same title as this editorial at the MIT symposium. The term *invisibles* has a dual meaning to me. A central theme of my research has been dealing with “micro signals” that are small in strength or scale by at least an order of magnitude and are nearly invisible, yet developing the theory and techniques to extract and utilize these invisible micro signals opens up new opportunities in a broad range of applications from security and forensics to data analytics to entertainment. One class of micro signals provides telltale traces of evidence in determining the origin and integrity of images, which is an active research area investigated by the Information Forensics and Security Technical Committee (IFS TC) of the IEEE Signal Processing Society (SPS) and the subject of the ongoing SP Cup 2018 competition (see page 175) and the latest outreach video “Multimedia

Forensics,” available online at the SPS Resource Center; please visit <http://rc.signalprocessingsociety.org/sp/product/conference-videos-and-slides/SPSVID00194>. Meanwhile, the profound role and contributions of signal processing are often invisible to the public, leading to the notion of “Signal Processing Inside.” In this issue, the second edition of the new “Community Voices” column presents the thoughts on such a topic by our magazine readers who are at various career stages and come from different regions and backgrounds.

Several other formal and informal gatherings have been held this year, celebrating the careers of signal processing pioneers and significant contributors: among them are Prof. Sanjit Mitra, who had a broad range of research interests over the years and nurtured signal processing activities in a number of underrepresented countries and regions; Prof. Mos Kaveh, who played a key role in developing statistical signal processing and served as the IEEE SPS president in 2010–2011; and Dr. John Cozzens, who led the signal processing program at the U.S. National Science Foundation for many years, just to name a few. Thanks to the persistent contributions of them and many others over the past decades, the field of signal processing has grown and our community has expanded both technically and geographically.

It has been a year and half since we launched the redesign of the print version of the magazine. I hope you enjoy the modern look and enhanced graphics of the magazine and its correspond-

ing electronic version. The second part of the redesign effort is for the online presence of our magazine. Although the timetable of the magazine’s web design was deferred to give priority to the major redesign of the SPS’s website, I am happy to report that the matching design for our magazine’s website is well underway. The first phase has been completed and launched this summer for the monthly “Inside Signal Processing eNewsletter” that complements the print version of the magazine. If you haven’t already, please check it out at <http://signalprocessingsociety.org/newsletter/>. My sincere thanks to Christian Debes, the area editor for eNewsletter, and Ervin Sejdic, who succeeded Christian in June 2017, and SPS Web Administrator Rupal Bhatt for their dedicated efforts.

The second phase of the website redesign is currently being implemented with the goal of creating a modern landing page that can host timely updates based on the magazine’s bimonthly content and well-organized resources for prospective authors.

This is the magazine’s final issue of 2017 and the last issue for which I serve as editor-in-chief. Looking back, this three-year journey has been a huge undertaking, and it could not have been possible without the hard work and support of many colleagues. A number of unsung heroes, whom ordinary readers may not have seen or known, contributed to the success of our magazine.

(continued on page 7)

Digital Object Identifier 10.1109/MSP.2017.2750306
Date of publication: 13 November 2017

PRESIDENT'S MESSAGE

Rabab Ward | SPS President | rababw@ece.ubc.ca

Signal Processing Is More than Its Beloved Name

Since its inception in 1948, the IEEE Signal Processing Society (SPS) has evolved in pace with the many technological changes and advancements in our field. In its early days, our Society—the first and oldest among the IEEE's Societies—was known as the *Professional Group on Audio of the Institute of Radio Engineers*. Over the course of four decades, our name has changed few times from *Audio* to *Audio* and *Electro-Acoustics* and then to *Acoustics, Speech, and Signal Processing*, and then to *Signal Processing* to reflect the field's growth and diversity, becoming the *IEEE Signal Processing Society* in 1989.

Since then, our scope of interest has been revised twice to reflect new theories and applications, and many SPS technical committees have also changed their names. Our Society has also developed many new workshops, conferences, specialized publications, journals, periodicals, and outreach programs in an effort to celebrate the achievements of our members, strengthen our industry networking opportunities, and also increase public awareness about signal processing.

We've made great strides, but our field is consistently evolving while eyeing the future. Over the past few years, members have suggested that we're perhaps due for another name change, that the term *signal processing* is obscure and doesn't adequately capture the scope, range, dynamic nature, and fundamental impact of our

chosen field on so many facets of everyday life. We are certainly not alone in this dilemma. For example, the famous mathematician Stanislaw Ulam wrote: "What exactly is mathematics? Many have tried but nobody has really succeeded in defining mathematics; it is always something else. Roughly speaking, people know that it deals with numbers, figures, with relations, operations, and that its formal procedures involving axioms, proofs, lemmas, and theorems have not changed since the time of Archimedes."

By comparison, our field is in its infancy, but it has grown and expanded rapidly to include many branches and subspecialties. So, in December 2013, our Society formed a committee, headed by Prof. Petar Djuric, to explore the possibility of a new name. The committee wrote a wonderful blog post about this topic (see https://signalprocessing.society.org/sites/default/files/uploads/get_involved/docs/Power_of_a_Name_Article_and_Comments.pdf), soliciting member feedback, and listed nine previously suggested name changes:

- 1) Society on Signal Science and Engineering
- 2) Society on Signal Processing and Data Science
- 3) Society on Signal and Data Science
- 4) Society on Signal Science and Processing
- 5) Society on Data Science and Processing
- 6) Society on Data and Signal Processing
- 7) Signal and Information Processing Society
- 8) Society for Data Science
- 9) Data Science Society.

Notice that the term *data science* is in five of these nine suggestions and *data* is in six, but the word *information* is only in one. The post elicited a lively discussion. Among the 58 respondents, some favored a name change, and *Signal and Information Processing* was the most popular of the proposed names, favored by 21 respondents. Yet the majority of respondents (28) preferred to stick with our current name, saying that while it may not be inclusive of everything that we do, it's the most succinct way to describe our complex, evolving field. Indeed, the very definition of a *signal* means the conveyer of some type of information, while the information within the signal is often related to knowledge and intelligence.

The respondents were certainly not unanimous that a name change would either increase or decrease our visibility among the general public, while also reflecting the monumental changes in our field since 1989. Whatever the ultimate decision, I agree with the recommendations that we would certainly benefit from improving the strength and clarity of our brand messaging, by articulating the impact of signal processing on so many fields, such as finance, seismology, satellite communication, medical instruments, and a wide range of commercial electronics and wearable technologies that billions of people use every day—at work, at play, and, in almost every facet of communication.

Highlights from Society members' responding to Prof. Djuric's blog post

Digital Object Identifier 10.1109/MSP.2017.2750307
Date of publication: 13 November 2017

about the name change proposal are given next. I'm pleasantly surprised by the number and depth of these comments. It shows that this topic is timely and of great interest to our community.

Signal processing is "present in nearly all the trendy mobile devices," according to one respondent, yet it's not well understood by our peers in science, industry, and the general public "is oblivious to the concept." It's a fitting paradox for signal processing, which is described in the book *Essentials of Digital Signal Processing* [1] as the "phantom technology because it is so pervasive and yet not well understood."

Another participant agreed, writing, "Signal processing is still a mystery to many of our peers, and it does not adequately reflect the current activities." Others pointed out that this dilemma hinders our ability to attract good students to the field, negotiate promotions at universities and corporations, and "build a visible ecosystem" upon which individuals could envision a career.

The diversity in theories and applications within our field can be viewed as both a benefit and a hurdle. One person wrote that our branding challenges "will get worse with signal processing getting more diverse and intangible" as the emphasis shifts from boards and circuits to software applications. Yet how do we strive to be both more inclusive and more succinct with our branding? Some participants suggested adding various qualifiers, the most popular of which were *data*, *data science*, *science*, *signal science*, *engineering*, and *information processing*.

The term *data* received a few favorables. With the increased emphasis and publicity on data in recent years, I wonder whether more respondents would have favored this term had the blog been posted a couple of years later. Others felt that while it's currently trendy and may have increasing funding opportunities, it may have a short shelf life and it's too broad and generic and too specific to computer science, data processing and "big data"—implying all are poorly understood by the general public. There are also educational differences to consider. One respondent pointed out that,

while signal processing necessitates an advanced scientific education and carefully conducted scientific protocols, it only takes a few courses in computer science and web programming to become a "data scientist."

The addition of the term *science* also received mixed reviews. As one respondent pointed out, "We are not scientists—we are engineers, and I for one am damn proud to call myself an engineer. Scientists take things apart; engineers put things together. Not only are these fields different, they are polar opposites." Someone else wrote, "Engineers is what we are, and signal processing is what we do." Several people opposed the addition of the term *engineering* to our name. One respondent pointed out that engineering is also misunderstood by the public as dealing with work that signal processors do not do, such as work related to engines.

Adding the term *information processing* was the most popular alternative among the respondents, primarily because it best conveyed the diversity of our field and our goal to "strive to be inclusive of all its members."

Yet other members called *information processing* redundant, saying that you cannot process information if it does not induce a signal; only signals that contain information can be processed and "signal processing" allows for information in a signal to be available in a "convenient format." As one respondent put it, "the word *signal* already indicates an information-bearing phenomenon, and *signal processing* already encompasses the decoding/encoding of any kind of information." However, I wish to add here that besides "processing," much of our work involves understanding and learning about the systems we study.

Among the various proponents of maintaining our current name, some voiced concern that a name change would dilute the brand name. The term *signal processing* is well established, featured in many journal titles, conference names, and the majority of academic programs. University electrical engineering departments teach subjects with *signal processing* in the title, and these courses are often first-year courses offered to under-

graduates in electrical and computer engineering, which piques the interest of young students, and sets them on the path to become signal processors.

Other supporters pointed out that the SPS is already a well-respected brand in the science and engineering community. "Let's keep the name and improve our outreach and publicity efforts," wrote one respondent. Another member agreed, writing, "Better outreach and publicity will fix this issue." However instead of "marketing the subtleties of a denoising algorithm that optimizes some supercool theoretical function," we should showcase the latest cutting-edge technologies. Another member agreed, writing, "When people ask 'What is SP?' I say it is 'everything that goes on inside a smartphone' and their eyes suddenly light up."

"Whatever the new name of the Society, I will still say that I am a signal processing guy," wrote one commentator. Another wrote, the term "*signal* must be kept since it represents the human instinct to communicate since the prehistoric age."

Reading the comments on this blog has given me, a proponent of a name change, much to think about. On one hand, a name change, e.g., by adding the term *data science*, would, at present, help us increase our visibility and capture the interests of students, friends, and other Societies, as well as the corporations and industries that provide employment and help fund our research and development. Our field has definitely evolved much further than processing signals measured by electronic devices and grew to processing, understanding, and learning from data, irrespective to whether or not it is obtained from physical or physiological processes. Also, many of the concepts and theories we have advanced have been abstracted for use in a large number of applications.

On the other hand, data science has much overlap with signal processing, mainly nonparametric, high-dimensional statistical signal processing (which involves big data and does not model the process). Thus it could be strongly argued that data science falls with the realm of signal processing.

Furthermore, signal processing has now become much larger and diverse. It has permeated a vast number of technologies and applications. Watch, for example, our video “What Is Signal Processing?” at <https://www.youtube.com/watch?v=EErkgr1MWw0> for some examples of these applications. The scope of our journals range from speech to networks, from forensics to imaging, from biomedical to multimedia, and so on. We have more than 185 Chapters in approximately 120 countries. This compels me to appreciate the good comments and relevant points made by

those who advocate for keeping our current name, as well as feel their devotion to the name. Whether or not we change our name, we should continue to expand our activities to stress our wide scope, for example, by initiating a new journal, workshop, distinguished speakers, summer schools, and educational material related to data science. I consider myself fortunate that I have been working in this exciting field and am equally proud to be attached to our beloved name *signal processing*.

Let us continue this important discussion. Please add your comments

to <https://signalprocessingsociety.org/get-involved/signal-processing-larger-its-beloved-name>



Reference

[1] R. G. Lyons and D. Lee Fugal, *Essentials of Digital Signal Processing*. Englewood Cliffs, NJ: Prentice Hall, 2014.



FROM THE EDITOR (continued from page 4)

Managing Editor Jessica Welsh and the IEEE Magazines Department production team are a driving force in interacting with authors and creating a professional look and feel for the articles. In addition, Senior Art Director Janet Dudar and Associate Art Director Gail Schnitzer help bring eye-catching artistic elements to each issue of the magazine.

I have also had the privilege to work very closely with a team of area editors who play a key role in the magazine operations: Shuguang Robert Cui screens dozens of feature article proposals each year and tirelessly coordinates the proposal reviews; Kenneth Lam leads a team of dedicated column associate editors to enrich the magazine content to serve our broad readership; Douglas O’Shaughnessy coordinates the special issue efforts, a signature tradition of the magazine; Andres Kwasinski, Erwin Sejdic, Christian Debes, and the associate editors on their teams who contributed to the electronic efforts that complement the print version of the magazine.

Our senior Editorial Board members bring a diverse set of expertise and perspectives and provide candid feedback and guidance; special thanks to 13 members who are completing their

three-year services: Mounir Ghogho, Lina Karam, Sven Lončarić, Brian Lovell, Stephen McLaughlin, Yi Ma, Henrique (Rico) Malvar, Athina Petropulu, Peter Ramadge, Shigeki Sagayama, Erchin Serpedin, Shihab Shamma, Gregory Wornell, and Dapeng Wu. In addition, special issue and cluster organizers work intensively to bring timely content to our readers, and each special issue, cluster, or series is a major undertaking by itself. My sincere thanks to all authors for contributing to the magazine, especially for the time and hard work it takes to make content accessible, and to the many reviewers who provided timely assessments and constructive comments to ensure the high technical and presentation quality of the articles. You can find an annual index of authors and articles associated with each year-end issue of the magazine in IEEE *Xplore*. The collective effort by authors, reviewers, and editors helped our magazine reach an all-time high in impact factor (9.65) and article influence score (4.02) in the most recent *Journal Citation Report*. Last but not the least, I thank SPS staff members Rebecca Wollman, Richard Baseil, Theresa Argiropoulos, Jessica Perry, and Deborah Blazek for their assistance, and I appreciate the

thoughtful feedback and support from our readers.

The SPS Executive Committee has appointed Prof. Robert Heath as the next editor-in-chief of the magazine, effective January 2018. Robert is a world-renowned expert and proliferate researcher on signal processing for communications. Please join me in welcoming him. As I pass the baton, I take this opportunity to thank all supporters of the magazine in the past and appreciate the continued support in the years to come. Together, we can continue to build this premier publication with a strong technical impact as well as indispensable benefits to our members and readers.

To quote Prof. Oppenheim: “There will always be signals, they will always need processing, and there will always be new applications, new mathematics, and new implementation technologies.” Let *IEEE Signal Processing Magazine* be your helpful companion in this everlasting journey of signal processing!




PANEL AND FORUM

Yonina C. Eldar, Alfred O. Hero III, Li Deng, Jeff Fessler, Jelena Kovačević, H. Vincent Poor, and Steve Young

Challenges and Open Problems in Signal Processing: Panel Discussion Summary from ICASSP 2017

This column summarizes the panel on open problems in signal processing, which took place on 5 March 2017 at the International Conference on Acoustics, Speech, and Signal Processing (ICASSP) in New Orleans, Louisiana. The goal of the panel was to draw attention to some of the challenges and open problems in various areas of signal processing and generate discussion on future research areas that can be of major significance and impact in signal processing. Five leading experts representing diverse areas within signal processing made up the panel:

- Li Deng, Microsoft: machine learning
- Jeff Fessler, the University of Michigan: medical imaging
- Jelena Kovačević, Carnegie Mellon University: graph signal processing
- H. Vincent Poor, Princeton University: wireless communication
- Steve Young, the University of Cambridge: speech and language processing.

It was organized and moderated by Yonina Eldar from the Technion and Alfred O. Hero III from the University of Michigan.

The panel drew a very large crowd and stimulated a vibrant discussion on directions, trends, and challenges of signal processing in the 21st century and in the era of big data. In this column, we summarize the main points raised by the panelists and the audience in each of

the aforementioned topics. Our goal and hope is to further the discussion on some of the main challenges and opportunities for signal processing in the coming years and to highlight areas where, as a community, working and collaborating together, we may have impact on theory, applications, and education.

Next, we summarize open problems in the previously mentioned areas, highlighted by the participants: open problems in machine learning, medical imaging, graph signal processing, physical layer wireless communications, and speech and language processing. A common cross-cutting theme that emerged was the opportunity to improve performance by the better integration of accurate physical models into state-of-the-art algorithms.

Open problems in machine learning

Machine learning aims to give computers the ability to learn by exploiting data instead of being explicitly programmed. There are many approaches in machine learning, including support vector machines, decision-tree learning, artificial neural networks, Bayesian networks, genetic algorithms, rule-based learning, and inductive logical programming, among others [3]. In recent years, the fastest growing area of machine learning comes from neural networks and related generative models, where carefully designed hierarchies are built into the overall machine-learning models to form multiple layers of latent representations that disentangle the confounding factors and complexity in the raw data. This type of

hierarchical model and the associated machine-learning algorithms are called *deep learning* [1], [2], which represents the most recent and influential advance in machine learning over the past decade. The first successful application of deep learning in real-world tasks came from speech recognition in our signal processing community and was published in this magazine [13], followed quickly with computer vision, natural language processing, robotics, speech synthesis, and image rendering [2].

Despite impressive empirical successes of deep learning and other machine-learning approaches, many open problems remain to be solved. Current deep-learning methods typically lack interpretability, in contrast to traditional machine-learning techniques based on linear models. In a number of applications, deep-learning methods achieve recognition accuracy close to or exceeding that of humans, but they require considerably more training data, power consumption, and computing resources than humans. In addition, although accuracy results are often statistically impressive, they are often unreliable on an individual basis. Finally, most of the current deep-learning models have no reasoning and explaining capabilities, making them vulnerable to disastrous failures or attacks without the ability to foresee and thus to prevent them.

To overcome these challenges, both fundamental and applied research is needed. One potential breakthrough in machine learning is in developing

interpretable deep-learning models with the aim of creating new algorithms and methods that can overcome current limitations of machine-learning systems in their lack of ability to explain the actions, decision, and prediction outcomes to human users while promising to perceive, learn, decide, and act on their own. This new class of machine-learning systems will allow users to understand and thus trust the system's outputs and to foresee and predict future system behaviors. To this end, neural networks and symbolic systems need to be integrated, enabling the machine-learning systems themselves to construct models that will explain how the world works. That is, they will discover by themselves the underlying causes or logical rules that shape their prediction and decision-making processes interpretable to human users in symbolic and natural language forms. An initial work in this direction makes use of an integrated neural-symbolic representation called *tensor-product neural memory cells*, which can be decoded back to symbolic form without loss of information after extensive learning in the neural-tensor domain.

Another potential breakthrough in machine-learning research is in new algorithms for reinforcement and unsupervised deep learning, which make use of weak or even no training signals paired to inputs to guide the learning. Effective reinforcement-learning algorithms will allow machine-learning systems to learn via interactions with possibly adversarial environments and with themselves.

The most challenging problem, however, is unsupervised learning, for which no satisfactory machine-learning algorithms have been devised so far in practical applications. The development of unsupervised learning techniques is significantly behind that of supervised and reinforcement deep learning. The most recent development in unsupervised learning exploits sequential output structure and advanced optimization methods to alleviate the need for using labels in training prediction systems [12].

Future advances in unsupervised learning include taking into account new sources of learning signals such as the structure of input data and building conditional generative models. In this context, the recent popular topic of generative adversarial networks [2] is a highly promising direction exploiting the long-standing concept of analysis by synthesis. A closely related open problem is multimodal deep learning with cross-domain information as low-cost supervision. Standard speech recognition, image recognition, and text classification methods make use of supervision labels within each of the speech, image, and text modalities separately. This is far from how children learn to recognize speech and classify text. For example, children often get a distant "supervision" signal for speech sounds by an adult pointing to an image scene or text.

A final future direction for tackling open problems in machine learning is the paradigm of learning-to-learn or metalearning; i.e., how to design a machine-learning system that improves or automatically discovers a learning algorithm. Learning-to-learn is a powerful emerging paradigm and a fertile research direction expected to impact a wide range of real-world applications.



Holcombe Department of Electrical and Computer Engineering Faculty Search in Computer Engineering and Electrical Engineering

The Holcombe Department of Electrical and Computer Engineering at Clemson University is seeking applicants for multiple computer engineering and electrical engineering tenure-track or tenured faculty positions at the rank of assistant professor or associate professor. The Department has a particular interest in applicants in the following technical areas: (1) machine learning, computer vision, artificial intelligence, signal processing, with collaborations in biomedical engineering, health science, or automotive engineering; (2) embedded computing, sensors, wearables; (3) high-performance computing with an emphasis on big data, high-performance networking, or accelerated computing architectures; and (4) cyber security and cyber-physical system security. Outstanding assistant-professor candidates will be considered for the Warren Owens Assistant Professorship.

The Holcombe Department of ECE is one of the largest and most active departments in Clemson University, with 32 primary and 14 affiliated full-time faculty members, approximately 550 undergraduates and 190 graduate students. Annual research expenditures exceed \$8.6 million. Many members of the faculty are known internationally; they include eight IEEE Fellows, three endowed chairs, and four named professors. Annual funded research expenditures exceed \$8.6 million. The Department and Clemson have highly successful computing-focused research programs in high-performance computing and networking; privacy, communications security, and secure control systems; and mobile health devices.

Clemson University is the largest land-grant institution in South Carolina, enrolling 18,600 undergraduates and 4,800 graduate students. Seven colleges house strong programs in architecture, engineering, science, agriculture, business, social sciences, arts and education. A faculty of 1,500 and staff of 3,700 support 84 undergraduate degree offerings, 73 master's degree programs and 40 Ph.D. programs. An annual operating budget of approximately \$956 million and an endowment of \$621 million fund programs and operations. The University has externally funded research expenditures of \$100 million per year. Research and economic development activities are enhanced by public-private partnerships at 4 innovation campuses and 6 research and education centers located throughout South Carolina. Clemson University is ranked 23th among national public universities by U.S. News & World Report.

Applicants must have an earned doctorate in electrical engineering, computer engineering, or a closely related field. Applicants should submit a current curriculum vitae, statements of research and teaching strategy, and a minimum of five references with full contact information. Application material should be submitted electronically at the following Web link:

<http://apply.interfolio.com/39731>

To ensure full consideration, applicants must apply by December 1, 2017; however, the search will remain open until the position is filled.

Clemson University is an AA/EEO employer and does not discriminate against any person or group on the basis of age, color, disability, gender, pregnancy, national origin, race, religion, sexual orientation, veteran status or genetic information. Clemson University is building a culturally diverse faculty committed to working in a multicultural environment and encourages applications from minorities and women.

Open problems in medical imaging

Medical image reconstruction is the process of forming interpretable images from the data recorded by an imaging system. Until recently, there have been two primary methods for image reconstruction: analytical and iterative. Analytical methods use idealized mathematical models for the imaging system. Typically, these techniques consider only the geometry and sampling properties of the imaging system and ignore details of the system physics and measurement noise. These reconstruction approaches have been used extensively because they require modest computation.

Over the past two decades, image reconstruction has evolved from the exclusive use of analytical methods to a wider use of model-based approaches that account for the physics and statistics. Usually the problems are ill posed, so that maximum-likelihood (ML) methods would propagate excessive noise from the measurements into the reconstructed image. Using priors or regularizers can overcome this limitation. A popular approach is to base iterative methods on maximum a posteriori (MAP) estimation. MAP estimation encompasses 1) modeling the system, 2) developing signal models to serve as priors, 3) developing faster optimization algorithms, and 4) assessing the quality of the reconstructed image.

The transition from analytical to iterative algorithms took place at widely different dates in different modalities. In positron emission tomography (PET) and single-photon emission computed tomography (SPECT), a seminal paper on an expectation maximization (EM) algorithm in the early 1980s led to more than a decade of research before a key acceleration method called *ordered subsets (OS)* (related to incremental gradients in the optimization field) helped lead to the commercial adoption of OS-EM for clinical PET and SPECT in about 1997, using an (unregularized) ML approach. This transition provided a dramatic improvement in image quality. Human PET scanners only recently began to provide MAP methods clinically using a modification of a Gaussian Markov random field prior and a convergent OS algorithm.

In X-ray computed tomography (CT), iterative image reconstruction first became available commercially for the CT part of SPECT-CT scanners in about 2010, using a different OS algorithm published a decade earlier. In 2012, the first U.S. Food and Drug Administration (FDA)-approved iterative MAP method targeted at reduced X-ray dose became available for clinical CT, building on an *IEEE Transactions on Signal Processing* paper from two decades earlier. This approach also uses a modified Gaussian MRF to make it edge preserving.

In MRI, researchers studied iterative techniques to quantify relaxation parameters, reconstruct data from multiple receive coils, and correct for magnetic field inhomogeneities. A turning point was the introduction of compressed sensing in about 2005, spawning an explosion of research that finally led to FDA approval of compressed sensing MRI products in 2017 using combinations of total variation regularization and wavelet sparsifying transforms. In all of the aforementioned examples, more than a decade passed between the key publication and commercial availability of the method!

Commercial MAP techniques use relatively simple priors defined mathematically. The emerging research trend is to explore signal models that are learned from data. In X-ray CT, there are numerous images acquired at “normal” X-ray doses from which one can learn signal models to use later for reconstructing images from low-dose data. Another data-driven option is to learn a sparse signal model during image reconstruction, rather than relying on training data, called *blind* or *adaptive dictionary* (or *transform*) *learning*. This data-driven evolution provides opportunities for signal processing researchers to explore signal models that better solve inverse problems, particularly from limited or noisy data.

One can “unroll the loop” of an iterative reconstruction algorithm and treat it as a sequence of processing steps akin to a deep neural network and then use data to train more aspects of the processing chain. Recent conferences have seen an explosion of such methods. There are many significant challenges because such algorithms are arguably even more nonlinear

(and opaque) than the edge-preserving regularization techniques used clinically today. Can one characterize the “resolution” and “noise” properties of such algorithms? What is the best training metric: MSE or diagnostic image quality? What if a patient has significantly different image features than those found in the training data? How well will a method trained for one system configuration (e.g., a certain set of coils in MRI or a certain set of angular views and pitch in CT) generalize to other configurations? Some experts have conjectured that “machine learning will transform radiology significantly within the next five years” but others point out there are significant technical and legal challenges. These questions and more should provide numerous research opportunities for signal processors interested in inverse problems like medical imaging [11].

Open problems in graph signal processing

Today’s data is being generated at an unprecedented rate from a diversity of sources. Examples include profile information in social networks, stimuli in brain connectivity networks, and traffic flow in city street networks, among others. A decade ago, a typical data set was supported on a regular lattice; today, the story is quite different. Data is supported on complex and irregular structures. Often, these structures are modeled by graphs, as they are able to describe both the structure and the data associated with that structure. For example, in an online social network, a user’s profile may contain the user’s date of birth, school attended, professional organizations, and more. Each of these attributes can form a subnetwork with different properties. Using graphs, we want to analyze data supported on such complex structures, allowing us to mine information from online social networks, transportation networks, the power grid, and more, in the same context. While these are representatives of physical-world graphs, other graphs may include abstract concept networks such as knowledge graphs and correlation graphs.

Data science on graphs has been considered from several angles by graph

theory, network science, and graph mining, all dealing with graph structure. More recently, the area of graph signal processing has emerged, formalizing the addition of metadata as signals on a graph [4]–[6]. Graph signal processing aims to extend classical signal processing tasks and tools to data on irregular structures modeled by graph signals (see Figure 1). The goal is to gain an understanding of the intrinsic structure of the data by using tools well understood on regular structures, such as filtering and Fourier transforms, and to perform tasks such as sampling, restoration, compression, and topology learning.

Signal processing on graphs is an active area of research; many challenges and opportunities still remain. For example, a number of basic concepts in statistical signal processing and sampling theory have not yet been entirely extended to graphs in a unified way. More advanced challenges include the scale of the data, its heterogeneity, distributed analysis and processing, fusing data from different scales and resolutions, and processing tensor values defined on nodes. Disparate communities such as network science, machine learning, and signal processing are all currently working on these challenges with the tendency to attack such problems either via learning methods or by building models; an important path for advancing this field and dealing effectively with the deluge of data is to combine the tools and integrate these different approaches.

Open problems in physical layer wireless communications

Wireless communications have been a major driver of signal processing research for at least the past three decades, spurred by the development of widespread consumer mobile communications and other applications, which today impact the lives of billions of people—indeed, most people alive today. Here we focus on research in the physical layer of mobile communication networks where signal processing has perhaps had the greatest impact.

Modern mobile communication networks have been through four generations to date, and the fifth generation (5G)

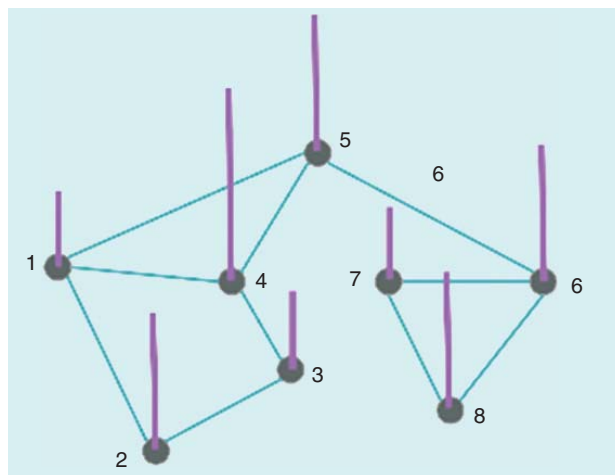
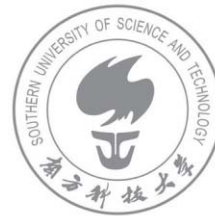


FIGURE 1. A graph signal models data (values on the graph nodes) supported on complex structures (graph nodes).



Professor/Associate Professor/Assistant Professorship in the Department of Electrical and Electronic Engineering

The University

Established in 2012, the Southern University of Science and Technology (SUSTech) is a public institution funded by the municipal of Shenzhen, a special economic zone city in China. Shenzhen is a major city located in Southern China, situated immediately north of Hong Kong Special Administrative Region. As one of China's major gateways to the world, Shenzhen is the country's fast-growing city in the past two decades. The city is the high-tech and manufacturing hub of southern China, home to the world's third-busiest container port, and the fourth-busiest airport on the Chinese mainland. A picturesque coastal city, Shenzhen is also a popular tourist destination and was named one of the world's 31 must-see tourist destinations in 2010 by The New York Times. The Southern University of Science and Technology is a pioneer in higher education reform in China. The mission of the University is to become a globally recognized institution which emphasizes academic excellence and promotes innovation, creativity and entrepreneurship. The teaching language at SUSTech is bilingual, either English or Putonghua. Set on five hundred acres of wooded landscape in the picturesque Nanshan (South Mountain) area, the new campus offers an ideal environment suitable for learning and research.

Call for Application

The Southern University of Science and Technology now invites applications for the faculty position in the Department of Electrical and Electronic Engineering. It is seeking to appoint a number of tenured or tenure track positions in all ranks. Candidates with research interests in all mainstream fields of electrical and electronic engineering will be considered, including but not limited to IC Design, Embedded Systems, Internet of Things, VR/AR, Signal and Information Processing, Control and Robotics, Big Data, AI, Communication/Networking, Microelectronics, and Photonics. SUSTech adopts the tenure track system, which offers the recruited faculty members a clearly defined career path. Candidates should have demonstrated excellence in research and a strong commitment to teaching. A doctoral degree is required at the time of appointment. Candidates for senior positions must have an established record of research, and a track-record in securing external funding as PI. As a State-level innovative city, Shenzhen has chosen independent innovation as the dominant strategy for its development. It is home to some of China's most successful high-tech companies, such as Huawei and Tencent. As a result, SUSTech considers entrepreneurship is one of the main directions of the university, and good starting supports will be provided for possible initiatives. SUSTech encourages candidates with intention and experience on entrepreneurship to apply.

Terms & Applications

To apply, please send curriculum vitae, description of research interests and statement on teaching to eehire@sustc.edu.cn. SUSTech offers competitive salaries, fringe benefits including medical insurance, retirement and housing subsidy, which are among the best in China. Salary and rank will commensurate with qualifications and experience. Candidates should also arrange for at least three letters of recommendation sending directly to the above email account. The search will continue until the position is filled. For informal discussion about the above posts, please contact Professor Xiaowei SUN, Head of Department of Electrical and Electronic Engineering, by phone 86-755-88018558 or email: sunxw@sustc.edu.cn.

is rapidly emerging. The key enablers of the most recent deployed generation of mobile networks, the so-called fourth generation (4G), have been the development of methods to exploit the spatial diversity afforded by the wireless medium in the forms of multiple-input, multiple output (MIMO) antenna systems, cooperation, and relaying; the exploitation of frequency diversity through the use of orthogonal frequency-division multiple access signaling; and the development of methods to approach link capacity via the iterative decoding of turbo or low-density parity-check codes. These signal processing advances have allowed 4G networks to meet the challenge of real-time multimedia communications that has been the primary advance of 4G over its predecessors.

The emerging generation of mobile networks, 5G, presents a number of new signal processing challenges. Beyond providing adequate capacity and reliability, 5G networks also add the issue of energy efficiency, required to support several new applications areas. These include the so-called Internet of Things (IoT), which is envisioned to involve orders-of-magnitude more terminals than 4G networks in highly densified configurations of low-complexity terminals; systems requiring autonomy or telecontrol, in which low latency and very high reliability are critical; and immersive experiences, such as virtual reality, which require very high bandwidth streaming [7].

These requirements give rise to a number of open problems and potential solutions. Solutions enabling densification and the consequent interference management include cloud radio access networks, massive MIMO systems, millimeter wave techniques, and transceivers that can harvest radio-frequency energy from their surroundings. Substantial capacity enhancements are also needed, and some techniques for providing greater capacity (in addition to densification of resources) include full duplex transmission and nonorthogonal multiple-access techniques, both of which will be enabled by sophisticated signal processing. Security is another issue in which signal processing has a key role to play; traditionally,

security has been a higher-layer issue, with encryption being a primary mechanism. However, with highly dense networks of low-complexity terminals connected via loosely organized networks, new methods are needed. Physical layer security is such a promising method, which relies on signal processing techniques, such as coding, beamforming, and signal design. Finally, many emerging applications, such as autonomous vehicles and factory automation, require low-latency, high-reliability communications via short packets. Since the existing theory of reliable data transmission is largely based on analyses in the asymptote of infinite block-length, new theories are needed to understand the limits of reliable communication in this regime. In addition, in applications such as autonomous driving, worst-case metrics may be more appropriate than the standard average-case analysis.

Open problems in speech and language processing

Spoken language processing encompasses methods and techniques for transforming and manipulating speech, text, and a wide variety of related symbolic representations. Examples are speech recognition (speech→words), natural language understanding (words→meaning), natural language generation (meaning→words), speech synthesis (words→speech), and machine translation (words in L1→words in L2). Modern applications of spoken language processing will typically incorporate many if not all of these component technologies [8]–[10]. For example, intelligent agents such as Siri and Alexa require all of the aforementioned technologies to support conversations over a wide range of topics in many languages.

Since virtually all spoken language processing involves classification and/or prediction, modern approaches depend heavily on statistical models and machine learning. A major breakthrough in recent years has been the widespread deployment of deep learning [9]. The ability of neural networks to automatically learn low-level features, the use of attention mechanisms to learn which features are important, and the flexibility to scale parameter sets both in width and depth has led to significant performance

improvements. For example, word error rates for real-time large vocabulary speaker-independent speech recognition are now routinely below 10%, and speech synthesis quality is acceptable for most applications.

The renaissance of neural networks has also been the catalyst for the development of a powerful toolbox of core network components (such as deep neural networks, long short-term memory networks, convolutional neural networks, and more) and development tools (such as TensorFlow, Torch, and others), which allow solutions to complex problems to be assembled, trained, and deployed quickly and at a relatively low cost.

Despite the undoubted progress witnessed over the last decade, there remain many challenges. The recognition of fluent conversations between human speakers and speech in high levels of background noise or in the presence of a competing talker still falls well short of human performance. Our ability to understand the meaning of natural language sentences, especially in the context of past interactions and a changing real-world environment, remains extremely limited.

Two emerging trends aimed at addressing some of the challenges are continuous representations and end-to-end training. In particular, there is currently a shift away from symbolic representations to continuous space representations. An already well-established example of this is the use of word embeddings. By projecting discrete words into a continuous high-dimensional space, many of the problems associated with synonyms, antonyms, and rare words are mitigated by the use of simple well-behaved distance metrics. The extension of embeddings to represent whole sentences and conversations enables variable-length sequences to be mapped into fixed-length vectors that can then be manipulated using conventional classification and prediction models. There is also increasing emphasis on end-to-end training. Conventional systems are typically built as a pipeline of processes for which each component interface needs to be explicitly defined, and training data needs to be appropriately labeled at every component interface. This is expensive and inevitably results in information loss as the

signal propagates through the pipeline. By treating component interfaces as hidden variables, and training end to end, costs are reduced and performance increases.

In summary, the extensive use of machine learning coupled with the availability of large-scale computing and very large data sets have led to a significant improvement across all areas of speech and language processing. Ultimately, however, the real challenge will concern our ability to extract and manipulate the underlying meaning of word sequences and this is an area that has so far remained rather elusive.

Authors

Yonina C. Eldar (yonina@ee.technion.ac.il) is a professor in the Department of Electrical Engineering at the Technion-Israel Institute of Technology, Haifa, where she holds the Edwards Chair in Engineering. She is also an adjunct professor at Duke University, Durham, North Carolina, a research affiliate with the Research Laboratory of Electronics at the Massachusetts Institute of Technology, and was a visiting professor at Stanford University, California. She is a member of the Israel Academy of Sciences and Humanities and is a Fellow of the IEEE and EURASIP. She has received many awards for excellence in research and teaching, including the IEEE Signal Processing Society Technical Achievement Award, the IEEE/AESS Fred Nathanson Memorial Radar Award, the IEEE Kiyo Tomiyasu Award, the Michael Bruno Memorial Award from the Rothschild Foundation, the Weizmann Prize for Exact Sciences, and the Wolf Foundation Krill Prize for Excellence in Scientific Research. She is the editor-in-chief of *Foundations and Trends in Signal Processing* and serves the IEEE on several technical and award committees.

Alfred O. Hero III (hero@eecs.umich.edu) received the B.S. degree (summa cum laude) from Boston University in 1980 and the Ph.D. degree from Princeton University in 1984, both in electrical engineering. He is the John H. Holland Distinguished University Professor of Electrical Engineering and Computer Science and the R. Jamison and Betty Williams Professor of

Engineering at the University of Michigan, Ann Arbor. He is also the codirector of the university's Michigan Institute for Data Science. He has served as president of the IEEE Signal Processing Society and as a member of the IEEE Board of Directors. He has received numerous awards for his scientific research and service to the profession, including several best paper awards, the IEEE Signal Processing Society Technical Achievement Award in 2013, and the 2015 Society Award. He is a Fellow of the IEEE.

Li Deng (l.deng@ieee.org) received his Ph.D. degree from the University of Wisconsin-Madison. He was a tenured professor from 1989 to 1999 at the University of Waterloo, Ontario, Canada, and then joined Microsoft Research, Redmond, Washington, where he was the chief scientist of artificial intelligence (AI) and partner research manager. He recently joined Citadel as its chief AI officer. He is a Fellow of the IEEE, the Acoustical Society of America, and the International Speech Communication Association. He was the editor-in-chief of *IEEE Signal Processing Magazine* (2009–2011) and *IEEE/ACM Transactions on Audio, Speech, and Language Processing* (2012–2014), for which he received the IEEE Signal Processing Society (SPS) Meritorious Service Award. He received the 2015 IEEE SPS Technical Achievement Award for “outstanding contributions to automatic speech recognition and deep learning” and numerous best paper and scientific awards for the contributions to AI, machine learning, multimedia signal processing, speech and human language technology, and their industrial applications.

Jeff Fessler (fessler@umich.edu) received his B.S.E.E. degree from Purdue University, West Lafayette, Indiana, in 1985 and his Ph.D. degree from Stanford University, California, in 1990. He is the William L. Root Professor of Electrical Engineering and Computer Sciences at the University of Michigan, where he has worked since 1990. He received the 2013 IEEE Edward Hoffman Medical Imaging Scientist Award and serves as an associate editor of *IEEE Transactions on*

Computational Imaging. He was the technical program cochair of the 2002 IEEE International Symposium on Biomedical Imaging (ISBI) and the general chair of ISBI 2007. His research group focuses on imaging problems. He is a Fellow of the IEEE.

Jelena Kovačević (jelenak@cmu.edu) received her M.S. and Ph.D. degrees in electrical engineering from Columbia University, New York, in 1988 and 1991, respectively. She is currently the Hamerschlag University Professor, head of the Department of Electrical and Computer Engineering, and a professor of biomedical engineering at Carnegie Mellon University (CMU), Pittsburgh, Pennsylvania. She received the IEEE Signal Processing Society Technical Achievement Award, the Dowd Fellowship at CMU, the Belgrade October Prize, and the E.I. Jury Award at Columbia University. She has coauthored a number of award-winning papers and is a coauthor of the textbooks *Wavelets and Subband Coding* and *Foundations of Signal Processing*. She is a Fellow of the IEEE and was the editor-in-chief of *IEEE Transactions on Image Processing*. Her research interests include applying data science to a number of domains such as biology, medicine, and smart infrastructure; she is an authority on multiresolution techniques such as wavelets and frames.

H. Vincent Poor (poor@princeton.edu) is the Michael Henry Strater University Professor of Electrical Engineering at Princeton University, New Jersey. His interests include information theory and signal processing, with applications in wireless networks and related fields. He is an IEEE Fellow, a member of the U.S. National Academy of Engineering and the U.S. National Academy of Sciences, and a Foreign Member of the Royal Society. He received the Technical Achievement and Society Awards of the IEEE Signal Processing Society in 2007 and 2011, respectively. Recent recognition of his work includes the 2017 IEEE Alexander Graham Bell Medal and honorary doctorates from several universities.

(continued on page 23)

COMMUNITY VOICES

Andres Kwasinski and Min Wu

What Is the Future of Signal Processing?

Views across our community

The goal of the “Community Voices” column in *IEEE Signal Processing Magazine (SPM)* is to encourage and share reflections from diverse members of our community on questions that are of interest to many of us. In this article, we posed the following question to our readers: “After half a century of development, some say signal processing is already matured in terms of theories and techniques, and perhaps would not have a new research breakthrough. Others have observed the problem of *signal processing inside* (a term that was coined by *SPM*’s then editor-in-chief, Prof. K.J. Ray Liu, in his editorial in the September 2004 issue). What are your thoughts on the future of signal processing?”

As we were shaping the question, we were inspired by discussions with Prof. Alan Oppenheim of the Massachusetts Institute of Technology (MIT) and Dr. Thomas A. Baran, who was chairing the organizing committee of MIT’s The Future of Signal Processing Symposium that honored Prof. Oppenheim’s career. We hope that you enjoy reading the responses from our community members around the world. These responses were selected from the online responses we received and have been edited for style, length, and clarity. Please let us know your ideas for future discussion topics by sending your e-mail to Andres Kwasinski (axkeec@rit.edu), area editor for social media and outreach.

Digital Object Identifier 10.1109/MSP.2017.2743841
Date of publication: 13 November 2017

Ahmed I. Humayun

According to the World Health Organization, health is defined as “a state of complete physical, mental, and social well-being, and not merely the absence of disease or infirmity.” Applications of signal processing have brought revolutionary progress in both the physical and mental healthcare domain. But social well-being still remains a lesser-defined term. We live in an era of luxury, where social connectivity has turned more digital than analog; it has discredited the sensory influence of touch, something that could be considered a challenge for digital social platforms in the future. Social interactions confined within quadrilateral screens are not always wholesome. Signal processing needs to confront the problems related to social health, both in the digital domain and physical realm, i.e., in workplaces. With an overload of data and channels all around us, we need a balanced diet of information that would properly address the social health crisis.

Ahmed I. Humayun (ahmed.imtiaz.prio@gmail.com) is studying for his bachelor’s degree in electrical and electronic engineering at the Bangladesh University of Engineering and Technology (BUET). He was on the BUET team that received an honorable mention in the IEEE SP Cup 2017 competition.

Guoru Ding

Signal processing has played an important role over the past decades and will continue to contribute to the development of human life. The fundamental theories of advanced signal processing are key in the future. The cross-disciplinary research between signal processing and other disciplines such as machine learning, data mining, and so on, is a trend, among many trends.

Guoru Ding (dr.guoru.ding@ieee.org) is an assistant professor at Southeast University, Nanjing, China.

Simona Lohan

What we could actually say about any research field on Earth—except maybe medicine and the fields related to the human mind—that they are mature enough and no huge new research breakthroughs are likely, this can be misleading, or a question of how we define a “breakthrough.” We are on the verge of a new digital revolution, the one in which many tasks, jobs, and even human relationships will be replaced by robots, drones, and other automated devices. Signal processing is and will continue to be the core of all such future devices, and it is likely that this robotization and automation will trigger unforeseen challenges that need

to be solved with new signal processing approaches and methodologies.

Whether signal processing will remain a tool, per se, or whether it will tend to integrate with social sciences and human psychology, this is a different question, and I answer positively: we already talk about in-body communications, where various nanometer-level devices can be implanted in the human body, and the human body (and mind) will act as the transmission channel, the catalyzer, or the receiver of various digitalized data. Digital signal processing will still play a crucial role in such a scenario, but can we talk about a “humanized” or “perception-tuned” signal processing? Another future trend is that, with the advent of global or ubiquitous “Internet of Everything,” privacy and security threats will increase many folds: a stalker from abroad could gain full video access to a remote home, or a remotely controlled drone (big enough) could even kidnap or injure a person. Again, signal processing mechanisms that are in use today for solving such privacy and security threats are probably not enough in tomorrow’s world, and new avenues of thinking need to emerge.

Simona Lohan (elena-simona.lohan@tut.fi) is an associate professor at the Tampere University of Technology, Finland.

Feng Liu



With the development of mathematics and computing techniques, the past few decades have shown that signal processing, more and more, has been considered as a relatively independent discipline, which studies the processing of waveforms or digits with mathematics. In turn, the signal processing discipline has also promoted the development of mathematics and computing techniques. As the computing techniques further develop, especially with graphic processing units, signal processing techniques such as deep neural networks become practical, which act more as “black boxes” but effectively achieve our complex goal. The future direction of signal processing

depends on the social need and the technical context.

Admittedly, mathematics plays today, and will play in the near future, a dominating and definitely positive role in the development of signal processing, with its results illuminating its path. However, it does not contradict with the notion that signal processing will advance with and inside the social need and technical context in the long run. Also, as to the old philosophy, a merited deed is influential but invisible like rain in the early spring: “sneaks into the tranquil night with the breeze and nurtures every spring life spontaneously and silently.”

Feng Liu (liuf@nankai.edu.cn) is a lecturer at Nankai University, Tianjin, China.

David A. Trejo Pizzo



Researchers have intensively investigated deep-learning algorithms for solving challenging problems in many areas such as image classification, speech recognition, signal processing, and natural language processing. In my job, I use deep-learning algorithms and signal processing to address new challenges in energy efficiency. Understanding the behavior of consumers and the trends in economy are essential for a country—with signal processing we can address this challenge better by taking noise out of market signals and huge databases with information about the energy consumption. Policy makers now will have a new tool that combines deep learning and signal processing to make better decisions. This means a breakthrough for engineers, where they become the next generation of policy makers by solving societal problems. Signal processing meets sociology by listening to the signals and erasing the outliers that are not clear with other tools. This joint venture of social scientists and engineers is a key to make smart cities a reality.

David A. Trejo Pizzo (dtrejopizzo@ieee.org) is a professor, Universidad del CEMA, Buenos Aires, Argentina.

Lav R. Varshney



I believe it is folly to doubt the creativity and hard work of signal processing researchers: there are surely new breakthroughs to come.

Whether due to new technologies leading to new research questions, internal knowledge shortcomings being filled, or new scientific phenomena requiring explanation, signal processing researchers will develop theories and techniques to address the challenges of the future.

Lav R. Varshney (varshney@illinois.edu) is an assistant professor at the University of Illinois at Urbana-Champaign.

Ajay V. Deshmukh



Signal processing, has seen much better times in its past and is continuing in the present. More and more, the future of signal processing is

looking very bright and long-lasting in time and in other dimensions—long-lasting because any information in nature and wherever it is available could be looked at as a signal. Signal genesis and its association with physical and other systems is also always going to be there. Therefore, one has to be optimistic about the future of signals as well as signal processing.

Although signal processing could be felt by many of us to be a mature topic, just because it has been offered as a semester-wide course many times across the globe, it is not the complete story. There are potential areas for novel advances in theory as well as practice. In fact, signal processing has brought mathematics into reality and connected it not only to physics but many other areas, for example, in engineering, medicine, biology, and agriculture. There are challenges, say, in the next 50 years, to provide signal processing solutions to industry problems. Some of the future key issues would be to put signals and systems together with other domains, nonlinear analysis would pick up, not only time but other dimensions

would be important, putting intelligence in systems, addressing diversified applications like safety, security, food, travel, water, and similar resources including those required in medicine, agriculture, automotive, astronomy and astrophysics, and so on. The context, would, however, change.

Ajay V. Deshmukh (ajay.deshmukh@rediffmail.com) is a principal at the Bajaj Institute of Technology, Wardha, India.

Volker Lohweg



Signal processing is, of course, in a mature state. Many concepts and theories are well established. However, I believe that signal processing has more to offer in the next ten to 20 years. Two aspects: sensors will be the key technology drivers and key enablers for many applications, but signal processing has to be on board because of data volume and necessary speed. These facts will definitely create new signal processing concepts.

Big data, which turn into smart data, have to cope with speed in the context of machine learning, classification, decision making, etc. Here we will also see new concepts and maybe new theories in the coming years.

Volker Lohweg (volker.lohweg@hs-owl.de) is a professor at the Institute of Industrial IT, Germany.

Abdelhak M. Zoubir



Signal processing will continue to grow. There are many challenges in real-life applications where signal processing is much needed. These areas include renewable energy, car engine monitoring, autonomous driving, synthetic aperture sonar, psychology, biomedical engineering, big data analytics, and synthetic biology, to mention a few. My view is that there is little room for theoretical breakthroughs in signal processing. This does not mean that there will not be emerging areas with great

potential for the advancement of knowledge in signal processing and solving the many unanswered theoretical questions, such as in adaptation and learning over sensor networks.

The future of signal processing and its success lies, in my view, in answering questions encountered in new application areas, for example, cultural heritage, and there is a trend to moving to new application areas outside the electrical engineering field. Indeed, many of my colleagues share this view with me, and we are putting much effort into solving new problems. In short, the future and potential of innovation in signal processing lies in interdisciplinary research (see [1]).

Abdelhak M. Zoubir (zoubir@ieee.org) is a professor at Technische Universität Darmstadt, Germany. He served as editor-in-chief of IEEE Signal Processing Magazine from 2012 to 2014.

Kush R. Varshney



Signal processing theory and methods are increasingly applicable to data science applications outside in society. With this emergence in socio-technical (not simply technical) solutions, new problems begging for research breakthroughs are popping up. We have started investigating some of these problems through the IBM Science for Social Good initiative, but we are just at the tip of the iceberg.

Kush R. Varshney (kryvarshn@us.ibm.com) is a member of research staff and the manager of Data Science Theory and Algorithms at the IBM Thomas J. Watson Research Center, Yorktown Heights, New York.

Muhammad Zubair Ahmad



Signal processing has historically been associated with the communication systems and one-dimensional signals. The rapidly developing field of machine learning has been perceived as the alternative to signal processing. Evidence of this is the

discussion held at Technion in 2014, under the title “Is Deep Learning the Final Frontier and the End of Signal Processing?” The computer science community and industry has successfully marketed machine learning as an alternative to signal processing. A major factor contributing to this is the fear of signal processing among undergraduates, which I have found common across countries. The research community at large has learned to fear signal processing and has been told that machine learning and data processing is a skill that can be acquired with minimal training.

Thus, I would call for a rebranding of signal processing as a paradigm fundamental to industrial applications and the modeling of physical reality. Anything ranging from the individual systems of the human body to the machines and automata developed by the humans can be modeled using this framework. Different modeling strategies based on statistical learning theory, analysis, differential geometry, topology, and group theory must be generalized under this universal framework.

The signal processing community must come up with a method of teaching that is alluring to young students. We need to convince young (as well as old) minds that the modern trend of sacrificing understanding for practicality may be economically good but is directly conflicting with the driving principles of the scientific community.

Muhammad Zubair Ahmad (ahmadmz@myumanitoba.ca) is a Ph.D. degree candidate at the University of Manitoba, Canada.

Thomas A. Baran



This question reminds me of a favorite saying of my Ph.D. advisor, Prof. Alan Oppenheim: “There will always be signals, they will always need processing, and there will always be new applications, new mathematics, and new implementation technologies.”

In October, we held a symposium titled “The Future of Signal Processing” at the

(continued on page 25)

John Edwards

Medical Optical Imaging

Signal processing leads to new methods of detecting life-threatening situations

Optical imaging, a medical technique that's used to obtain detailed images of organs, tissues, cells, and molecules in the presence of visible light, is an emerging technology with the potential to enhance patient treatment, diagnosis, and disease prevention.

Offering numerous advantages over radiological imaging techniques, optical imaging uses nonionizing radiation to reduce a patient's radiation exposure, thereby allowing for more frequent studies over time. Optical imaging also has the ability to differentiate suspicious soft tissues from native soft tissues as well as tissues labeled with either exogenous or endogenous contrast media. Scattering differences and photon absorption provide specific tissue contrasts and potential capabilities for studying functional and molecular level activities. Optical imaging is a multimodal and highly responsive imaging technique that can be easily combined with other imaging approaches to create complete multidimensional views of objects and areas of interest.

Signal processing is now helping to make optical imaging even more useful and versatile, allowing more detailed images to be captured and expanding the technology's use into new patient treatment and medical research areas.

Identifying cancer biomarkers

At the University of Arizona, Prof. Jennifer Barton is using advanced optical

imaging to identify imaging biomarkers of ovarian cancer, one of the most deadly gynecological cancers. The project's goal is to extend lives while preserving quality of life. "Right now, there is no effective ovarian cancer screening technology that is useful for all women," says Barton, a professor of biomedical engineering and interim director of the BIO5 Institute.

When detected early, ovarian cancer can often be treated effectively with surgery and chemotherapy. Yet, given the lack of good tools for catching it at its early stages, fewer than half of women diagnosed survive five years.

Barton is collaborating with researchers in the university's departments of physiology, medical imaging, and obstetrics and gynecology to identify imaging biomarkers, subtle changes in tissue that can be detected by sensitive optical methods, for ovarian cancer in mice. Barton's team has developed a tiny, highly flexible falloposcope—a wand-like imaging device that uses high-resolution optical imaging techniques—to obtain in vivo images of ovaries and fallopian tubes (Figure 1). By analyzing physical and biochemical changes over time to create a road map of the changes that happen during ovarian cancer, the researchers hope to be able to detect cancer in the fallopian tubes, where many researchers believe it originates.

"My optical imaging work utilizes two modalities: optical coherence tomography (OCT) and multispectral fluorescence imaging (MFI)," Barton says. OCT measures the interference of

broadband light from a reference mirror with light from the tissue. "The reflectivity of tissue as a function of depth is then encoded in the interference frequency, where that interference is measured as a function of wavelength," she notes. "In my current setup, we use a spectrometer detector and measure the spatial frequency modulations on a linear charge-coupled device (CCD) array."

Resampling is necessary to convert from a measured function of wavelength to a function of wavenumber. Then a Fourier transform is performed to obtain the reflectivity as a function of depth. "Each measurement off the linear CCD array provides one depth scan, or column of a cross-sectional image," Barton says. "We have to scan the beam in one or two dimensions to obtain a 2-D or 3-D image."

For MFI, the researchers have to control which excitation wavelength is used to illuminate the tissue. "In our falloposcope, lasers are coupled into a single high numerical aperture multimode fiber that directs light to the tissue," Barton says. An imaging fiber bundle collects the reflected or fluorescence light, which is measured with a high-sensitivity CCD camera. A filter wheel in front of the camera selects out reflected light or fluorescence light at a specific wavelength range. "We need to adjust gain and exposure time of the CCD as the signals in reflected light are orders of magnitude higher than fluorescence light," she notes.

Frame grabbers are typically used for the OCT linear CCD array, the MFI

Digital Object Identifier 10.1109/MSP.2017.2743118
Date of publication: 13 November 2017



UNIVERSITY OF ARIZONA

FIGURE 1. University of Arizona Prof. Jennifer Barton holding a highly flexible falloscope her research team has developed to image the biomarkers of ovarian cancer, one of the most deadly gynecological cancers.

CCD camera, and a multipurpose data acquisition (DAQ) board to generate control signals for scanning or excitation of a source/filter wheel and to generate any needed synchronization signals. “It is always a struggle to increase signal-to-noise, dynamic range, and contrast in imaging systems,” Barton says. These attributes affect how fast—and deep—one can image—in OCT. “We are limited in the amount of light power we can put on the tissue, so signal processing techniques that efficiently extract the signal from noise, background, and unwanted artifacts are always important,” she explains.

“In the past, systems were slow enough that one didn’t have to pay too much attention to data acquisition and signal processing,” Barton observes. “Nowadays, with linear CCDs running at 100-KHz frame rates and 2k pixels, there needs to be more careful consideration of hardware and software processing,” she notes. “This is not extraordinary as compared to some signal processing applications, but it means that imaging teams have to have new skill sets.”

The team is now seeking additional funding to build hospital-ready falloscopes so that research can be conducted on human subjects. Barton is hopeful that the technology will lead to an earlier and more accurate diagnosis of ovarian

cancer. “Our technique can either serve as a primary screening method, or as a follow-up to other tests,” she says.

Imaging arteries

Plaque accumulating inside artery walls can cause arteries to thicken and harden. When a plaque accumulation ruptures, it can restrict or even block blood flow, leading to a heart attack, stroke, or other serious medical issues. Accurate diagnoses are limited by the fact that there are no imaging tools available to consistently and accurately detect plaque at risk of rupturing in living patients.

An enhanced imaging technology—intravascular photoacoustic (IVPA) imaging—can generate three-dimensional images of artery interiors, potentially helping physicians to diagnose plaques on the verge of rupturing. The drawback is that developers have so far struggled to develop imaging instruments that are capable of illuminating arteries to a useful depth and at fast enough speeds while also meeting clinical requirements.

Now, using signal processing and other advanced tools and approaches, a team of researchers from Purdue University, the Indiana University School of Medicine, and the Shanghai Institute of Optics and Fine Mechanics has developed a new type of collinear

catheter (Figure 2), featuring a design that promises to greatly improve the sensitivity and imaging depth of IVPA imaging.

“Our photoacoustic catheter probe integrates both photoacoustic and ultrasound modalities within a very tiny space—1 mm in diameter in our most updated version,” says Yingchun Cao, a postdoctoral fellow working in the research group led by Prof. Ji-Xin Cheng of Purdue University. “The most important feature of our catheter is that we used a collinear design for the optical-acoustic wave overlap to greatly improve the imaging sensitivity and depth,” notes Cao, who is the lead author of a research paper on the project.

IVPA imaging functions by measuring ultrasound signals from molecules exposed to a light beam from a fast-pulsing laser. The new collinear probe allows the optical beam and sound wave to share the same path throughout the imaging process, rather than cross-overlapping as in previous designs. The approach increases the instrument’s sensitivity as well as the imaging depth, enabling high-quality IVPA imaging of a human coronary artery over 6 mm in depth from the lumen, the normally open channel within arteries, to the perivascular fat that surrounds the outside of most arteries and veins. “This research can be used ... to help the doctor for accurate diagnosis of plaque vulnerability and even for imaging-guided intravascular surgery or drug delivery,” Cao says.

“The unique advantage of our research,” Cao says, “is we can provide quantitative information of lipid deposit within the artery wall, including the size and depth of lipid core with sufficient spatial resolution. The coregistered ultrasound image integrated in the technique can provide morphological structure of the artery for accurate position identification of lipid deposit. “In our most recent research, we can accurately distinguish different lipid compositions by a self-developed numerical approach,” Cao notes.

Signal processing is an important part of the research. “A high-quality real-time

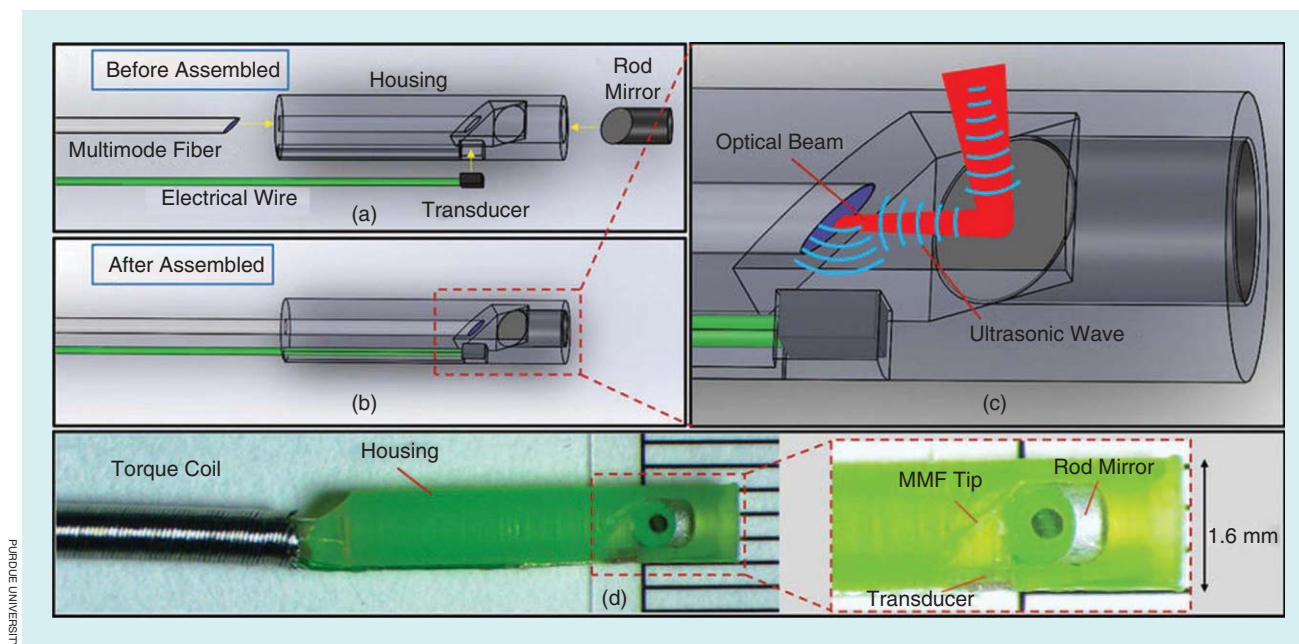


FIGURE 2. A new catheter probe, developed by researchers at Purdue University, the Indiana University School of Medicine, and the Shanghai Institute of Optics and Fine Mechanics, can generate three-dimensional images of artery interiors, potentially helping physicians to diagnose plaque on the verge of rupturing. (a) The main components of the collinear catheter before assembly. (b) The assembled catheter probe. (c) A zoomed-in view of the catheter tip shows the collinear overlap between optical and ultrasonic waves. (d) The fabricated 1.6-mm catheter probe and the detailed structure of the catheter tip (inset).

displayed image at video-rate or quasi-video-rate speed requires a number of advanced signal processing techniques, including noise shielding,” Cao says. “In our current system we use a preamplification device to boost the signal, data sectioning to select the effective data we need, a programmable sampling rate to reduce the data amount, a median filter to remove the random noise speckle and a bandpass filter to remove other noise,” Cao says. The team also uses a Hilbert transform to obtain amplitude information, polar coordinate projection for fast coordinate transformation, logarithmic compression and Tagged Image File Format (TIFF) imaging compression to save storage space.

The biggest signal processing-related challenges facing the researchers are enabling effective noise filtering, fast image display, and saving image data to a hard disk, if necessary. “Our imaging system can work at a high frame rate, say, 16 frames per second,” Cao says. “That means in every second a huge amount of data will be generated and saved to a computer.”

The biggest overall technical challenge is the contradiction between catheter size

and sensitivity. The current diameter of 1 mm is for the bare catheter without a protective sheath. After integrating the sheath, the diameter is around 1.6 mm, which is slightly large for a coronary application. The team is now working to shrink the diameter of the catheter, including a sheath, down to ~1 mm to meet the clinical requirement. “The further decrease of the catheter size will result in both apparent photoacoustic and ultrasound loss, because both of these waves are reflected by a micro-mirror imbedded in the catheter,” Cao says. Another challenge facing the researchers is the optical wave scattering that occurs when the signal travels through blood, which greatly reduces light intensity and photoacoustic sensitivity during in vivo applications.

“I believe this technology is very promising for future clinical diagnosis of human coronary artery disease,” Cao says, noting that the research is still at a very early stage. “But we are confident to overcome these technical problems

and, hopefully, it can go to clinic in the next few years,” he adds.

Peering inside cells

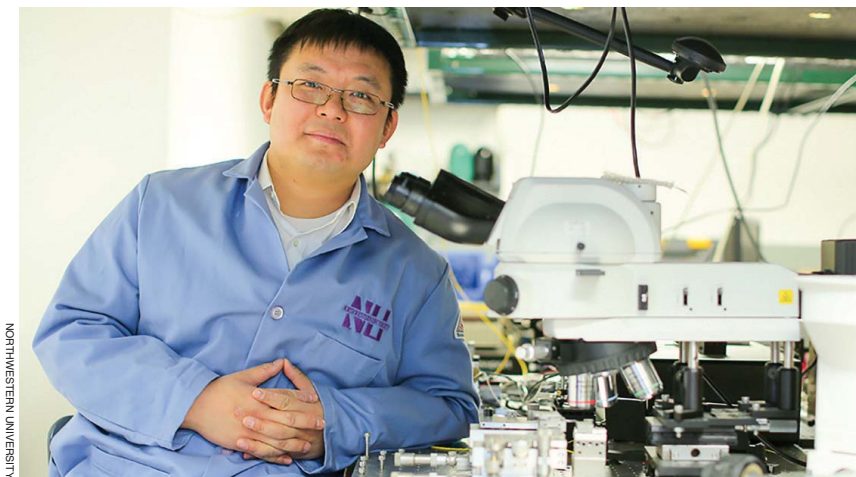
Building on research that won an international team the 2014 Nobel Prize in Chemistry, Northwestern University engineers say they have developed an improved version of a superresolution fluorescence microscopy technique that

makes it possible to study complex molecular processes in cells.

The new optical imaging technology—spectroscopic photon localization microscopy (SPLM)—is simpler and less expensive than its two predecessors while also offering four times

that resolution, claim the researchers. Like the earlier technologies, SPLM is designed to control how fluorescence molecules emit, ensuring that no spatially adjacent molecules emit simultaneously. As a result, each random fluorescence emission can be considered to very likely to come from a single molecule. Based

The biggest signal processing-related challenges facing the researchers are enabling effective noise filtering, fast image display, and saving image data to a hard disk, if necessary.



NORTHWESTERN UNIVERSITY

FIGURE 3. Northwestern University Prof. Hao Zhang says his research team has developed an improved version of a superresolution fluorescence microscopy technique that's used to study molecular processes in living cells.

on this assumption, only the centers of the detected individual molecular emissions are extracted and stored in each acquisition. This process is then repeated thousands of times to accumulate all the “emission centers” into a final image. The spatial resolution is proportional to the number of photons in each emission.

SPLM brings, for the first time, spectroscopic analysis to photon localization microscopy, allowing researchers to image multiple molecular labels simultaneously. The emission spectra of these molecular labels do not have to be significantly different and, in fact, can be largely overlapping. SPLM can analyze the full profile of each emission spectrum to distinguish molecular labels, numerically improving spatial resolution by combining photons from different imaging frames. Additionally, SPLM allows the imaging of multiple molecules, such as DNA, using their intrinsic fluorescence emissions.

“Our contribution to this technology is that we add an additional spectroscopic imaging capability to photon localization,” says Hao F. Zhang, professor of biomedical engineering in Northwestern’s McCormick School of Engineering (Figure 3). “The earlier technologies cannot distinguish wavelength differences from those emissions.”

To solve this deficiency, the researchers needed to design molecular labels with desired, separated emission spectra and use optical filters to separate photon

emissions with different wavelengths. One technical constraint the team faced is the limited number of filters that can be incorporated into a single system, which restricts the number of molecules that can be simultaneously imaged. Additionally, spatial resolution cannot be further improved once a particular molecular label has been selected. “We added a specially designed optical grating to the detecting optical path so that both the intensity of emitted photons and their associated optical spectra are detected at the same time,” Zhang says.

Optical grating is an optical dispersive component. “When light passes through or is reflected by an optical grating, two beams will be generated simultaneously due to multiple interference,” Zhang says. One beam, referred to as the *zeroth-order diffraction beam*, discloses the incident beam’s intensity. The second beam, referred to as the *first-order diffraction beam*, reveals the optical spectrum. “We detect both the zeroth- and first-order beams using the same high-sensitivity array detector to obtain the molecular location and its emission spectrum simultaneously,” Zhang explains. “Because no optical filter is used in the detection, and the complete profile of photon emission is detected, the number of molecular contrasts is, in principle, unlimited.” Additionally, based on the individual emission spectrum, the system can combine imaging frames to numerically increase the number of

photons for localization, which improves spatial resolution.

Signal processing plays a critical role in building SPLM’s high-quality images. “For example, during the photon localization process, we need to find the best way to fit the point spread function of each photon emission,” Zhang says. “To make individual emissions recognizable among different camera frames, we needed to design pattern recognition algorithms to identify optical spectral features among thousands of frames and determine whether they are from the same single molecule or not.”

SPLM uses a high-sensitivity array photon detector to capture two images at different regions of the array detection simultaneously. The photo detector has an internal amplification capability. Each single frame acquisition takes about 10 ms, and several thousand such single frames may be generated during a single session. “Once all these frames are acquired and stored, we identify the center of each single-molecular emission and its associate optical spectrum,” Zhang says. “We applied sophisticated signal conditioning operations to, for example, reduce background noise and remove detector dark current.” Gaussian fitting provides the emission center and associated optical spectrum.

Zhang says the signal processing used in SPLM is far more comprehensive than what’s currently available in the field. “The most unique part is that our method takes advantage of the optical spectrum of all single-molecular emissions, besides their locations, into consideration,” he says. “As a result, we are not constrained by the limited number of detection channels and pseudo coloring; we know the full spectra of all the detected molecules. There are no alternative approaches because “the optical spectra information is undetectable otherwise,” he adds. Zhang says the team is developing an open-source image processing package for SPLM.

Author

John Edwards (jedwards@johnedwardsmedia.com) is a technology writer based in the Phoenix, Arizona, area.

SP

Top Downloads in IEEE Xplore

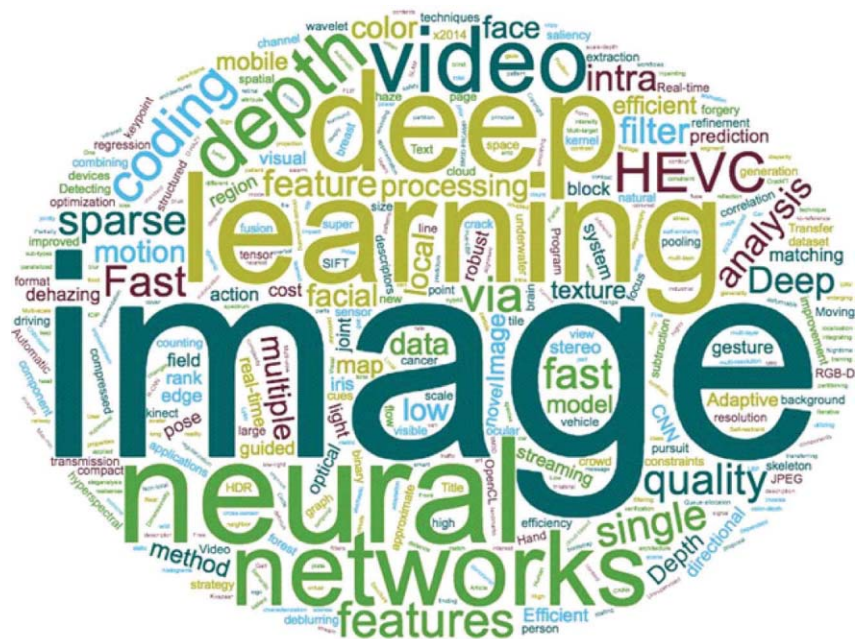
Each "Reader's Choice" column focuses on a different publication of the IEEE Signal Processing Society (SPS). This month we are highlighting articles featured at the IEEE International Conference on Image Processing (ICIP).

The ICIP, sponsored by the IEEE SPS, is the premier forum for the presentation of technological advances and research results in the fields of theoretical, experimental, and applied image and video processing.

This issue's "Reader's Choice" column lists the top 15 articles from ICIP 2014, ICIP 2015, and ICIP 2016, indicated by the year listed after the abstract, that were the most downloaded from January 2015 to June 2017. Please send your suggestions and comments on this column to Associate Editor Changshui Zhang (zcs@mail.tsinghua.edu.cn).

Hand Gesture Recognition with Leap Motion and Kinect Devices

Marin, G.; Dominio, F.; Zanuttigh, P.
This paper proposes a novel hand gesture recognition scheme explicitly targeted to leap motion data. An adhoc feature set based on the positions and orientation of the fingertips is fed into a multiclass support vector machine classifier to recognize the performed gestures. A set of features is also ex-



tracted from the depth computed from the Kinect and combined with the leap motion ones to improve the recognition performance.

2014

Local Binary Pattern Network: A Deep Learning Approach for Face Recognition

Xi, M.; Chen, L.; Polajnar, D.; Tong, W.

In this paper, a novel unsupervised deep-learning-based methodology, named *local binary pattern network (LBPNet)*, is proposed to efficiently extract and compare high-level overcomplete features in a multilayer

hierarchy. The LBPNet retains the same topology of the convolutional neural network, whereas the trainable kernels are replaced by the off-the-shelf computer vision descriptor (i.e., LBP).

2016

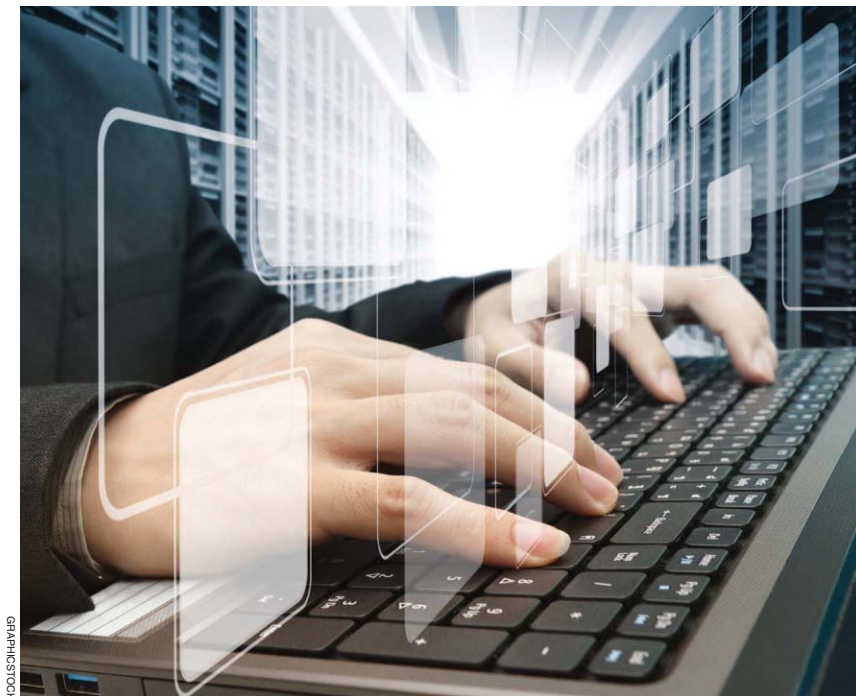
Road Crack Detection Using Deep Convolutional Neural Network

Zhang, L.; Yang, F.; Zhang, Y.D.; Zhu, Y.J.

A deep-learning-based method for crack detection is proposed in this paper. A supervised deep convolutional neural network is trained to classify each image patch in the collected images.

2016

Digital Object Identifier 10.1109/MSP.2017.2743111
Date of publication: 13 November 2017



GRAPHICSTOCK

A Deep Neural Network for Image Quality Assessment

Bosse, S.; Maniry, D.; Wiegand, T.; Samek, W.

This paper presents a no-reference image quality assessment method based on a deep convolutional neural network (CNN). The CNN takes un-preprocessed image patches as an input and estimates the quality without employing any domain knowledge. By that, features and natural scene statistics are learned purely data driven and combined with pooling and regression in one framework.

2016

Fast Multidimensional Image Processing with OpenCL

Oliveira Dantas, D.; Danilo Passos Leal, H.; Oliveira Barros Sousa, D.

VisionGL is an open-source library that provides a set of image processing functions and can help the programmer by automatically generating code. The objective of this work is to augment VisionGL by adding multidimensional image processing support with OpenCL for high performance through use of graphic processing units.

2016

Dimensionality Reduction of Brain Imaging Data Using Graph Signal Processing

Rui, L.; Nejati, H.; Cheung, N.-M.

This paper presents a new dimensionality reduction method based on the recent graph signal processing theory for the task of classifying the brain imaging signals recording the cortical activities in response to visual stimuli. Authors propose using the resting-state measurements (i.e., before onset of the stimulus) of the subjects to build a connectivity graph. The graph Laplacian and graph-based filtering are then applied to learn the low-dimensional linear subspace for the task-state measurements (i.e., after onset of the stimulus).

2016

Moving Object Segmentation Using Depth and Optical Flow in Car Driving Sequences

Kao, J.-Y.; Tian, D.; Mansour, H.; Vetro, A.; Ortega, A.

In this paper, based on an analysis of motion vanishing points of the scene and estimated depth, a geometric model that relates extracted two-dimensional (2-D) motion to a three-dimensional (3-D) motion field relative to the camera is

derived. A constrained optimization problem that considers group sparsity is formulated to recover the 3-D motion field from the 2-D motion. The recovered 3-D motion field is then clustered to provide the segmentation of moving objects.

2016

ORB-SLAM Map Initialization Improvement Using Depth

Fujimoto, S.; Hu, Z.; Chapuis, R.; Aufrère, R.

Map initialization and scale ambiguity are well-known challenging problems for visual simultaneous localization and mapping. In this paper, a triangulation is used on red, green, and blue feature points for getting three-dimensional points from out of the limited area in depth. The authors combined both advantages of triangulation and depth to improve the performance of robustness to initialization and tracking.

2016

Deep Learning Network for Blind Image Quality Assessment

Gu, K.; Zhai, G.; Yang, X.; Zhang, W.

The authors in this paper introduce a new deep-learning-based image quality index (DIQI) for blind quality assessment. Extensive studies are conducted on the new TID2013 database and confirm the effectiveness of their DIQI relative to classical full-reference and state-of-the-art reduced- and no-reference IQA approaches.

2014

Depth Augmented Stereo Panorama for Cinematic Virtual Reality with Focus Cues

Thatte, J.; Boin, J.-B.; Lakshman, H.; Wetzstein, G.; Girod, B.

Cinematic virtual reality aims to provide immersive visual experiences of real-world scenes on head-mounted displays. The authors propose a new content representation, depth augmented stereo panorama, which permits generating light fields across the observer's pupils, achieving an order of magnitude reduction in data requirements compared to the existing techniques.

2016

ICIP 2016 Competition on Mobile Ocular Biometric Recognition

Rattani, A.; Derakhshani, R.; Saripalle, S.K.; Gottemukkula, V.

The aim of this competition is to evaluate and compare the performance of mobile ocular biometric recognition schemes in visible light on a large scale database (VISOB Data Set ICIP 2016 Challenge Version) using standard evaluation methods. Four different teams from universities across the world participated in this competition, submitting five algorithms altogether. The best results were obtained by a team from Norwegian Biometrics Laboratory (NTNU, Norway).

2016

Semantic Context and Depth-Aware Object Proposal Generation

Zhang, H.; He, X.; Porikli, F.; Kneip, L.

This paper presents a context-aware object proposal generation method for stereo images. The authors propose to incorporate additional geometric and

high-level semantic context information into the proposal generation.

2016

Super-Resolution of Compressed Videos Using Convolutional Neural Networks

Kappeler, A.; Yoo, S.; Dai, Q.; Katsaggelos, A.K.

In this paper, for the problem of compressed video superresolution, the authors propose a CNN that is trained on both the spatial and the temporal dimensions of compressed videos to enhance their spatial resolution. Consecutive frames are motion compensated and used as input to a CNN that provides superresolved video frames as output.

2016

Classification of Hyperspectral Image Based on Deep Belief Networks

Li, T.; Zhang, J.; Zhang, Y.

In this paper, deep-learning frameworks, the restricted Boltzmann machine

model, and its deep structure deep belief networks are introduced in hyperspectral image processing as the feature extraction and classification approach.

2014

Image Character Recognition Using Deep Convolutional Neural Network Learned from Different Languages

Bai, J.; Chen, Z.; Feng, B.; Xu, B.

This paper proposes a shared-hidden-layer deep convolutional neural network (SHL-CNN) for image character recognition. In SHL-CNN, the hidden layers are made common across characters from different languages, performing a universal feature extraction process that aims at learning common character traits existing in different languages, such as strokes, while the final softmax layer is made language dependent, trained based on characters from the destination language only.

2014



PANEL AND FORUM (continued from page 13)

Steve Young (sjy@eng.cam.ac.uk) is a professor of information engineering at the University of Cambridge, United Kingdom, and a senior member of technical staff at Apple. His main research interests lie in the area of statistical spoken language systems, including speech recognition, speech synthesis, and dialog management. He is the recipient of a number of awards including an IEEE Signal Processing Society Technical Achievement Award and the IEEE James L. Flanagan Speech and Audio Processing Award. He is a Fellow of the IEEE and the U.K. Royal Academy of Engineering. In addition to his academic career, he has also founded a number of successful start-ups in the speech technology area.

References

- [1] L. Deng and D. Yu, "Deep learning: Methods and applications," in *Foundations and Trends in Signal Processing Series*. Boston, MA: NOW Publishers, 2014.
- [2] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. Cambridge, MA: MIT Press, 2016.
- [3] L. Deng and X. Li, "Machine learning paradigms for speech recognition: An overview," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 21, pp. 1060–1089, May 2013.
- [4] D. I. Shuman, S. K. Narang, P. Frossard, A. Ortega, and P. Vandergheynst, "The emerging field of signal processing on graphs: Extending high dimensional data analysis to networks and other irregular domains," *IEEE Signal Process. Mag.*, vol. 30, pp. 83–98, May 2013.
- [5] A. Sandryhaila and J. M. F. Moura, "Big data processing with signal processing on graphs," *IEEE Signal Process. Mag.*, vol. 31, no. 5, pp. 80–90, Sept. 2014.
- [6] S. Chen, R. Varma, A. Sandryhaila, and J. Kovačević, "Discrete signal processing on graphs: Sampling theory," *IEEE Trans. Signal Process.*, vol. 63, no. 24, pp. 6510–6523, Dec. 2015.
- [7] F.-L. Luo and C. Zhang, Eds. *Signal Processing for 5G: Algorithms and Implementations*. Chichester, U.K.: Wiley-IEEE Press, 2016.
- [8] R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, and P. Kuksa, "Natural language processing (almost) from scratch," *J. Mach. Learn. Res.*, vol. 12, pp. 2493–2537, Aug. 2011.
- [9] L. Deng, "Industrial technology advances: Deep learning: From speech recognition to language and multimodal processing," *APSIPA Trans. Signal Inform. Process.*, vol. 5, pp. 1–15, Jan. 2016.
- [10] N. Mrksic, I. Vulic, D. O. Seaghdha, I. Leviant, R. Reichart, M. Gasic, A. Korhonen, and S. Young, "Semantic specialization of distributional word vector spaces using monolingual and cross-lingual constraints," *Trans. Assoc. Computat. Linguistics*, vol. 5, pp. 309–324, Sept. 2017.
- [11] J. A. Fessler. (2017). Medical image reconstruction: A brief overview of past milestones and future directions. [Online]. Available: <http://arxiv.org/abs/1707.05927>
- [12] Y. Liu, J. Chen, and L. Deng, "Unsupervised sequence classification using sequential output statistics," in *Proc. NIPS*, Dec. 2017.
- [13] G. Hinton, et al., "Deep neural networks for acoustic modeling in speech recognition," *IEEE Signal Process. Mag.*, vol. 29, no. 6, pp. 82–97, Nov. 2012.



FROM THE GUEST EDITORS

Fatih Porikli, Shiguang Shan, Cees Snoek,
Rahul Sukthankar, and Xiaogang Wang

Deep Learning for Visual Understanding

In the past decade, there has been a transformative and permanent revolution in computer vision cultivated by the rein-vigorated adoption of deep learning for visual understanding tasks. Driven by the increasing availability of large annotated data sets, efficient training techniques, and faster computational platforms, deep-learning-based solutions have been progressively employed in a broader spectrum of applications from image classification to activity recognition.

Deep learning, in general, refers to a range of artificial neural networks that consist of multiple layers, mimicking the structure and cognitive process of the human brain. Instead of relying on hand-crafted features, they allow the acquisition of knowledge directly from data. They regress intricate objective functions in a nested hierarchy, where more sophisticated representations with larger receptive fields computed in terms of less abstract ones with localized supports. Deep learning also makes it possible to incorporate explicit domain knowledge and replace a large variety of conventional algorithmic blocks with trainable differentiable modules. These all give deep learning an exceptional power and flexibility in modeling the relationship between the input data and target output.

Efforts are now shifting toward the remaining challenges. For instance, the majority of current methods have

been designed to solve supervised learning problems where data comes with its labeled attributes and how to reliably apply deep learning to unsupervised settings in a similar degree of success is an active area of research. Similarly, recent efforts aim at working with small data, focusing on how to take advantage of large quantities of unlabeled examples as well as with a few labeled samples.

Another area where deep agents may play a significant role is to integrate positive and negative rewards into deep learning to choose the actions that yield the best cumulative reward by interacting with the environment. Also, the fusion of multimodal and structured data into existing deep-learning models would open up more extended application domains.

This special issue of *IEEE Signal Processing Magazine (SPM)* is therefore devoted to providing survey articles on the latest advances in deep learning for visual understanding. Its objective is to encourage a diverse audience of researchers and enthusiasts toward an effective participation in the solution of analogous problems in other signal processing fields by inseminating similar ideas.

The range of articles in this two-part special issue indicates the breadth of the computer vision discipline. (Part two will be published in January 2018.) Many fundamental areas are surveyed from the computer vision perspective, including

- reinforcement learning
- learning with limited and no supervision (unsupervised learning)

- weakly supervised learning
 - zero- and few-shot learning
 - domain adaptation
 - multimodal learning
 - metric learning
 - generative adversarial networks
 - recurrent networks
 - regression with Bayesian networks
 - model compression and robustness.
- In addition, in-depth overviews of several deep-learning-based computer vision applications are provided, including
- inverse problems such as superresolution and image enhancement
 - picture quality prediction
 - saliency detection
 - image and video segmentation with conditional random fields
 - image-to-text generation
 - visual question answering
 - face image analytics.

We would like to wholeheartedly thank all of the contributing authors and reviewers of this special issue. We also sincerely appreciate *SPM*'s editor-in-chief, Prof. Min Wu, Managing Editor Jessica Welsh, and the entire magazine's editorial staff for their extremely valuable support.

Meet the guest editors



Fatih Porikli (fatih.porikli@anu.edu.au) received his B.Sc. degree in electrical engineering from Bilkent University, Turkey, in 1992 and his Ph.D. degree in electrical and computer engineering from

New York University in 2002. He is an IEEE Fellow and a professor at Australian National University. He is also the chief scientist at Huawei, Santa Clara, California. Previously, he served as the Computer Vision Research group leader at National ICT Australia and distinguished scientist at Mitsubishi Electric Research Laboratories. His research interests include computer vision and machine learning with commercial applications in autonomous vehicles, video surveillance, visual inspection, robotics, and medical systems. He received the R&D100 Scientist of the Year Award in 2006, won five Best Paper Awards at IEEE conferences, and invented 71 patents.



Shiguang Shan (sgshan@ict.ac.cn) received his B.S.E. and M.S.E. degrees in computer science from Harbin Institute of

Technology, China, in 1997 and 1999, respectively. He received his Ph.D. degree in computer science from the Institute of Computing Technology, Chinese Academy of Sciences (CAS), Beijing, in 2004, where he has been a full professor since 2010 and is now the deputy director of the CAS Key Lab of Intelligent Information Processing. His research interests include computer vision, pattern recognition, and machine learning. He has published more than 200 papers in these areas. He served as area chair for many international conferences and is an associate editor of several

journals, including *IEEE Transactions on Image Processing*, *Computer Vision and Image Understanding*, *Neurocomputing*, and *Pattern Recognition Letters*.



Cees Snoek (cgmsnoek@uva.nl) received the M.Sc. degree in business information systems in 2000 and the Ph.D. degree in computer science in 2005, both from the University of Amsterdam, The Netherlands.

He is currently a director of the QUVA Lab, the joint research lab of Qualcomm and the University of Amsterdam, on deep learning and computer vision. He is also a principal engineer/manager at Qualcomm and an associate professor at the University of Amsterdam. His research interests focus on video and image recognition. He has published more than 200 refereed book chapters, journal, and conference papers. He received a Veni Talent Award, a Fulbright Junior Scholarship, a Vidi Talent Award, and The Netherlands Prize for Computer Science Research, all for research excellence.



Rahul Sukthankar (rahulsukthankar@gmail.com) received his B.S.E. degree in computer science from Princeton University, New Jersey, in 1991 and his Ph.D. degree in robotics from Carnegie Mellon, Pitts-

burgh, Pennsylvania, in 1997. He leads

research efforts in computer vision, machine learning, and robotics at Google. He is also an adjunct research professor with the Robotics Institute at Carnegie Mellon and courtesy faculty at the University of Central Florida. Previously, he was a senior principal researcher at Intel Labs, a senior researcher at HP/Compaq Labs, and a research scientist at Just Research. He has organized several workshops and conferences and currently serves as the editor-in-chief of *Machine Vision and Applications*.



Xiaogang Wang (xgwang@ee.cuhk.edu.hk) received his bachelor's degree in electronic engineering and information science from the Special Class of Gifted Young at the University of Science and Technology of China in 2001, his M.Phil. degree in information engineering from the Chinese University of Hong Kong in 2004, and his Ph.D. degree in computer science from the Massachusetts Institute of Technology in 2009. He has been an associate professor in the Department of Electronic Engineering at the Chinese University of Hong Kong since August 2009. He received the PAMI Young Research Award Honorable Mention in 2016. He is the associate editor of *Image and Visual Computing Journal*, *Computer Vision and Image Understanding*, and *IEEE Transactions on Circuit Systems and Video Technology*.

SP

Community Voices (continued from page 16)

Massachusetts Institute of Technology in honor of Al Oppenheim's 80th birthday, with the goal of bringing together experts in industry and academia to think progressively and speculate about the future of the field moving forward.

Over a dozen speakers provided a range of thought-provoking insights about the continued impact of the field in the decades ahead, in terms of applica-

tions, mathematics for new algorithms, and new implementation technologies. We would love to share this with those in the signal processing community who were unable to attend. A collection of video recordings and thoughts from the symposium will be available at <https://futureofsp.eecs.mit.edu/>. It was an exciting event, and we hope that the videos continue to stimulate further creative discussion within the community!

Thomas A. Baran (tom.baran@gmail.com) is a cofounder and chief executive officer of Lumii and a research affiliate at the Massachusetts Institute of Technology.

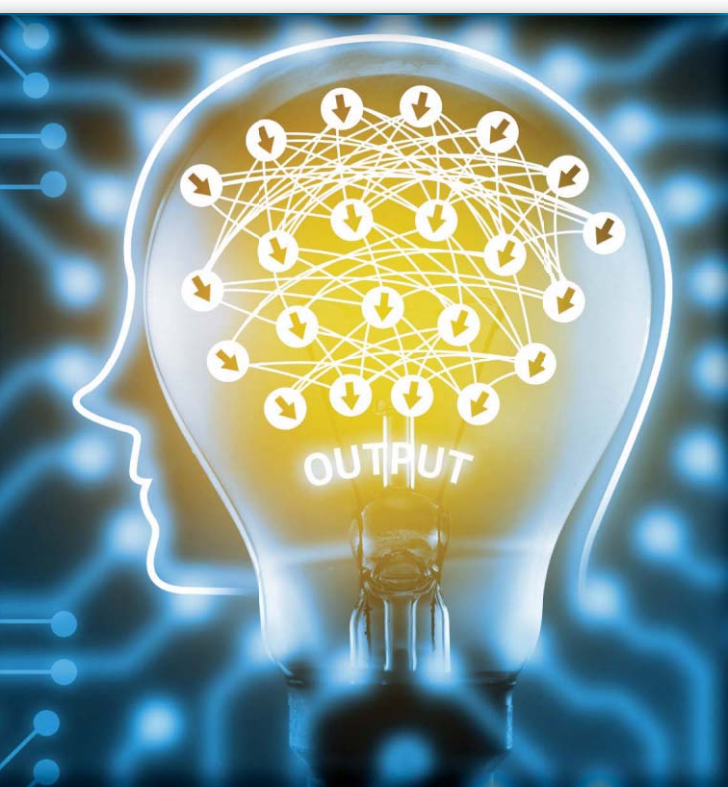
Reference

[1] A. Zoubir, "Interdisciplinary research: A catalyst for innovation" [From the Editor], *IEEE Signal Process. Mag.*, vol. 29, no. 3, pp. 2–4, 2012. SP

Kai Arulkumaran, Marc Peter Deisenroth,
Miles Brundage, and Anil Anthony Bharath

Deep Reinforcement Learning

A brief survey



©ISTOCKPHOTO.COM/ZAPP2PHOTO

Deep reinforcement learning (DRL) is poised to revolutionize the field of artificial intelligence (AI) and represents a step toward building autonomous systems with a higher-level understanding of the visual world. Currently, deep learning is enabling reinforcement learning (RL) to scale to problems that were previously intractable, such as learning to play video games directly from pixels. DRL algorithms are also applied to robotics, allowing control policies for robots to be learned directly from camera inputs in the real world. In this survey, we begin with an introduction to the general field of RL, then progress to the main streams of value-based and policy-based methods. Our survey will cover central algorithms in deep RL, including the deep Q -network (DQN), trust region policy optimization (TRPO), and asynchronous advantage actor critic. In parallel, we highlight the unique advantages of deep neural networks, focusing on visual understanding via RL. To conclude, we describe several current areas of research within the field.

Introduction

One of the primary goals of the field of AI is to produce fully autonomous agents that interact with their environments to learn optimal behaviors, improving over time through trial and error. Crafting AI systems that are responsive and can effectively learn has been a long-standing challenge, ranging from robots, which can sense and react to the world around them, to purely software-based agents, which can interact with natural language and multimedia. A principled mathematical framework for experience-driven autonomous learning is RL [78]. Although RL had some successes in the past [31], [53], [74], [81], previous approaches lacked scalability and were inherently limited to fairly low-dimensional problems. These limitations exist because RL algorithms share the same complexity issues as other algorithms: memory complexity, computational complexity, and, in the case of machine-learning algorithms, sample complexity [76]. What we have witnessed in recent years—the rise of deep learning, relying on the powerful function approximation and representation learning properties of deep neural networks—has provided us with new tools to overcoming these problems.

Digital Object Identifier 10.1109/MSP.2017.2743240
Date of publication: 13 November 2017

The advent of deep learning has had a significant impact on many areas in machine learning, dramatically improving the state of the art in tasks such as object detection, speech recognition, and language translation [39]. The most important property of deep learning is that deep neural networks can automatically find compact low-dimensional representations (features) of high-dimensional data (e.g., images, text, and audio). Through crafting inductive biases into neural network architectures, particularly that of hierarchical representations, machine-learning practitioners have made effective progress in addressing the curse of dimensionality [7]. Deep learning has similarly accelerated progress in RL, with the use of deep-learning algorithms within RL defining the field of DRL. The aim of this survey is to cover both seminal and recent developments in DRL, conveying the innovative ways in which neural networks can be used to bring us closer toward developing autonomous agents. For a more comprehensive survey of recent efforts in DRL, we refer readers to the overview by Li [43].

Deep learning enables RL to scale to decision-making problems that were previously intractable, i.e., settings with high-dimensional state and action spaces. Among recent work in the

field of DRL, there have been two outstanding success stories. The first, kickstarting the revolution in DRL, was the development of an algorithm that could learn to play a range of Atari 2600 video games at a superhuman level, directly from image pixels [47]. Providing solutions for the instability of function approximation techniques in RL, this work was the first to convincingly demonstrate that RL agents could be trained on raw, high-dimensional observations, solely based on a reward signal. The second standout success was the development of a hybrid DRL system, AlphaGo, that defeated a human world champion in Go [73], paralleling the historic achievement of IBM's Deep Blue in chess two decades earlier [9]. Unlike the handcrafted rules that have dominated chess-playing systems, AlphaGo comprised neural networks that were trained using supervised learning and RL, in combination with a traditional heuristic search algorithm.

DRL algorithms have already been applied to a wide range of problems, such as robotics, where control policies for robots can now be learned directly from camera inputs in the real world [41], [42], succeeding controllers that used to be hand-engineered or learned from low-dimensional features of the robot's state. In Figure 1, we showcase just some of the domains that DRL has

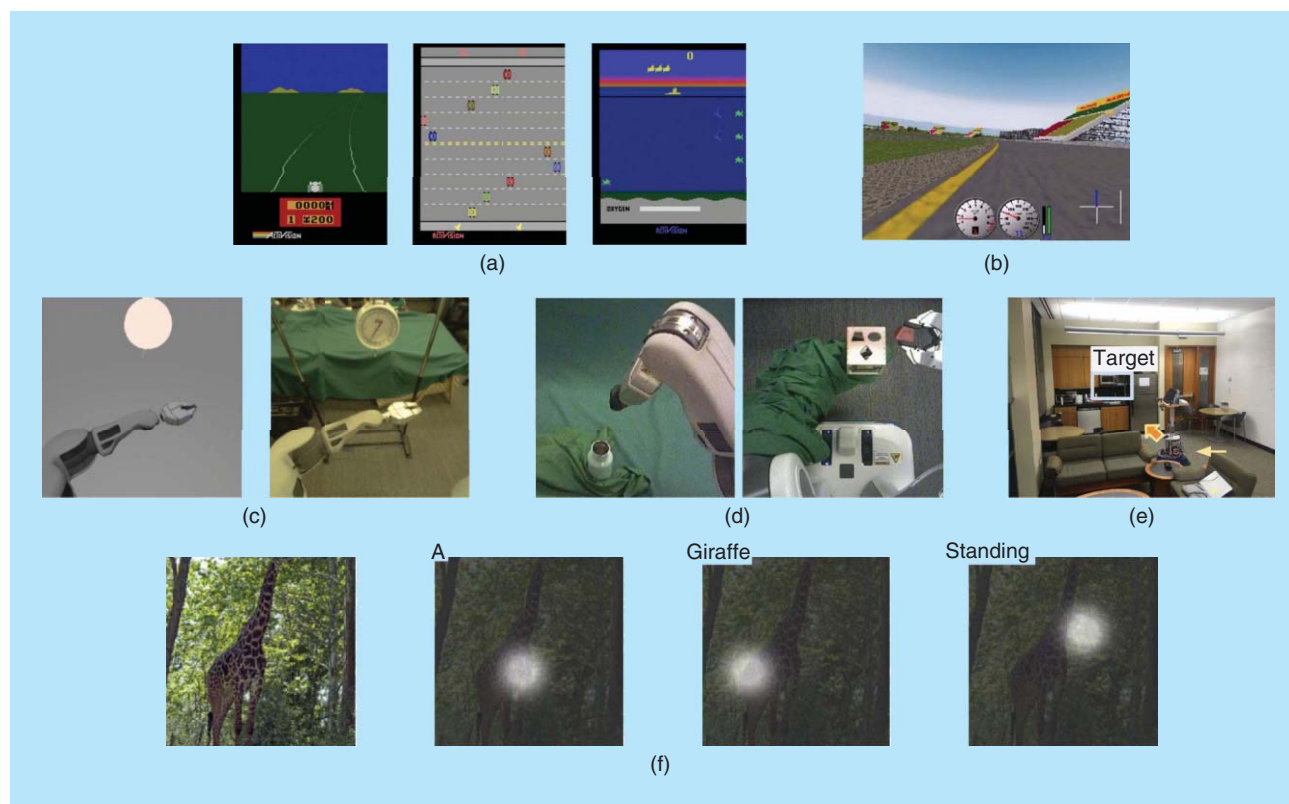


FIGURE 1. A range of visual RL domains. (a) Three classic Atari 2600 video games, *Enduro*, *Freeway*, and *Seaquest*, from the Arcade Learning Environment (ALE) [5]. Due to the range of supported games that vary in genre, visuals, and difficulty, the ALE has become a standard test bed for DRL algorithms [20], [47], [48], [55], [70], [75], [92]. The ALE is one of several benchmarks that are now being used to standardize evaluation in RL. (b) The TORCS car racing simulator, which has been used to test DRL algorithms that can output continuous actions [33], [44], [48] (as the games from the ALE only support discrete actions). (c) Utilizing the potentially unlimited amount of training data that can be amassed in robotic simulators, several methods aim to transfer knowledge from the simulator to the real world [11], [64], [84]. (d) Two of the four robotic tasks designed by Levine et al. [41]: screwing on a bottle cap and placing a shaped block in the correct hole. Levine et al. [41] were able to train visuomotor policies in an end-to-end fashion, showing that visual servoing could be learned directly from raw camera inputs by using deep neural networks. (e) A real room, in which a wheeled robot trained to navigate the building is given a visual cue as input and must find the corresponding location [100]. (f) A natural image being captioned by a neural network that uses RL to choose where to look [99]. (b)–(f) reproduced from [41], [44], [84], [99], and [100], respectively.

been applied to, ranging from playing video games [47] to indoor navigation [100].

Reward-driven behavior

Before examining the contributions of deep neural networks to RL, we will introduce the field of RL in general. The essence of RL is learning through interaction. An RL agent interacts with its environment and, upon observing the consequences of its actions, can learn to alter its own behavior in response to rewards received. This paradigm of trial-and-error learning has its roots in behaviorist psychology and is one of the main foundations of RL [78]. The other key influence on RL is optimal control, which has lent the mathematical formalisms (most notably dynamic programming [6]) that underpin the field.

In the RL setup, an autonomous agent, controlled by a machine-learning algorithm, observes a state \mathbf{s}_t from its environment at time step t . The agent interacts with the environment by taking an action \mathbf{a}_t in state \mathbf{s}_t . When the agent takes an action, the environment and the agent transition to a new state, \mathbf{s}_{t+1} , based on the current state and the chosen action. The state is a sufficient statistic of the environment and thereby comprises all the necessary information for the agent to take the best action, which can include parts of the agent such as the position of its actuators and sensors. In the optimal control literature, states and actions are often denoted by \mathbf{x}_t and \mathbf{u}_t , respectively.

The best sequence of actions is determined by the rewards provided by the environment. Every time the environment transitions to a new state, it also provides a scalar reward r_{t+1} to the agent as feedback. The goal of the agent is to learn a policy (control strategy) π that maximizes the expected return (cumulative, discounted reward). Given a state, a policy returns an action

to perform; an optimal policy is any policy that maximizes the expected return in the environment. In this respect, RL aims to solve the same problem as optimal control. However, the challenge in RL is that the agent needs to learn about the consequences of actions in the environment by trial and error, as, unlike in optimal control, a model of the state transition dynamics is not available to the agent. Every interaction with the environment yields information, which the agent uses to update its knowledge. This perception-action-learning loop is illustrated in Figure 2.

Markov decision processes

Formally, RL can be described as a Markov decision process (MDP), which consists of

- a set of states \mathcal{S} , plus a distribution of starting states $p(\mathbf{s}_0)$
- a set of actions \mathcal{A}
- transition dynamics $\mathcal{T}(\mathbf{s}_{t+1}|\mathbf{s}_t, \mathbf{a}_t)$ that map a state-action pair at time t onto a distribution of states at time $t + 1$
- an immediate/instantaneous reward function $\mathcal{R}(\mathbf{s}_t, \mathbf{a}_t, \mathbf{s}_{t+1})$
- a discount factor $\gamma \in [0, 1]$, where lower values place more emphasis on immediate rewards.

In general, the policy π is a mapping from states to a probability distribution over actions $\pi: \mathcal{S} \rightarrow p(\mathcal{A} = \mathbf{a} | \mathcal{S})$. If the MDP is episodic, i.e., the state is reset after each episode of length T , then the sequence of states, actions, and rewards in an episode constitutes a trajectory or rollout of the policy. Every rollout of a policy accumulates rewards from the environment, resulting in the return $R = \sum_{t=0}^{T-1} \gamma^t r_{t+1}$. The goal of RL is to find an optimal policy, π^* that achieves the maximum expected return from all states:

$$\pi^* = \operatorname{argmax}_{\pi} \mathbb{E}[R | \pi]. \tag{1}$$

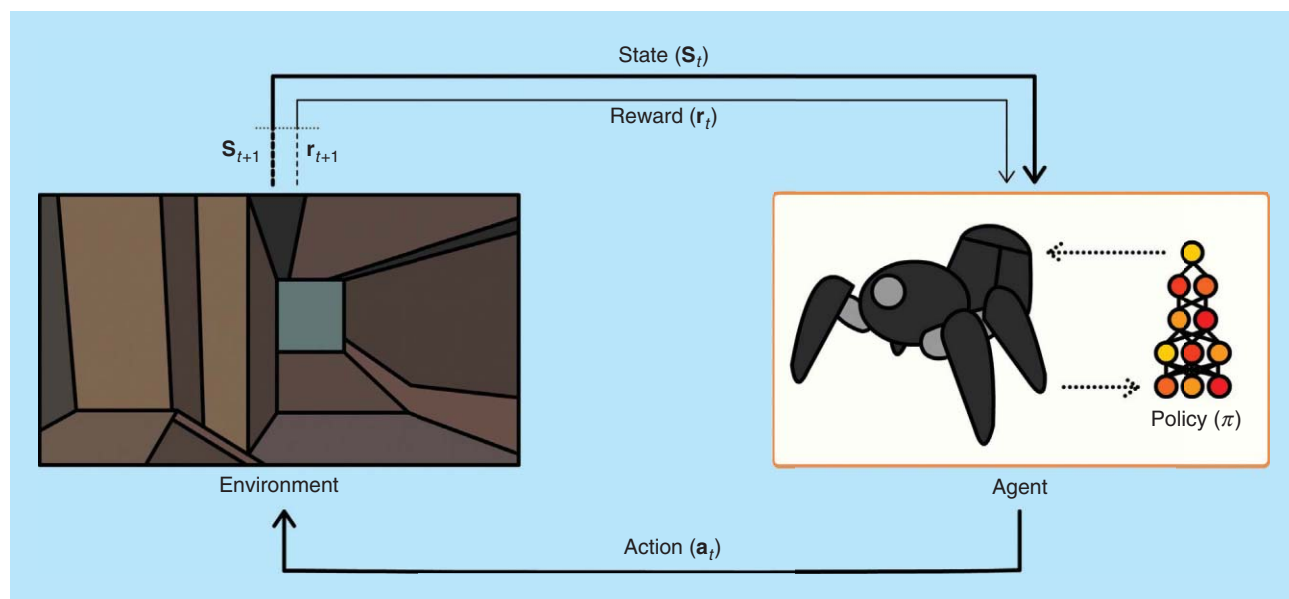


FIGURE 2. The perception-action-learning loop. At time t , the agent receives state \mathbf{s}_t from the environment. The agent uses its policy to choose an action \mathbf{a}_t . Once the action is executed, the environment transitions a step, providing the next state, \mathbf{s}_{t+1} , as well as feedback in the form of a reward, r_{t+1} . The agent uses knowledge of state transitions, of the form $(\mathbf{s}_t, \mathbf{a}_t, \mathbf{s}_{t+1}, r_{t+1})$, to learn and improve its policy.

It is also possible to consider nonepisodic MDPs, where $T = \infty$. In this situation, $\gamma < 1$ prevents an infinite sum of rewards from being accumulated. Furthermore, methods that rely on complete trajectories are no longer applicable, but those that use a finite set of transitions still are.

A key concept underlying RL is the Markov property—only the current state affects the next state, or, in other words, the future is conditionally independent of the past given the present state. This means that any decisions made at \mathbf{s}_t can be based solely on \mathbf{s}_{t+1} , rather than $\{\mathbf{s}_0, \mathbf{s}_1, \dots, \mathbf{s}_{t-1}\}$. Although this assumption is held by the majority of RL algorithms, it is somewhat unrealistic, as it requires the states to be fully observable. A generalization of MDPs are partially observable MDPs (POMDPs), in which the agent receives an observation $\mathbf{o}_t \in \Omega$, where the distribution of the observation $p(\mathbf{o}_{t+1} | \mathbf{s}_{t+1}, \mathbf{a}_t)$ is dependent on the current state and the previous action [27]. In a control and signal processing context, the observation would be described by a measurement/observation mapping in a state-space model that depends on the current state and the previously applied action.

POMDP algorithms typically maintain a belief over the current state given the previous belief state, the action taken, and the current observation. A more common approach in deep learning is to utilize recurrent neural networks (RNNs) [20], [21], [48], [96], which, unlike feedforward neural networks, are dynamical systems.

Challenges in RL

It is instructive to emphasize some challenges faced in RL:

- The optimal policy must be inferred by trial-and-error interaction with the environment. The only learning signal the agent receives is the reward.
- The observations of the agent depend on its actions and can contain strong temporal correlations.
- Agents must deal with long-range time dependencies: often the consequences of an action only materialize after many transitions of the environment. This is known as the (temporal) *credit assignment problem* [78].

We will illustrate these challenges in the context of an indoor robotic visual navigation task: if the goal location is specified, we may be able to estimate the distance remaining (and use it as a reward signal), but it is unlikely that we will know exactly what series of actions the robot needs to take to reach the goal. As the robot must choose where to go as it navigates the building, its decisions influence which rooms it sees and, hence, the statistics of the visual sequence captured. Finally, after navigating several junctions, the robot may find itself in a dead end. There is a range of problems, from learning the consequences of actions to balancing exploration versus exploitation, but ultimately these can all be addressed formally within the framework of RL.

RL algorithms

So far, we have introduced the key formalism used in RL, the MDP, and briefly noted some challenges in RL. In the following, we will distinguish between different classes of RL algorithms.

There are two main approaches to solving RL problems: methods based on value functions and methods based on policy search. There is also a hybrid actor-critic approach that employs both value functions and policy search. Next, we will explain these approaches and other useful concepts for solving RL problems.

Value functions

Value function methods are based on estimating the value (expected return) of being in a given state. The state-value function $V^\pi(\mathbf{s})$ is the expected return when starting in state \mathbf{s} and following π subsequently:

$$V^\pi(\mathbf{s}) = \mathbb{E}[R | \mathbf{s}, \pi]. \quad (2)$$

The optimal policy, π^* , has a corresponding state-value function $V^*(\mathbf{s})$, and vice versa; the optimal state-value function can be defined as

$$V^*(\mathbf{s}) = \max_{\pi} V^\pi(\mathbf{s}) \quad \forall \mathbf{s} \in \mathcal{S}. \quad (3)$$

If we had $V^*(\mathbf{s})$ available, the optimal policy could be retrieved by choosing among all actions available at \mathbf{s}_t and picking the action \mathbf{a} that maximizes $\mathbb{E}_{\mathbf{s}_{t+1} \sim \mathcal{T}(\mathbf{s}_t, \mathbf{a})}[V^*(\mathbf{s}_{t+1})]$.

In the RL setting, the transition dynamics \mathcal{T} are unavailable. Therefore, we construct another function, the state-action value or quality function $Q^\pi(\mathbf{s}, \mathbf{a})$, which is similar to V^π , except that the initial action \mathbf{a} is provided and π is only followed from the succeeding state onward:

$$Q^\pi(\mathbf{s}, \mathbf{a}) = \mathbb{E}[R | \mathbf{s}, \mathbf{a}, \pi]. \quad (4)$$

The best policy, given $Q^\pi(\mathbf{s}, \mathbf{a})$, can be found by choosing \mathbf{a} greedily at every state: $\operatorname{argmax}_{\mathbf{a}} Q^\pi(\mathbf{s}, \mathbf{a})$. Under this policy, we can also define $V^*(\mathbf{s})$ by maximizing $Q^\pi(\mathbf{s}, \mathbf{a})$: $V^*(\mathbf{s}) = \max_{\mathbf{a}} Q^\pi(\mathbf{s}, \mathbf{a})$.

Dynamic programming

To actually learn Q^π , we exploit the Markov property and define the function as a Bellman equation [6], which has the following recursive form:

$$Q^\pi(\mathbf{s}_t, \mathbf{a}_t) = \mathbb{E}_{\mathbf{s}_{t+1}}[r_{t+1} + \gamma Q^\pi(\mathbf{s}_{t+1}, \pi(\mathbf{s}_{t+1}))]. \quad (5)$$

This means that Q^π can be improved by bootstrapping, i.e., we can use the current values of our estimate of Q^π to improve our estimate. This is the foundation of Q -learning [94] and the state-action-reward-state-action (SARSA) algorithm [62]:

$$Q^\pi(\mathbf{s}_t, \mathbf{a}_t) \leftarrow Q^\pi(\mathbf{s}_t, \mathbf{a}_t) + \alpha \delta, \quad (6)$$

where α is the learning rate and $\delta = Y - Q^\pi(\mathbf{s}_t, \mathbf{a}_t)$ the temporal difference (TD) error; here, Y is a target as in a standard regression problem. SARSA, an on-policy learning algorithm, is used to improve the estimate of Q^π by using transitions generated by the behavioral policy (the policy derived from Q^π), which results in setting $Y = r_t + \gamma Q^\pi(\mathbf{s}_{t+1}, \mathbf{a}_{t+1})$. Q -learning

is off-policy, as Q^π is instead updated by transitions that were not necessarily generated by the derived policy. Instead, Q -learning uses $Y = r_t + \gamma \max_a Q^\pi(s_{t+1}, \mathbf{a})$, which directly approximates Q^* .

To find Q^* from an arbitrary Q^π , we use generalized policy iteration, where policy iteration consists of policy evaluation and policy improvement. Policy evaluation improves the estimate of the value function, which can be achieved by minimizing TD errors from trajectories experienced by following the policy. As the estimate improves, the policy can naturally be improved by choosing actions greedily based on the updated value function. Instead of performing these steps separately to convergence (as in policy iteration), generalized policy iteration allows for interleaving the steps, such that progress can be made more rapidly.

Sampling

Instead of bootstrapping value functions using dynamic programming methods, Monte Carlo methods estimate the expected return (2) from a state by averaging the return from multiple rollouts of a policy. Because of this, pure Monte Carlo methods can also be applied in non-Markovian environments. On the other hand, they can only be used in episodic MDPs, as a rollout has to terminate for the return to be calculated. It is possible to get the best of both methods by combining TD learning and Monte Carlo policy evaluation, as is done in the TD(λ) algorithm [78]. Similarly to the discount factor, the λ in TD(λ) is used to interpolate between Monte Carlo evaluation and bootstrapping. As demonstrated in Figure 3, this results in

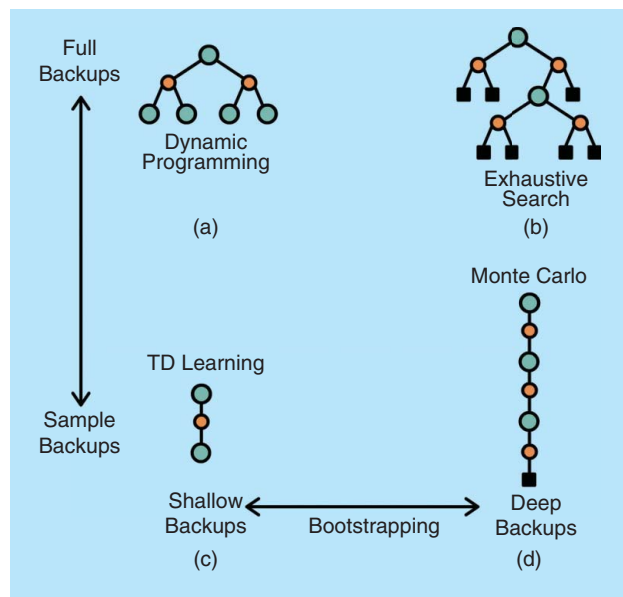


FIGURE 3. Two dimensions of RL algorithms based on the backups used to learn or construct a policy. At the extremes of these dimensions are (a) dynamic programming, (b) exhaustive search, (c) one-step TD learning, and (d) Monte Carlo approaches. Bootstrapping extends from (c) one-step TD learning to n -step TD learning methods [78], with (d) pure Monte Carlo approaches not relying on bootstrapping at all. Another possible dimension of variation is (c) and (d) choosing to sample actions versus (a) and (b) taking the expectation over all choices. (Figure recreated based on [78].)

an entire spectrum of RL methods based around the amount of sampling utilized.

Another major value-function-based method relies on learning the advantage function $A^\pi(s, \mathbf{a})$ [3]. Unlike producing absolute state-action values, as with Q^π , A^π instead represents relative state-action values. Learning relative values is akin to removing a baseline or average level of a signal; more intuitively, it is easier to learn that one action has better consequences than another than it is to learn the actual return from taking the action. A^π represents a relative advantage of actions through the simple relationship $A^\pi = Q^\pi - V^\pi$ and is also closely related to the baseline method of variance reduction within gradient-based policy search methods [97]. The idea of advantage updates has been utilized in many recent DRL algorithms [19], [48], [71], [92].

Policy search

Policy search methods do not need to maintain a value function model but directly search for an optimal policy π^* . Typically, a parameterized policy π_θ is chosen, whose parameters are updated to maximize the expected return $\mathbb{E}[R|\theta]$ using either gradient-based or gradient-free optimization [12]. Neural networks that encode policies have been successfully trained using both gradient-free [17], [33] and gradient-based [22], [41], [44], [70], [71], [96], [97] methods. Gradient-free optimization can effectively cover low-dimensional parameter spaces, but, despite some successes in applying them to large networks [33], gradient-based training remains the method of choice for most DRL algorithms, being more sample efficient when policies possess a large number of parameters.

When constructing the policy directly, it is common to output parameters for a probability distribution; for continuous actions, this could be the mean and standard deviations of Gaussian distributions, while for discrete actions this could be the individual probabilities of a multinomial distribution. The result is a stochastic policy from which we can directly sample actions. With gradient-free methods, finding better policies requires a heuristic search across a predefined class of models. Methods such as evolution strategies essentially perform hill climbing in a subspace of policies [65], while more complex methods, such as compressed network search, impose additional inductive biases [33]. Perhaps the greatest advantage of gradient-free policy search is that it can also optimize nondifferentiable policies.

Policy gradients

Gradients can provide a strong learning signal as to how to improve a parameterized policy. However, to compute the expected return (1) we need to average over plausible trajectories induced by the current policy parameterization. This averaging requires either deterministic approximations (e.g., linearization) or stochastic approximations via sampling [12]. Deterministic approximations can be only applied in a model-based setting where a model of the underlying transition dynamics is available. In the more common model-free RL setting, a Monte Carlo estimate of the expected return is

determined. For gradient-based learning, this Monte Carlo approximation poses a challenge since gradients cannot pass through these samples of a stochastic function. Therefore, we turn to an estimator of the gradient, known in RL as the REINFORCE rule [97]. Intuitively, gradient ascent using the estimator increases the log probability of the sampled action, weighted by the return. More formally, the REINFORCE rule can be used to compute the gradient of an expectation over a function f of a random variable X with respect to parameters θ :

$$\nabla_{\theta} \mathbb{E}_X[f(X; \theta)] = \mathbb{E}_X[f(X; \theta) \nabla_{\theta} \log p(X)]. \quad (7)$$

As this computation relies on the empirical return of a trajectory, the resulting gradients possess a high variance. By introducing unbiased estimates that are less noisy, it is possible to reduce the variance. The general methodology for performing this is to subtract a baseline, which means weighting updates by an advantage rather than the pure return. The simplest baseline is the average return taken over several episodes [97], but there are many more options available [71].

Actor-critic methods

It is possible to combine value functions with an explicit representation of the policy, resulting in actor-critic methods, as shown in Figure 4. The “actor” (policy) learns by using feedback from the “critic” (value function). In doing so, these methods tradeoff variance reduction of policy gradients with bias introduction from value function methods [32], [71].

Actor-critic methods use the value function as a baseline for policy gradients, such that the only fundamental difference between actor-critic methods and other baseline methods is that actor-critic methods utilize a learned value function. For this reason, we will later discuss actor-critic methods as a subset of policy gradient methods.

Planning and learning

Given a model of the environment, it is possible to use dynamic programming over all possible actions [Figure 3(a)], sample trajectories for heuristic search (as was done by AlphaGo [73]), or even perform an exhaustive search [Figure 3(b)]. Sutton and Barto [78] define *planning* as any method that utilizes a model to produce or improve a policy. This includes distribution models, which include \mathcal{T} and \mathcal{R} , and sample models, from which only samples of transitions can be drawn.

In RL, we focus on learning without access to the underlying model of the environment. However, interactions with the environment could be used to learn value functions, policies, and also a model. Model-free RL methods learn directly from interactions with the environment, but model-based RL methods can simulate transitions using the learned model, resulting in increased sample efficiency. This is particularly important in domains where each interaction with the environment is expensive. However, learning a model introduces extra complexities, and there is always the

Searching directly for a policy represented by a neural network with very many parameters can be difficult and can suffer from severe local minima.

danger of suffering from model errors, which in turn affects the learned policy. Although deep neural networks can potentially produce very complex and rich models [14], [55], [75], sometimes simpler, more data-efficient methods are preferable [19]. These considerations also play a role in actor-critic methods with learned value functions [32], [71].

The rise of DRL

Many of the successes in DRL have been based on scaling up prior work in RL to high-dimensional problems. This is due to the learning of low-dimensional feature representations and the powerful function approximation properties of neural networks. By means of representation learning, DRL can deal efficiently with the curse of dimensionality, unlike tabular and traditional nonparametric methods [7]. For instance, convolutional neural networks (CNNs) can be used as components of RL agents, allowing them to learn directly from raw, high-dimensional visual inputs. In general, DRL is based on training deep neural networks to approximate the optimal policy π^* and/or the optimal value functions V^* , Q^* , and A^* .

Value functions

The well-known function approximation properties of neural networks led naturally to the use of deep learning to regress functions for use in RL agents. Indeed, one of the earliest success stories in RL is TD-Gammon, a neural network that reached expert-level performance in backgammon in the early 1990s [81]. Using TD methods, the network took in the state of the board to predict the probability of black or white winning. Although this simple idea has been echoed in later work [73], progress in RL research has favored the explicit use of value functions, which can capture the

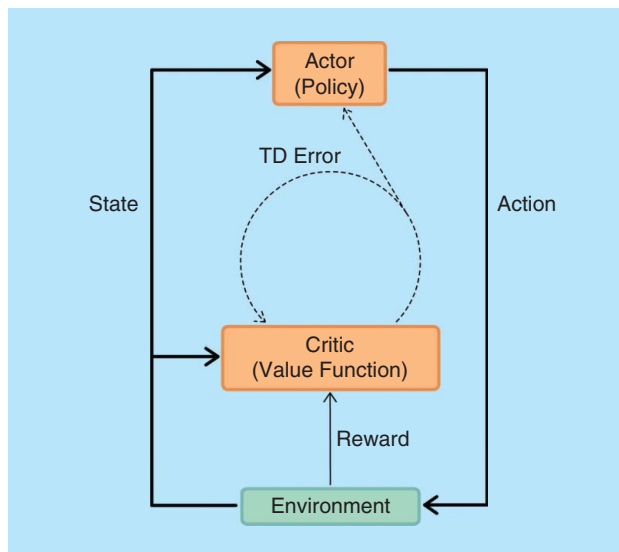


FIGURE 4. The actor-critic setup. The actor (policy) receives a state from the environment and chooses an action to perform. At the same time, the critic (value function) receives the state and reward resulting from the previous interaction. The critic uses the TD error calculated from this information to update itself and the actor. (Figure recreated based on [78].)

structure underlying the environment. From early value function methods in DRL, which took simple states as input [61], current methods are now able to tackle visually and conceptually complex environments [47], [48], [70], [100].

Function approximation and the DQN

We begin our survey of value-function-based DRL algorithms with the DQN [47], illustrated in Figure 5, which achieved scores across a wide range of classic Atari 2600 video games [5] that were comparable to that of a professional video games tester. The inputs to the DQN are four gray-scale frames of the game, concatenated over time, which are initially processed by several convolutional layers to extract spatiotemporal features, such as the movement of the ball in *Pong* or *Breakout*. The final feature map from the convolutional layers is processed by several fully connected layers, which more implicitly encode the effects of actions. This contrasts with more traditional controllers that use fixed preprocessing steps, which are therefore unable to adapt their processing of the state in response to the learning signal.

A forerunner of the DQN—neural-fitted Q (NFQ) iteration—involved training a neural network to return the Q -value given a state-action pair [61]. NFQ was later extended to train a network to drive a slot car using raw visual inputs from a camera over the race track, by combining a deep autoencoder to reduce the dimensionality of the inputs with a separate branch to predict Q -values [38]. Although the previous network could have been trained for both reconstruction and RL tasks simultaneously, it was both more reliable and computationally efficient to train the two parts of the network sequentially.

The DQN [47] is closely related to the model proposed by Lange et al. [38] but was the first RL algorithm that was demonstrated to work directly from raw visual inputs and on a wide variety of environments. It was designed such that the final fully connected layer outputs $Q^\pi(\mathbf{s}, \cdot)$ for all action values in a discrete set of actions—in this case, the various directions of the joystick and the fire button. This not only enables the best action, $\operatorname{argmax}_{\mathbf{a}} Q^\pi(\mathbf{s}, \mathbf{a})$, to be chosen after a single forward pass of the network, but also allows the network to more easily encode action-independent knowledge in the lower, convolutional layers. With merely the goal of

maximizing its score on a video game, the DQN learns to extract salient visual features, jointly encoding objects, their movements, and, most importantly, their interactions. Using techniques originally developed for explaining the behavior of CNNs in object recognition tasks, we can also inspect what parts of its view the agent considers important (see Figure 6).

The true underlying state of the game is contained within 128 bytes of Atari 2600 random-access memory. However, the DQN was designed to directly learn from visual inputs (210×160 pixel 8-bit RGB images), which it takes as the state \mathbf{s} . It is impractical to represent $Q^\pi(\mathbf{s}, \mathbf{a})$ exactly as a lookup table: when combined with 18 possible actions, we obtain a Q -table of size $|\mathcal{S}| \times |\mathcal{A}| = 18 \times 256^{3 \times 210 \times 160}$. Even if it were feasible to create such a table, it would be sparsely populated, and information gained from one state-action pair cannot be propagated to other state-action pairs. The strength of the DQN lies in its ability to compactly represent both high-dimensional observations and the Q -function using deep neural networks. Without this ability, tackling the discrete Atari domain from raw visual inputs would be impractical.

The DQN addressed the fundamental instability problem of using function approximation in RL [83] by the use of two techniques: experience replay [45] and target networks. Experience replay memory stores transitions of the form $(\mathbf{s}_t, \mathbf{a}_t, \mathbf{s}_{t+1}, r_{t+1})$ in a cyclic buffer, enabling the RL agent to sample from and train on previously observed data offline. Not only does this massively reduce the number of interactions needed with the environment, but batches of experience can be sampled, reducing the variance of learning updates. Furthermore, by sampling uniformly from a large memory, the temporal correlations that can adversely affect RL algorithms are broken. Finally, from a practical perspective, batches of data can be efficiently processed in parallel by modern hardware, increasing throughput. While the original DQN algorithm used uniform sampling [47], later work showed that prioritizing samples based on TD errors is more effective for learning [67].

The second stabilizing method, introduced by Mnih et al. [47], is the use of a target network that initially contains the weights of the network enacting the policy but is kept frozen for a large period of time. Rather than having to calculate the

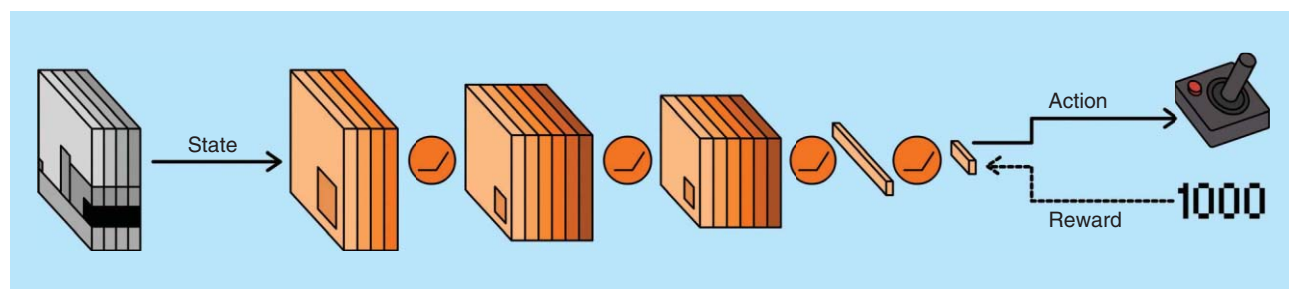


FIGURE 5. The DQN [47]. The network takes the state—a stack of gray-scale frames from the video game—and processes it with convolutional and fully connected layers, with ReLU nonlinearities in between each layer. At the final layer, the network outputs a discrete action, which corresponds to one of the possible control inputs for the game. Given the current state and chosen action, the game returns a new score. The DQN uses the reward—the difference between the new score and the previous one—to learn from its decision. More precisely, the reward is used to update its estimate of Q , and the error between its previous estimate and its new estimate is backpropagated through the network.

TD error based on its own rapidly fluctuating estimates of the Q -values, the policy network uses the fixed target network. During training, the weights of the target network are updated to match the policy network after a fixed number of steps. Both experience replay and target networks have gone on to be used in subsequent DRL works [19], [44], [50], [93].

Q-function modifications

Considering that one of the key components of the DQN is a function approximator for the Q -function, it can benefit from fundamental advances in RL. In [86], van Hasselt showed that the single estimator used in the Q -learning update rule overestimates the expected return due to the use of the maximum action value as an approximation of the maximum expected action value. Double- Q learning provides a better estimate through the use of a double estimator [86]. While double- Q learning requires an additional function to be learned, later work proposed using the already available target network from the DQN algorithm, resulting in significantly better results with only a small change in the update step [87].

Yet another way to adjust the DQN architecture is to decompose the Q -function into meaningful functions, such as constructing Q^π by adding together separate layers that compute the state-value function V^π and advantage function A^π [92]. Rather than having to come up with accurate Q -values for all actions, the duelling DQN [92] benefits from a single baseline for the state in the form of V^π and easier-to-learn relative values in the form of A^π . The combination of the duelling DQN with prioritized experience replay [67] is one of the state-of-the-art techniques in discrete action settings. Further insight into the properties of A^π by Gu et al. [19] led them to modify the DQN with a convex advantage layer that extended the algorithm to work over sets of continuous actions, creating the normalized advantage function (NAF) algorithm. Benefiting from experience replay, target networks, and advantage updates, NAF is one of several state-of-the-art techniques in continuous control problems [19].

Policy search

Policy search methods aim to directly find policies by means of gradient-free or gradient-based methods. Prior to the current surge of interest in DRL, several successful methods in DRL eschewed the commonly used backpropagation algorithm in favor of evolutionary algorithms [17], [33], which are gradient-free policy search algorithms. Evolutionary methods rely on evaluating the performance of a population of agents. Hence, they are expensive for large populations or agents with many parameters. However, as black-box optimization methods, they can be used to optimize arbitrary, nondifferentiable models and naturally allow for more exploration in the parameter space. In combination with a compressed representation of neural network weights, evolutionary algorithms can even be used to train large networks; such a technique resulted in the first deep neural network to learn an RL task, straight from high-dimensional visual inputs [33]. Recent work has reignited interest in evolutionary methods for RL as they can

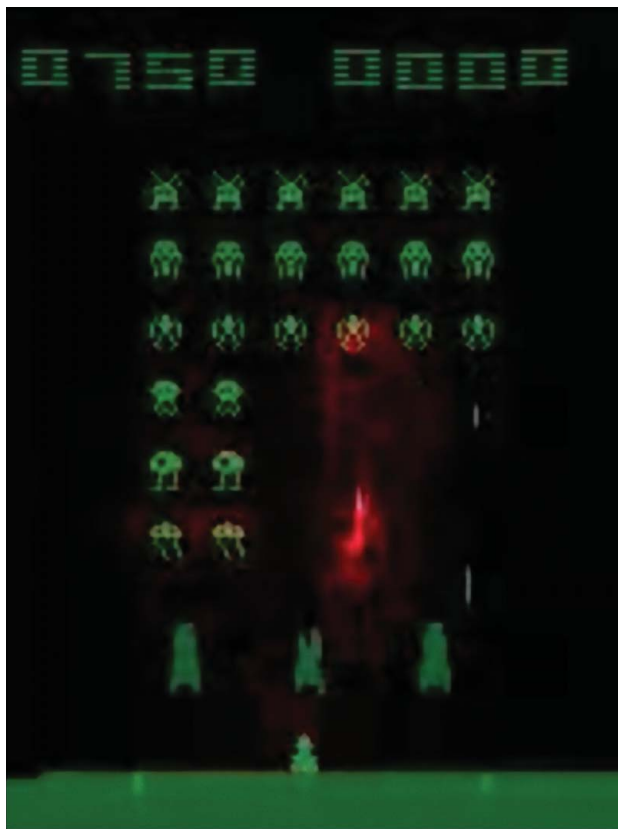


FIGURE 6. A saliency map of a trained DQN [47] playing *Space Invaders* [5]. By backpropagating the training signal to the image space, it is possible to see what a neural-network-based agent is attending to. In this frame, the most salient points—shown with the red overlay—are the laser that the agent recently fired and also the enemy that it anticipates hitting in a few time steps.

potentially be distributed at larger scales than techniques that rely on gradients [65].

Backpropagation through stochastic functions

The workhorse of DRL, however, remains backpropagation. The previously discussed REINFORCE rule [97] allows neural networks to learn stochastic policies in a task-dependent manner, such as deciding where to look in an image to track [69] or caption [99] objects. In these cases, the stochastic variable would determine the coordinates of a small crop of the image and hence reduce the amount of computation needed. This usage of RL to make discrete, stochastic decisions over inputs is known in the deep-learning literature as *hard attention* and is one of the more compelling uses of basic policysearch methods in recent years, having many applications outside of traditional RL domains.

Compounding errors

Searching directly for a policy represented by a neural network with very many parameters can be difficult and suffer from severe local minima. One way around this is to use guided policy search (GPS), which takes a few sequences of actions from another controller (which could be constructed using a separate method, such

as optimal control). GPS learns from them by using supervised learning in combination with importance sampling, which corrects for off-policy samples [40]. This approach effectively biases the search toward a good (local) optimum. GPS works in a loop, by optimizing policies to match sampled trajectories and optimizing trajectory distributions to match the policy and minimize costs. Levine et al. [41] showed that it was possible to train visuomotor policies for a robot “end to end,” straight from the RGB pixels of the camera to motor torques, and, hence, provide one of the seminal works in DRL.

A more commonly used method is to use a trust region, in which optimization steps are restricted to lie within a region where the approximation of the true cost function still holds. By preventing updated policies from deviating too wildly from previous policies, the chance of a catastrophically bad update is lessened, and many algorithms that use trust regions guarantee or practically result in monotonic improvement in policy performance. The idea of constraining each policy gradient update, as measured by the Kullback–Leibler (KL) divergence between the current and proposed policy, has a long history in RL [28]. One of the newer algorithms in this line of work, TRPO, has been shown to be relatively robust and applicable to domains with high-dimensional inputs [70]. To achieve this, TRPO optimizes a surrogate objective function—specifically, it optimizes an (importance sampled) advantage estimate, constrained using a quadratic approximation of the KL divergence. While TRPO can be used as a pure policy gradient method with a simple baseline, later work by Schulman et al. [71] introduced generalized advantage estimation (GAE), which proposed several, more advanced variance reduction baselines. The combination of TRPO and GAE remains one of the state-of-the-art RL techniques in continuous control.

Actor-critic methods

Actor-critic approaches have grown in popularity as an effective means of combining the benefits of policy search methods with learned value functions, which are able to learn from full returns and/or TD errors. They can benefit from improvements in both policy gradient methods, such as GAE [71], and value function methods, such as target networks [47]. In the last few years, DRL actor-critic methods have been scaled up from learning simulated physics tasks [22], [44] to real robotic visual navigation tasks [100], directly from image pixels.

One recent development in the context of actor-critic algorithms is deterministic policy gradients (DPGs) [72], which extend the standard policy gradient theorems for stochastic policies [97] to deterministic policies. One of the major advantages of DPGs is that, while stochastic policy gradients integrate over both state and action spaces, DPGs only integrate over the state space, requiring fewer samples in problems with large action spaces. In the initial work on DPGs, Silver et al. [72] introduced and demonstrated an off-policy actor-critic algorithm that vastly improved upon a stochastic policy gradient equivalent in high-dimensional continuous control problems. Later work introduced deep DPG, which utilized neural networks to operate on high-dimensional,

visual state spaces [44]. In the same vein as DPGs, Heess et al. [22] devised a method for calculating gradients to optimize stochastic policies by “reparameterizing” [30], [60] the stochasticity away from the network, thereby allowing standard gradients to be used (instead of the high-variance REINFORCE estimator [97]). The resulting stochastic value gradient (SVG) methods are flexible and can be used both with (SVG(0) and SVG(1)) and without (SVG(∞)) value function critics, and with (SVG(∞) and SVG(1)) and without (SVG(0)) models. Later work proceeded to integrate DPGs and SVGs with RNNs, allowing them to solve continuous control problems in POMDPs, learning directly from pixels [21]. Together, DPGs and SVGs can be considered algorithmic approaches for improving learning efficiency in DRL.

An orthogonal approach to speeding up learning is to exploit parallel computation. By keeping a canonical set of parameters that are read by and updated in an asynchronous fashion by multiple copies of a single network, computation can be efficiently distributed over both processing cores in a single central processing unit (CPU), and across CPUs in a cluster of machines. Using a distributed system, Nair et al. [51] developed a framework for training multiple DQNs in parallel, achieving both better performance and a reduction in training time. However, the simpler asynchronous advantage actor-critic (A3C) algorithm [48], developed for both single and distributed machine settings, has become one of the most popular DRL techniques in recent times. A3C combines advantage updates with the actor-critic formulation and relies on asynchronously updated policy and value function networks trained in parallel over several processing threads. The use of multiple agents, situated in their own, independent environments, not only stabilizes improvements in the parameters, but conveys an additional benefit in allowing for more exploration to occur. A3C has been used as a standard starting point in many subsequent works, including the work of Zhu et al. [100], who applied it to robotic navigation in the real world through visual inputs.

There have been several major advancements on the original A3C algorithm that reflect various motivations in the field of DRL. The first is actor-critic with experience replay [93], which adds off-policy bias correction to A3C, allowing it to use experience replay to improve sample complexity. Others have attempted to bridge the gap between value and policy-based RL, utilizing theoretical advancements to improve upon the original A3C [50], [54]. Finally, there is a growing trend toward exploiting auxiliary tasks to improve the representations learned by DRL agents and, hence, improve both the learning speed and final performance of these agents [26], [46].

Current research and challenges

To conclude, we will highlight some current areas of research in DRL and the challenges that still remain. Previously, we have focused mainly on model-free methods, but we will now examine a few model-based DRL algorithms in more detail. Model-based RL algorithms play an important role in making RL data efficient and in trading off exploration and exploitation. After tackling exploration strategies, we shall then address hierarchical

RL (HRL), which imposes an inductive bias on the final policy by explicitly factorizing it into several levels. When available, trajectories from other controllers can be used to bootstrap the learning process, leading us to imitation learning and inverse RL (IRL). For the final topic, we will look at multiagent systems, which have their own special considerations.

Model-based RL

The key idea behind model-based RL is to learn a transition model that allows for simulation of the environment without interacting with the environment directly. Model-based RL does not assume specific prior knowledge. However, in practice, we can incorporate prior knowledge (e.g., physics-based models [29]) to speed up learning. Model learning plays an important role in reducing the number of required interactions with the (real) environment, which may be limited in practice. For example, it is unrealistic to perform millions of experiments with a robot in a reasonable amount of time and without significant hardware wear and tear. There are various approaches to learn predictive models of dynamical systems using pixel information. Based on the deep dynamical model [90], where high-dimensional observations are embedded into a lower-dimensional space using autoencoders, several model-based DRL algorithms have been proposed for learning models and policies from pixel information [55], [91], [95]. If a sufficiently accurate model of the environment can be learned, then even simple controllers can be used to control a robot directly from camera images [14]. Learned models can also be used to guide exploration purely based on simulation of the environment, with deep models allowing these techniques to be scaled up to high-dimensional visual domains [75].

Although deep neural networks can make reasonable predictions in simulated environments over hundreds of time steps [10], they typically require many samples to tune the large number of parameters they contain. Training these models often requires more samples (interaction with the environment) than simpler models. For this reason, Gu et al. [19] train locally linear models for use with the NAF algorithm—the continuous equivalent of the DQN [47]—to improve the algorithm's sample complexity in the robotic domain where samples are expensive. It seems likely that the usage of deep models in model-based DRL could be massively spurred by general advances in improving the data efficiency of neural networks.

Exploration versus exploitation

One of the greatest difficulties in RL is the fundamental dilemma of exploration versus exploitation: When should the agent try out (perceived) nonoptimal actions to explore the environment (and potentially improve the model), and when should it exploit the optimal action to make useful progress? Off-policy algorithms, such as the DQN [47], typically use the simple ϵ -greedy exploration policy, which chooses a random action with probability $\epsilon \in [0, 1]$, and the optimal action otherwise. By decreasing ϵ over time, the agent progresses toward exploitation. Although adding independent noise for exploration is usable in continuous control problems, more sophisticated strategies inject noise that is corre-

lated over time (e.g., from stochastic processes) to better preserve momentum [44].

The observation that temporal correlation is important led Osband et al. [56] to propose the bootstrapped DQN, which maintains several Q -value “heads” that learn different values through a combination of different weight initializations and bootstrapped sampling from experience replay memory. At the beginning of each training episode, a different head is chosen, leading to temporally extended exploration. Usunier et al. [85] later proposed a similar method that performed exploration in policy space by adding noise to a single output head, using zero-order gradient estimates to allow backpropagation through the policy.

One of the main principled exploration strategies is the upper confidence bound (UCB) algorithm, based on the principle of “optimism in the face of uncertainty” [36]. The idea behind UCB is to pick actions that maximize $\mathbb{E}[R] + \kappa\sigma[R]$, where $\sigma[R]$ is the standard deviation of the return and $\kappa > 0$. UCB therefore encourages exploration in regions with high uncertainty and moderate expected return. While easily achievable in small tabular cases, the use of powerful density models has allowed this algorithm to scale to high-dimensional visual domains with DRL [4].

UCB can also be considered one way of implementing intrinsic motivation, which is a general concept that advocates decreasing uncertainty/making progress in learning about the environment [68]. There have been several DRL algorithms that try to implement intrinsic motivation via minimizing model prediction error [57], [75] or maximizing information gain [25], [49].

Hierarchical RL

In the same way that deep learning relies on hierarchies of features, HRL relies on hierarchies of policies. Early work in this area introduced options, in which, apart from primitive actions (single time-step actions), policies could also run other policies (multitime-step “actions”) [79]. This approach allows top-level policies to focus on higher-level goals, while subpolicies are responsible for fine control. Several works in DRL have attempted HRL by using one top-level policy that chooses between subpolicies, where the division of states or goals in to subpolicies is achieved either manually [1], [34], [82] or automatically [2], [88], [89]. One way to help construct subpolicies is to focus on discovering and reaching goals, which are specific states in the environment; they may often be locations, to which an agent should navigate. Whether utilized with HRL or not, the discovery and generalization of goals is also an important area of ongoing research [35], [66], [89].

Imitation learning and inverse RL

One may ask why, if given a sequence of “optimal” actions from expert demonstrations, it is not possible to use supervised learning in a straightforward manner—a case of “learning from demonstration.” This is indeed possible and is known as *behavioral cloning* in traditional RL literature. Taking advantage of the stronger signals available in supervised learning problems, behavioral cloning enjoyed success in earlier

neural network research, with the most notable success being ALVINN, one of the earliest autonomous cars [59]. However, behavioral cloning cannot adapt to new situations, and small deviations from the demonstration during the execution of the learned policy can compound and lead to scenarios where the policy is unable to recover. A more generalizable solution is to use provided trajectories to guide the learning of suitable state-action pairs but fine-tune the agent using RL [23].

The goal of IRL is to estimate an unknown reward function from observed trajectories that characterize a desired solution [52]; IRL can be used in combination with RL to improve upon demonstrated behavior. Using the power of deep neural networks, it is now possible to learn complex, nonlinear reward functions for IRL [98]. Ho and Ermon [24] showed that policies are uniquely characterized by their occupancies (visited state and action distributions) allowing IRL to be reduced to the problem of measure matching. With this insight, they were able to use generative adversarial training [18] to facilitate reward-function learning in a more flexible manner, resulting in the generative adversarial imitation learning algorithm.

Multiagent RL

Usually, RL considers a single learning agent in a stationary environment. In contrast, multiagent RL (MARL) considers multiple agents learning through RL and often the nonstationarity introduced by other agents changing their behaviors as they learn [8]. In DRL, the focus has been on enabling (differentiable) communication between agents, which allows them to cooperate. Several approaches have been proposed for this purpose, including passing messages to agents sequentially [15], using a bidirectional channel (providing ordering with less signal loss) [58], and an all-to-all channel [77]. The addition of communication channels is a natural strategy to apply to MARL in complex scenarios and does not preclude the usual practice of modeling cooperative or competing agents as applied elsewhere in the MARL literature [8].

Conclusion: Beyond pattern recognition

Despite the successes of DRL, many problems need to be addressed before these techniques can be applied to a wide range of complex real-world problems [37]. Recent work with (nondeep) generative causal models demonstrated superior generalization over standard DRL algorithms [48], [63] in some benchmarks [5], achieved by reasoning about causes and effects in the environment [29]. For example, the schema networks of Kankys et al. [29] trained on the game *Break-out* immediately adapted to a variant where a small wall was placed in front of the target blocks, while progressive (A3C) networks [63] failed to match the performance of the schema networks even after training on the new domain. Although DRL has already been combined with AI techniques, such as search [73] and planning [80], a deeper integration with other traditional AI approaches promises benefits such as bet-

ter sample complexity, generalization, and interpretability [16]. In time, we also hope that our theoretical understanding of the properties of neural networks (particularly within DRL) will improve, as it currently lags far behind practice.

To conclude, it is worth revisiting the overarching goal of all of this research: the creation of general-purpose AI systems that can interact with and learn from the world around them. Interaction with the environment is simultaneously the advantage and disadvantage of RL. While there are many challenges in seeking to understand our complex and ever-changing world, RL allows

us to choose how we explore it. In effect, RL endows agents with the ability to perform experiments to better understand their surroundings, enabling them to learn even high-level causal relationships. The availability of high-quality visual renderers and physics engines now enables us to take steps in this direction, with works that try to learn intuitive models of physics in visual environments [13]. Challenges remain before this will be possible in the real world, but steady progress

is being made in agents that learn the fundamental principles of the world through observation and action. Perhaps, then, we are not too far away from AI systems that learn and act in more human-like ways in increasingly complex environments.

Acknowledgments

Kai Arulkumaran would like to acknowledge Ph.D. funding from the Department of Bioengineering at Imperial College London. This research has been partially funded by a Google Faculty Research Award to Marc Deisenroth.

Authors

Kai Arulkumaran (ka709@imperial.ac.uk) received a B.A. degree in computer science at the University of Cambridge in 2012 and an M.Sc. degree in biomedical engineering from Imperial College London in 2014, where he is currently a Ph.D. degree candidate in the Department of Bioengineering. He was a research intern at Twitter's Magic Pony and Microsoft Research in 2017. His research focus is deep reinforcement learning and transfer learning for visuomotor control.

Marc Peter Deisenroth (m.deisenroth@imperial.ac.uk) received an M.Eng. degree in computer science at the University of Karlsruhe in 2006 and a Ph.D. degree in machine learning at the Karlsruhe Institute of Technology in 2009. He is a lecturer of statistical machine learning in the Department of Computing at Imperial College London and with PROWLER.io. He was awarded an Imperial College Research Fellowship in 2014 and received Best Paper Awards at the International Conference on Robotics and Automation 2014 and the International Conference on Control, Automation, and Systems 2016. He is a recipient of a Google Faculty Research Award and a Microsoft Ph.D. Scholarship. His research is centered around data-efficient machine learning for autonomous decision making.

Miles Brundage (miles.brundage@philosophy.ox.ac.uk) received a B.A. degree in political science at George

In effect, RL endows agents with the ability to perform experiments to better understand their surroundings, enabling them to learn even high-level causal relationships.

Washington University, Washington, D.C., in 2010. He is a Ph.D. degree candidate in the Human and Social Dimensions of Science and Technology Department at Arizona State University and a research fellow at the University of Oxford's Future of Humanity Institute. His research focuses on governance issues related to artificial intelligence.

Anil Anthony Bharath (a.bharath@ic.ac.uk) received his B. Eng. degree in electronic and electrical engineering from University College London in 1988 and his Ph.D. degree in signal processing from Imperial College London in 1993, where he is currently a reader in the Department of Bioengineering. He is also a fellow of the Institution of Engineering and Technology and a cofounder of Cortexica Vision Systems. He was previously an academic visitor in the Signal Processing Group at the University of Cambridge in 2006. His research interests are in deep architectures for visual inference.

References

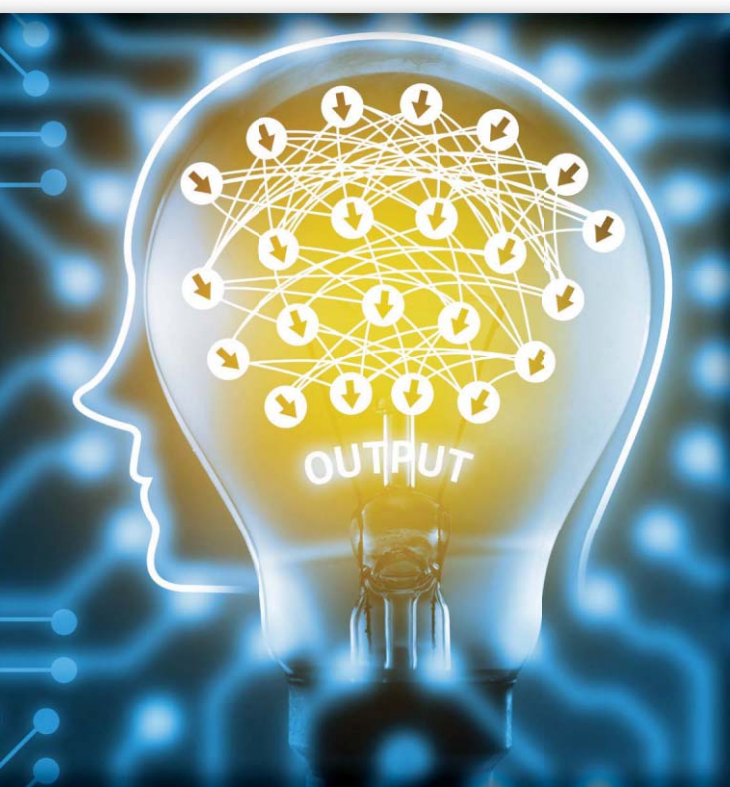
- [1] K. Arulkumaran, N. Dilothanakul, M. Shanahan, and A. A. Bharath, "Classifying options for deep reinforcement learning," in *Proc. IJCAI Workshop Deep Reinforcement Learning: Frontiers and Challenges*, 2016.
- [2] P. Bacon, J. Harb, and D. Precup, "The option-critic architecture," in *Proc. Association Advancement Artificial Intelligence*, 2017, pp. 1726–1734.
- [3] L. C. Baird III, "Advantage updating," Defense Tech. Inform. Center, Tech. Report D-A280 862, Fort Belvoir, VA, 1993.
- [4] M. Bellemare, S. Srinivasan, G. Ostrovski, T. Schaul, D. Saxton, and R. Munos, "Unifying count-based exploration and intrinsic motivation," in *Proc. Neural Information Processing Systems*, 2016, pp. 1471–1479.
- [5] M. G. Bellemare, Y. Naddaf, J. Veness, and M. Bowling, "The arcade learning environment: an evaluation platform for general agents," in *Proc. Int. Joint Conf. Artificial Intelligence*, 2015, pp. 253–279.
- [6] R. Bellman, "On the theory of dynamic programming," *Proc. Nat. Acad. Sci.*, vol. 38, no. 8, pp. 716–719, 1952.
- [7] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: a review and new perspectives," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 8, pp. 1798–1828, 2013.
- [8] L. Busoniu, R. Babuska, and B. De Schutter, "A comprehensive survey of multiagent reinforcement learning," *IEEE Trans. Syst., Man, Cybern.*, vol. 38, no. 2, pp. 156–172, 2008.
- [9] M. Campbell, A. J. Hoane, and F. Hsu, "Deep Blue," *Artificial Intell.*, vol. 134, no. 1–2, pp. 57–83, 2002.
- [10] S. Chiappa, S. Racaniere, D. Wierstra, and S. Mohamed, "Recurrent environment simulators," in *Proc. Int. Conf. Learning Representations*, 2017.
- [11] P. Christiano, Z. Shah, I. Mordatch, J. Schneider, T. Blackwell, J. Tobin, P. Abbeel, and W. Zaremba, (2016). Transfer from simulation to real world through learning deep inverse dynamics model. *arXiv*. [Online]. Available: <https://arxiv.org/abs/1610.03518>
- [12] M. P. Deisenroth, G. Neumann, and J. Peters, "A survey on policy search for robotics," *Foundations and Trends in Robotics*, vol. 2, no. 1–2, pp. 1–142, 2013.
- [13] M. Denil, P. Agrawal, T. D. Kulkarni, T. Erez, P. Battaglia, and N. de Freitas, "Learning to perform physics experiments via deep reinforcement learning," in *Proc. Int. Conf. Learning Representations*, 2017.
- [14] C. Finn, X. Y. Tan, Y. Duan, T. Darrell, S. Levine, and P. Abbeel, "Deep spatial autoencoders for visuomotor learning," in *Proc. IEEE Int. Conf. Robotics and Automation*, 2016, pp. 512–519.
- [15] J. Foerster, Y. M. Assael, N. de Freitas, and S. Whiteson, "Learning to communicate with deep multi-agent reinforcement learning," in *Proc. Neural Information Processing Systems*, 2016, pp. 2137–2145.
- [16] M. Garnelo, K. Arulkumaran, and M. Shanahan, "Towards deep symbolic reinforcement learning," in *NIPS Workshop on Deep Reinforcement Learning*, 2016.
- [17] F. Gomez and J. Schmidhuber, "Evolving modular fast-weight networks for control," in *Proc. Int. Conf. Artificial Neural Networks*, 2005, pp. 383–389.
- [18] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Proc. Neural Information Processing Systems*, 2014, pp. 2672–2680.
- [19] S. Gu, T. Lillicrap, I. Sutskever, and S. Levine, "Continuous deep Q-learning with model-based acceleration," in *Proc. Int. Conf. Learning Representations*, 2016.
- [20] M. Hausknecht and P. Stone, "Deep recurrent Q-learning for partially observable MDPs," in *Association for the Advancement of Artificial Intelligence Fall Symp. Series*, 2015.
- [21] N. Heess, J. J. Hunt, T. P. Lillicrap, and D. Silver, "Memory-based control with recurrent neural networks," in *NIPS Workshop on Deep Reinforcement Learning*, 2015.
- [22] N. Heess, G. Wayne, D. Silver, T. Lillicrap, T. Erez, and Y. Tassa, "Learning continuous control policies by stochastic value gradients," in *Proc. Neural Information Processing Systems*, 2015, pp. 2944–2952.
- [23] T. Hester, M. Vecerik, O. Pietquin, M. Lanctot, T. Schaul, B. Piot, A. Sendonaris, G. Dulac-Arnold, et al. (2017). Learning from demonstrations for real world reinforcement learning. *arXiv*. [Online]. Available: <https://arxiv.org/abs/1704.03732>
- [24] J. Ho and S. Ermon, "Generative adversarial imitation learning," in *Proc. Neural Information Processing Systems*, 2016, pp. 4565–4573.
- [25] R. Houthoofd, X. Chen, Y. Duan, J. Schulman, F. de Turck, and P. Abbeel, "VIME: Variational information maximizing exploration," in *Proc. Neural Information Processing Systems*, 2016, pp. 1109–1117.
- [26] M. Jaderberg, V. Mnih, W. M. Czarnecki, T. Schaul, J. Z. Leibo, D. Silver, and K. Kavukcuoglu, "Reinforcement learning with unsupervised auxiliary tasks," in *Proc. Int. Conf. Learning Representations*, 2017.
- [27] L. P. Kaelbling, M. L. Littman, and A. R. Cassandra, "Planning and acting in partially observable stochastic domains," *Artificial Intell.*, vol. 101, no. 1, pp. 99–134, 1998.
- [28] S. M. Kakade, "A natural policy gradient," in *Proc. Neural Information Processing Systems*, 2002, pp. 1531–1538.
- [29] K. Kanksy, T. Silver, D. A. Mély, M. Eldawy, M. Lázaro-Gredilla, X. Lou, N. Dofman, S. Sidor, S. Phoenix, and D. George, "Schema networks: zero-shot transfer with a generative causal model of intuitive physics," in *Proc. Int. Conf. Machine Learning*, 2017, pp. 1809–1818.
- [30] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," in *Proc. Int. Conf. Learning Representations*, 2014.
- [31] N. Kohl and P. Stone, "Policy gradient reinforcement learning for fast quadrupedal locomotion," in *Proc. IEEE Int. Conf. Robotics and Automation*, 2004, pp. 2619–2624.
- [32] V. R. Konda and J. N. Tsitsiklis, "On actor-critic algorithms," *SIAM J. Control Optim.*, vol. 42, no. 4, pp. 1143–1166, 2003.
- [33] J. Koutník, G. Cuccu, J. Schmidhuber, and F. Gomez, "Evolving large-scale neural networks for vision-based reinforcement learning," in *Proc. Conf. Genetic and Evolutionary Computation*, 2013, pp. 1061–1068.
- [34] T. D. Kulkarni, K. Narasimhan, A. Saeedi, and J. Tenenbaum, "Hierarchical deep reinforcement learning: Integrating temporal abstraction and intrinsic motivation," in *Proc. Neural Information Processing Systems*, 2016, pp. 3675–3683.
- [35] T. D. Kulkarni, A. Saeedi, S. Gautam, and S. J. Gershman, "Deep successor reinforcement learning," in *NIPS Workshop on Deep Reinforcement Learning*, 2016.
- [36] T. L. Lai and H. Robbins, "Asymptotically efficient adaptive allocation rules," *Adv. Appl. Math.*, vol. 6, no. 1, pp. 4–22, 1985.
- [37] B. M. Lake, T. D. Ullman, J. B. Tenenbaum, and S. J. Gershman, "Building machines that learn and think like people," *Behavioral Brain Sci.*, pp. 1–101, 2016. [Online]. Available: <https://www.cambridge.org/core/journals/behavioral-and-brain-sciences/article/building-machines-that-learn-and-think-like-people/A9535B1D745A0377E16C590E14B94993>
- [38] S. Lange, M. Riedmiller, and A. Voigtlander, "Autonomous reinforcement learning on raw visual input data in a real world application," in *Proc. Int. Joint Conf. Neural Networks*, 2012, pp. 1–8.
- [39] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.
- [40] S. Levine and V. Koltun, "Guided policy search," in *Proc. Int. Conf. Learning Representations*, 2013.
- [41] S. Levine, C. Finn, T. Darrell, and P. Abbeel, "End-to-end training of deep visuomotor policies," *J. Mach. Learning Res.*, vol. 17, no. 39, pp. 1–40, 2016.
- [42] S. Levine, P. Pastor, A. Krizhevsky, and D. Quillen, "Learning hand-eye coordination for robotic grasping with deep learning and large-scale data collection," in *Proc. Int. Symp. Experimental Robotics*, 2016, pp. 173–184.
- [43] Y. Li. (2017). Deep reinforcement learning: An overview. *arXiv*. [Online]. Available: <https://arxiv.org/abs/1701.07274>
- [44] T. P. Lillicrap, J. J. Hunt, A. Pritzel, N. Heess, T. Erez, Y. Tassa, D. Silver, and D. Wierstra, "Continuous control with deep reinforcement learning," in *Proc. Int. Conf. Learning Representations*, 2016.
- [45] L. Lin, "Self-improving reactive agents based on reinforcement learning, planning and teaching," *Mach. Learning*, vol. 8, no. 3–4, pp. 293–321, 1992.
- [46] P. Mirowski, R. Pascanu, F. Viola, H. Soyer, A. Ballard, A. Banino, M. Denil, R. Goroshin, et al., "Learning to navigate in complex environments," in *Proc. Int. Conf. Learning Representations*, 2017.
- [47] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, et al., "Human-level control through deep reinforcement learning," *Nature*, vol. 518, no. 7540, pp. 529–533, 2015.

- [48] V. Mnih, A. P. Badia, M. Mirza, A. Graves, T. P. Lillicrap, T. Harley, D. Silver, and K. Kavukcuoglu, "Asynchronous methods for deep reinforcement learning," in *Proc. Int. Conf. Learning Representations*, 2016.
- [49] S. Mohamed and D. J. Rezende, "Variational information maximisation for intrinsically motivated reinforcement learning," in *Proc. Neural Information Processing Systems*, 2015, pp. 2125–2133.
- [50] O. Nachum, M. Norouzi, K. Xu, and D. Schuurmans. (2017). Bridging the gap between value and policy based reinforcement learning. *arXiv*. [Online]. Available: <https://arxiv.org/abs/1702.08892>
- [51] A. Nair, P. Srinivasan, S. Blackwell, C. Alcicek, R. Fearon, A. de Maria, V. Panneershelvam, M. Suleyman, et al., "Massively parallel methods for deep reinforcement learning," in *ICML Workshop on Deep Learning*, 2015.
- [52] A. Y. Ng and S. J. Russell, "Algorithms for inverse reinforcement learning," in *Proc. Int. Conf. Machine Learning*, 2000, pp. 663–670.
- [53] A. Y. Ng, A. Coates, M. Diehl, V. Ganapathi, J. Schulte, B. Tse, E. Berger, and E. Liang, "Autonomous inverted helicopter flight via reinforcement learning," in *Proc. Int. Symp. Experimental Robotics*, 2006, pp. 363–372.
- [54] B. O'Donoghue, R. Munos, K. Kavukcuoglu, and V. Mnih, "PGQ: Combining policy gradient and Q-learning," in *Proc. Int. Conf. Learning Representations*, 2017.
- [55] J. Oh, X. Guo, H. Lee, R. L. Lewis, and S. Singh, "Action-conditional video prediction using deep networks in Atari games," in *Proc. Neural Information Processing Systems*, 2015, pp. 2863–2871.
- [56] I. Osband, C. Blundell, A. Pritzel, and B. van Roy, "Deep exploration via bootstrapped DQN," in *Proc. Neural Information Processing Systems*, 2016, pp. 4026–4034.
- [57] D. Pathak, P. Agrawal, A. A. Efros, and T. Darrell, "Curiosity-driven exploration by self-supervised prediction," in *Proc. Int. Conf. Machine Learning*, 2017, pp. 2778–2787.
- [58] P. Peng, Q. Yuan, Y. Wen, Y. Yang, Z. Tang, H. Long, and J. Wang. (2017). Multiagent bidirectionally-coordinated nets for learning to play StarCraft combat games. *arXiv*. [Online]. Available: <https://arxiv.org/abs/1703.10069>
- [59] D. A. Pomerleau, "ALVINN, an autonomous land vehicle in a neural network," in *Proc. Neural Information Processing Systems*, 1989, pp. 305–313.
- [60] D. J. Rezende, S. Mohamed, and D. Wierstra, "Stochastic backpropagation and approximate inference in deep generative models," in *Proc. Int. Conf. Machine Learning*, 2014, pp. 1278–1286.
- [61] M. Riedmiller, "Neural fitted q iteration—First experiences with a data efficient neural reinforcement learning method," in *Proc. European Conf. Machine Learning*, 2005, pp. 317–328.
- [62] G. A. Rummery and M. Niranjan, "On-line Q-learning using connectionist systems," Dept. Engineering, Univ. Cambridge, MA, Tech. Rep. CUED/F-INFENG/TR 166, 1994.
- [63] A. A. Rusu, N. C. Rabinowitz, G. Desjardins, H. Soyer, J. Kirkpatrick, K. Kavukcuoglu, R. Pascanu, and R. Hadsell. (2016). Progressive neural networks. *arXiv*. [Online]. Available: <https://arxiv.org/abs/1606.04671>
- [64] A. A. Rusu, M. Vecerik, T. Rothhörn, N. Heess, R. Pascanu, and R. Hadsell. (2016). Sim-to-real robot learning from pixels with progressive nets. *arXiv*. [Online]. Available: <https://arxiv.org/abs/1610.04286>
- [65] T. Salimans, J. Ho, X. Chen, and I. Sutskever. (2017). Evolution strategies as a scalable alternative to reinforcement learning. *arXiv*. [Online]. Available: <https://arxiv.org/abs/1703.03864>
- [66] T. Schaul, D. Horgan, K. Gregor, and D. Silver, "Universal value function approximators," in *Proc. Int. Conf. Machine Learning*, 2015, pp. 1312–1320.
- [67] T. Schaul, J. Quan, I. Antonoglou, and D. Silver, "Prioritized experience replay," in *Proc. Int. Conf. Learning Representations*, 2016.
- [68] J. Schmidhuber, "A possibility for implementing curiosity and boredom in model-building neural controllers," in *Proc. Int. Conf. Simulation Adaptive Behavior*, 1991, pp. 222–227.
- [69] J. Schmidhuber and R. Huber, "Learning to generate artificial fovea trajectories for target detection," *Int. J. Neural Syst.*, vol. 2, no. 01n02, pp. 125–134, 1991. [Online]. Available: <http://www.worldscientific.com/doi/abs/10.1142/S012906579100011X>
- [70] J. Schulman, S. Levine, P. Abbeel, M. Jordan, and P. Moritz, "Trust region policy optimization," in *Proc. Int. Conf. Machine Learning*, 2015, pp. 1889–1897.
- [71] J. Schulman, P. Moritz, S. Levine, M. Jordan, and P. Abbeel, "High-dimensional continuous control using generalized advantage estimation," in *Proc. Int. Conf. Learning Representations*, 2016.
- [72] D. Silver, G. Lever, N. Heess, T. Degris, D. Wierstra, and M. Riedmiller, "Deterministic policy gradient algorithms," in *Proc. Int. Conf. Machine Learning*, 2014, pp. 387–395.
- [73] D. Silver, A. Huang, C. J. Maddison, A. Guez, L. Sifre, G. van den Driessche, J. Schrittwieser, I. Antonoglou, et al., "Mastering the game of go with deep neural networks and tree search," *Nature*, vol. 529, no. 7587, pp. 484–489, 2016.
- [74] S. Singh, D. Litman, M. Kearns, and M. Walker, "Optimizing dialogue management with reinforcement learning: Experiments with the NJFun system," *J. Artificial Intell. Res.*, vol. 16, pp. 105–133, Feb. 2002.
- [75] B. C. Stadie, S. Levine, and P. Abbeel, "Incentivizing exploration in reinforcement learning with deep predictive models," in *NIPS Workshop on Deep Reinforcement Learning*, 2015.
- [76] A. L. Strehl, L. Li, E. Wiewiora, J. Langford, and M. L. Littman, "PAC model-free reinforcement learning," in *Proc. Int. Conf. Machine Learning*, 2006, pp. 881–888.
- [77] S. Sukhbaatar, A. Szlam, and R. Fergus, "Learning multiagent communication with backpropagation," in *Proc. Neural Information Processing Systems*, 2016, pp. 2244–2252.
- [78] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*. Cambridge, MA: MIT Press, 1998.
- [79] R. S. Sutton, D. Precup, and S. Singh, "Between MDPs and semi-MDPs: A framework for temporal abstraction in reinforcement learning," *Artificial Intell.*, vol. 112, no. 1–2, pp. 181–211, 1999.
- [80] A. Tamar, Y. Wu, G. Thomas, S. Levine, and P. Abbeel, "Value iteration networks," in *Proc. Neural Information Processing Systems*, 2016, pp. 2154–2162.
- [81] G. Tesaro, "Temporal difference learning and TD-gammon," *Commun. ACM*, vol. 38, no. 3, pp. 58–68, 1995.
- [82] C. Tessler, S. Givony, T. Zahavy, D. J. Mankowitz, and S. Mannor, "A deep hierarchical approach to lifelong learning in Minecraft," in *Proc. Association for the Advancement Artificial Intelligence*, 2017, pp. 1553–1561.
- [83] J. N. Tsitsiklis and B. van Roy, "Analysis of temporal-difference learning with function approximation," in *Proc. Neural Information Processing Systems*, 1997, pp. 1075–1081.
- [84] E. Tzeng, C. Devin, J. Hoffman, C. Finn, X. Peng, S. Levine, K. Saenko, and T. Darrell, "Towards adapting deep visuomotor representations from simulated to real environments," in *Workshop Algorithmic Foundations Robotics*, 2016.
- [85] N. Usunier, G. Synnaeve, Z. Lin, and S. Chintala, "Episodic exploration for deep deterministic policies: An application to StarCraft micromanagement tasks," in *Proc. Int. Conf. Learning Representations*, 2017.
- [86] H. van Hasselt, "Double Q-learning," in *Proc. Neural Information Processing Systems*, 2010, pp. 2613–2621.
- [87] H. van Hasselt, A. Guez, and D. Silver, "Deep reinforcement learning with double Q-learning," in *Proc. Association for the Advancement of Artificial Intelligence*, 2016, pp. 2094–2100.
- [88] A. Vezhnevets, V. Mnih, S. Osindero, A. Graves, O. Vinyals, J. Agapiou, and K. Kavukcuoglu, "Strategic attentive writer for learning macro-actions," in *Proc. Neural Information Processing Systems*, 2016, pp. 3486–3494.
- [89] A. S. Vezhnevets, S. Osindero, T. Schaul, N. Heess, M. Jaderberg, D. Silver, and K. Kavukcuoglu, "FeUdal networks for hierarchical reinforcement learning," in *Proc. Int. Conf. Machine Learning*, 2017, pp. 3540–3549.
- [90] N. Wahlström, T. B. Schön, and M. P. Deisenroth, "Learning deep dynamical models from image pixels," in *Proc. IFAC Symp. System Identification*, 2015, pp. 1059–1064.
- [91] N. Wahlström, T. B. Schön, and M. P. Deisenroth, "From pixels to torques: policy learning with deep dynamical models," in *ICML Workshop on Deep Learning*, 2015.
- [92] Z. Wang, N. de Freitas, and M. Lanctot, "Dueling network architectures for deep reinforcement learning," in *Proc. Int. Conf. Learning Representations*, 2016.
- [93] Z. Wang, V. Bapst, N. Heess, V. Mnih, R. Munos, K. Kavukcuoglu, and N. de Freitas, "Sample efficient actor-critic with experience replay," in *Proc. Int. Conf. Learning Representations*, 2017.
- [94] C. J. C. H. Watkins and P. Dayan, "Q-learning," *Mach. Learning*, vol. 8, no. 3–4, pp. 279–292, 1992.
- [95] M. Watter, J. Springenberg, J. Boedecker, and M. Riedmiller, "Embed to control: A locally linear latent dynamics model for control from raw images," in *Proc. Neural Information Processing Systems*, 2015, pp. 2746–2754.
- [96] D. Wierstra, A. Förster, J. Peters, and J. Schmidhuber, "Recurrent policy gradients," *Logic J. IGPL*, vol. 18, no. 5, pp. 620–634, 2010.
- [97] R. J. Williams, "Simple statistical gradient-following algorithms for connectionist reinforcement learning," *Mach. Learning*, vol. 8, no. 3–4, pp. 229–256, 1992.
- [98] M. Wulfmeier, P. Ondruska, and I. Posner, "Maximum entropy deep inverse reinforcement learning," in *NIPS Workshop on Deep Reinforcement Learning*, 2015.
- [99] K. Xu, J. Ba, R. Kiros, K. Cho, A. C. Courville, R. Salakhutdinov, R. S. Zemel, and Y. Bengio, "Show, attend and tell: Neural image caption generation with visual attention," in *Proc. Int. Conf. Machine Learning*, 2015, pp. 2048–2057.
- [100] Y. Zhu, R. Mottaghi, E. Kolve, J. J. Lim, A. Gupta, L. Fei-Fei, and A. Farhadi, "Target-driven visual navigation in indoor scenes using deep reinforcement learning," in *Proc. IEEE Int. Conf. Robotics and Automation*, 2017, pp. 3357–3364.

Seunghoon Hong, Suha Kwak, and Bohyung Han

Weakly Supervised Learning with Deep Convolutional Neural Networks for Semantic Segmentation

Understanding semantic layout of images with minimum human supervision



©STOCKPHOTO.COM/ZAPP2PHOTO

Semantic segmentation is a popular visual recognition task whose goal is to estimate pixel-level object class labels in images. This problem has been recently handled by deep convolutional neural networks (DCNNs), and the state-of-the-art techniques achieve impressive records on public benchmark data sets. However, learning DCNNs demand a large number of annotated training data while segmentation annotations in existing data sets are significantly limited in terms of both quantity and diversity due to the heavy annotation cost. Weakly supervised approaches tackle this issue by leveraging weak annotations such as image-level labels and bounding boxes, which are either readily available in existing large-scale data sets for image classification and object detection or easily obtained thanks to their low annotation costs. The main challenge in weakly supervised semantic segmentation then is the incomplete annotations that miss accurate object boundary information required to learn segmentation. This article provides a comprehensive overview of weakly supervised approaches for semantic segmentation. Specifically, we describe how the approaches overcome the limitations and discuss research directions worthy of investigation to improve performance.

Introduction

Over the past few years, we observed significant advances in visual recognition techniques, which are particularly attributed to the recent development of DCNNs [25]. DCNNs learn a feature hierarchy directly from raw data, and the learned features are, in general, richer and more powerful than manually designed ones that had been widely used before the era of deep learning. Also, DCNNs can further improve their capacity by optimizing their decision makers (e.g., classifier) and the feature extractors jointly in an end-to-end manner. These potentials of DCNNs are realized recently thanks to the development of novel learning algorithms, large-scale training data sets, and computer hardware supporting massively parallel computation. The success story of DCNNs in visual recognition includes image classifiers surpassing human-level performance [14], [15], object detectors meeting both excellent accuracy and real-time speed [30], [41], and other models

Digital Object Identifier 10.1109/MSP.2017.2742558
Date of publication: 13 November 2017

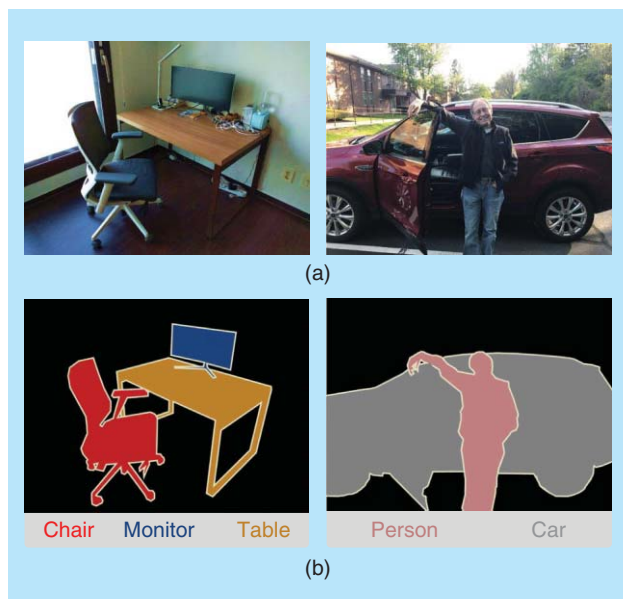


FIGURE 1. (a) Example images and (b) their semantic segmentation ground truths. Compared to image-level class labels and instance bounding boxes, the pixel-wise labels provide more dense and comprehensive description of image content.

outperforming previous state of the art in many other computer vision tasks such as human pose estimation [7], [53], face recognition [45], [48], and so on.

The great success of DCNNs also leads to another challenging visual recognition task called *semantic segmentation*. The goal of semantic segmentation is to assign semantic class labels to every pixel in images, where semantic classes typically include a diverse range of object categories (e.g., person, dog, bus, bike) and background components (e.g., sky, road, building, mountain). As illustrated in Figure 1, the result of semantic segmentation is pixel-level masks of each semantic class, which describe image content more comprehensively than image-level class labels given by image classification and object bounding boxes predicted by object detection.

Such a detailed image description is essential to build an intelligent system that is as competitive as human visual cognitive ability. Also, due to the emergence of computer vision applications that require comprehensive understanding of visual input, such as medical image analysis, autonomous driving, robotics, and human computer interaction, the demand for accurate semantic segmentation algorithms has been increasing continually.

In return for its detailed high-level prediction capability, however, semantic segmentation involves several critical challenges to be resolved. One has significant appearance variations of semantic classes caused by large intraclass variation, occlusion, deformation, illumination change, and viewpoint variation that are commonly observed in real-world images. Being invariant to these factors is challenging especially for semantic segmentation that has to predict class labels in a pixel level. Also, semantic segmentation must consider structured dependency among class labels of pixels during prediction (i.e., assigning the same class labels to spatially adjacent pixels), but

this constraint in semantic segmentation is difficult to handle in practice due to a prohibitively large search space for possible segmentation results.

Fortunately, DCNNs provide solutions to the aforementioned issues. The rich hierarchical feature representations of DCNNs is robust against significant appearance variations. Also, several architectures of DCNNs have been proposed to predict structured output naturally by considering the structured dependency either implicitly [5], [32], [33] or explicitly [27], [31], [57]. Furthermore, during their training, the feature representation and the structured prediction of the networks are jointly optimized in end-to-end manners. All of these factors are critical to overcome the previously mentioned difficulties in semantic segmentation. Consequently, DCNNs have achieved substantial progress in semantic segmentation, improving previous records based on handcrafted features significantly on public benchmarks including PASCAL Visual Object Classes (VOC) [11].

Despite the great success of DCNNs on public benchmarks, there still remains a critical obstacle in the way of their applications to semantic segmentation in an uncontrolled and realistic environment: lack of annotated training images. It has been known that, since a DCNN has a large number of tunable parameters, it accordingly demands a large number of annotated data for training models with good generalization performance. For semantic segmentation, however, collecting large-scale annotations is significantly labor intensive because people have to manually draw pixel-level masks for every semantic categories per image to carry out the annotation. Also, collecting annotations for semantic segmentation is practically limited for some applications. An example is medical image analysis, for which domain expert knowledge is essential to accurate annotations. For these reasons, existing data sets often suffer from lack of annotated examples and class diversity, and it is also difficult to maintain good quality of segmentation annotations in terms of both accuracy and consistency. Therefore, it is not straightforward to extend the existing models based on DCNNs to cover more classes while maintaining high accuracy.

To resolve the issues related to training data collection and make semantic segmentation more scalable and generally applicable, researchers are interested in weakly supervised learning. In this setting, the objective is to train a robust model for semantic segmentation using the annotations that are much weaker than pixel-wise labels. Examples of weak supervision for semantic segmentation are illustrated in Figure 2. The clear advantage of weak annotations is that they are much cheaper to obtain than the standard segmentation annotations. Some types of weak annotations such as image-level class labels and bounding boxes are even readily available in existing large-scale data sets [10], [29] for image classification and object detection. Thus, with such weakly annotated images, we can greatly enlarge or easily create training data sets for semantic segmentation. The main issue of weakly supervised semantic segmentation is then how to fill the gap between the level of supervision and that of prediction. The supervisory signal

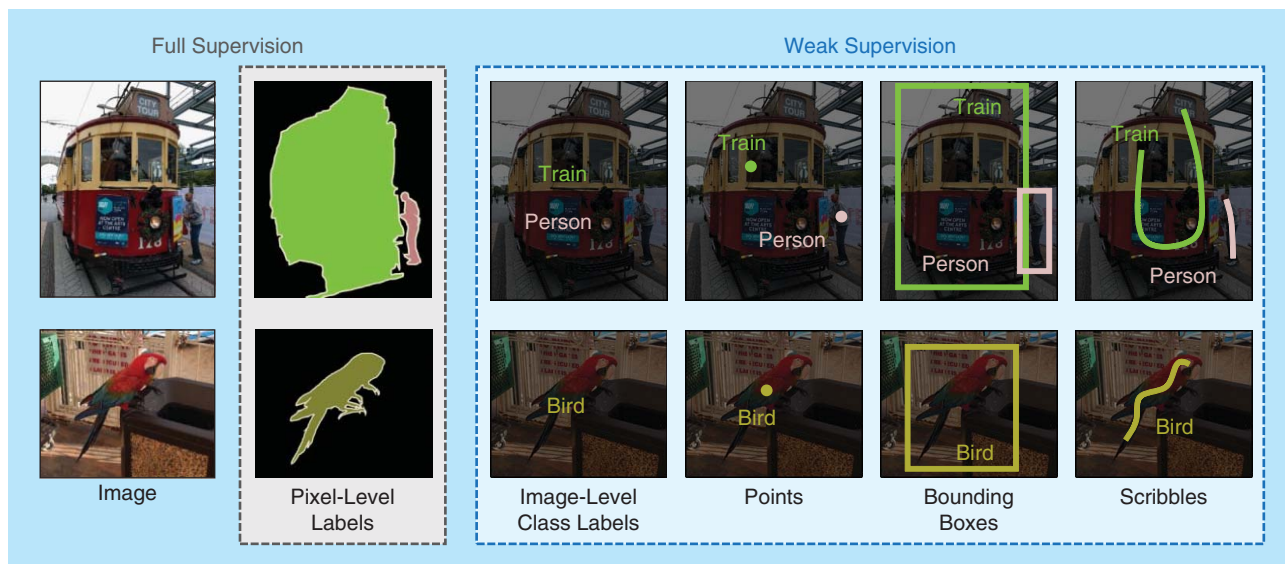


FIGURE 2. Illustrations of various weak annotations employed for weakly supervised semantic segmentation.

indicating object location and shape is critical in learning to predict segmentation masks but is partly or totally absent in weak supervision. The success of weakly supervised approaches depends heavily on the way to compensate the missing information during training.

The purpose of this article is to provide an introduction to weakly supervised semantic segmentation and a thorough review of recent approaches in this line of research. In particular, we narrow our focus to approaches based on DCNNs. There exist weakly supervised approaches proposed before the era of deep learning [50]–[52], [56]. They attempt to associate pixels with image-level class labels by first computing region-based classification scores and further refining them using similarities between local image regions based on various handcrafted visual cues. However, their performance is, in general, limited due to lack of robust appearance models. On the other hand, DCNNs provide more natural ways to associate pixels with image labels and more robust feature representations for appearance modeling. In addition, DCNNs are flexible enough to integrate various types of weak supervision and additional information that may be useful to improve segmentation performance.

DCNNs for semantic segmentation

This section provides an overview of approaches for semantic segmentation based on DCNNs. The objective of semantic segmentation is to infer semantic class labels of every pixel in an image. To achieve this goal, many existing approaches pose the task as dense local area classification and modify a DCNN designed for image classification to predict class scores for every local area in an input image.

The most popular choice of network architecture in this direction is a fully convolutional network (FCN) [32]. The FCN is based on a DCNN pretrained for large-scale image classification, but its architecture is fully convolutional (i.e.,

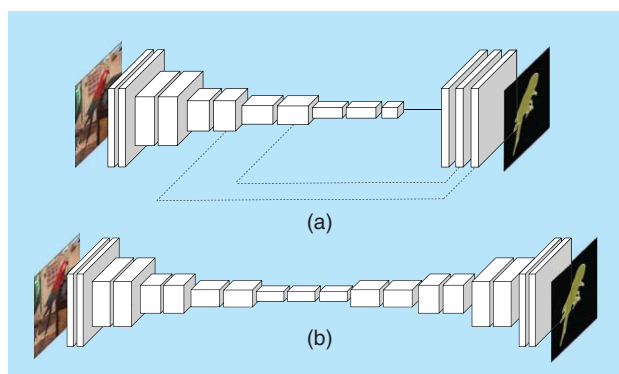


FIGURE 3. Illustrations of popular DCNN architectures used for semantic segmentation. (a) A fully convolutional network makes all network components fully convolutional, thus converts a CNN trained for image classification to produce class scores on local image regions. Optional skip-connections from lower layers (dashed lines) are used to reconstruct spatial information lost by spatial poolings. (b) The network architecture of a deep convolutional encoder-decoder network. On top of the convolutional network, a stack of deconvolutional layers are used to reconstruct fine object segmentation masks using many network parameters.

having no fully connected layer) as it interprets fully connected layers of the classification DCNN as 1×1 convolution filters so that it can handle input images with arbitrary sizes. The output of the network then has a form of a class score map over the image. Since the output score map is low resolution, due to multiple pooling operations in the network, a single deconvolution layer is employed on top of the class score map to enlarge the size of the output map to that of the input image. The overall network architecture of the FCN is illustrated in Figure 3(a). Since the output of the network corresponds to pixel-wise class prediction scores, it is possible to learn the entire model parameters in an end-to-end manner by computing classification loss on every pixel location using pixel-level ground-truth labels.

Also, it provides an efficient inference mechanism that directly produces pixel-wise class predictions with a single forward pass of an input image regardless of its size.

Later approaches based on FCN-style architectures improve prediction accuracy by considering low-level image structures and building a deeper network architecture. One popular way to refine the predicted labels is to apply postprocessing based on a graphical model such as fully connected conditional random field (CRF) [23]. It regularizes the predicted labels to coincide with visually similar pixels, thus it encourages the inferred labels to preserve underlying image structures such as object boundaries. Specifically, Chen et al. [5] propose to integrate prediction results of the network with fully connected CRF and estimate the final pixel-wise labels by solving an optimization problem based on the CRF model. The approach is improved in [57] by decomposing CRF with several steps of differentiable operations and modeling each step with an individual neural network, where all networks are nicely integrated into a single DCNN for end-to-end training. The idea of integrating CRF is further extended by modeling pairwise relationship between output units to consider spatial context between semantic classes [27], [31]. Apart from CRF, there have been approaches to improve the performance by considering multiscale predictions. For example, [5] employs a set of convolution filters sampled from multiple scales to capture objects with variable sizes, while [6] performs semantic segmentation in multiple scales and aggregate the results via weighted summation, where the weights are also predicted for each scale by an independent model.

On the other hand, some approaches propose building a deep encoder-decoder network for precise per-pixel class prediction [33], [42]. A typical architecture of the deep encoder-decoder network is illustrated in Figure 3(b). Contrary to the FCN-style architecture that has a single up-sampling layer, they employ a deep decoder on top of the encoder output to recover the original input image resolution. Specifically, Noh et al. [33] propose a deconvolution network, which has a symmetric architecture of encoder and decoder, where the decoder is implemented by stacks of deconvolution layers and unpooling operations. The similar architecture is employed in [42] together with an efficient data augmentation technique for the task of biomedical image segmentation.

These approaches have been successful in semantic segmentation even on real-world images [11], [29] when a sufficiently large number of training images with pixel-wise annotations are available. However, such annotations require a tremendous amount of labeling cost and is available only in a few data sets with a limited number of semantic categories. To resolve the difficulties in training data collection and design more flexible and scalable models for semantic segmentation, approaches based on weakly supervised semantic segmentation have been proposed to utilize much weaker labels than pixel-wise ones.

Weakly supervised semantic segmentation

This section introduces weakly supervised semantic segmentation and discusses relevant approaches based on DCNNs.

The goal of weakly supervised semantic segmentation is to leverage weak annotations instead of pixel-wise ones to learn models for semantic segmentation. Weak labels for semantic segmentation include, but is not limited to, image-level class label, bounding box, scribble, and point supervision as illustrated in Figure 2. These labels are easier to collect than pixel-wise labels as they require much less annotation cost. For example, the annotation time of image-level class labels is that of only one-tenth of pixel-level segmentation annotations [4]. Thus, one can easily build a weakly annotated image data set for diverse semantic categories on a large scale, and such a training data set will, accordingly, allow to learn a model for semantic segmentation in the wild.

The main challenge in weakly supervised semantic segmentation is that the weak labels provide only a part of the supervision required for semantic segmentation. For example, none of weak labels presented in Figure 2 provide information about object shape, which is a critical evidence required to learn a model predicting segmentation masks. Therefore, to train a model for segmentation with incomplete supervision in weakly labeled data, the latent per-pixel ground truth as well as the model parameters should be jointly estimated during training. In the following sections, we introduce various types of weak labels employed in the literature and discuss the related approaches in detail.

Image-level class label

Image-level class label is the simplest form of weak supervision for semantic segmentation as it indicates only presence or absence of a semantic entity in an image. Because it requires the least amount of human annotation cost and is already available in existing large-scale data sets such as ImageNet [10], image-level class label has been most extensively exploited in weakly supervised semantic segmentation. However, learning segmentation networks from only image-level class labels is very challenging because spatial information about target objects are missing.

Some existing approaches resolve this issue by considering pixel-level labels as latent variables, and optimize parameters of segmentation network jointly with the latent pixel-level labels. Specifically, they consider outputs from a convolution layer of DCNN as confidence scores for latent per-pixel labels. Since the image-level label is the only available supervision in this weakly supervised setting, they aggregate the output scores over all pixels using a global pooling operation (e.g., max pooling or average pooling) to generate image-level class score. Then the network is trained to maximize image classification performance using image-level labels as ground truth. Within this framework, Pathak et al. [37] formulate the task as a multiple instance learning problem, where a global max pooling operation is applied to enforce the constraint that each image should contain at least one pixel corresponding to the positive class. With the same motivation, Papandreou et al. [35] adopt a recursive refinement procedure based on expectation-maximization, where the latent pixel-level labels are predicted by the learned

model and, in turn, used to update the model as new ground-truth annotations.

Since the supervision by image-level class labels is too coarse for segmentation, the quality of the obtained results is often not satisfactory in the above approaches. This issue has been tackled by incorporating additional cues to simulate supervision for object location and shape. To incorporate localization cue, techniques for discriminative localization based on DCNN are employed [58]. By carefully investigating the contribution of each hidden units to the output class score of the network, one can identify coarse locations of discriminative parts of each class in an image. Then the outputs from discriminative localization are used to choose seeds indicating a position on the area of a semantic class, and the seeds are expanded to neighboring pixels to estimate pixel-wise area of the class [22], [34], [46]. To incorporate shape information, superpixels are utilized as units for label assignment [24], [38]. A superpixel is a group of neighboring pixels that are similar in visual appearance (e.g., color) and is often obtained by clustering pixels based on low-level visual similarity. Superpixels are beneficial by encoding shape information, as they naturally reflect a low-level image structure such as object boundary. Pinheiro and Collobert [38] employ superpixels to smooth pixel-wise class labels within each superpixel as postprocessing. Kwak et al. [24] exploit superpixels as the layout of a pooling operation in the DCNN. Another popular approach to refining pixel-level prediction is applying the fully connected CRF as in the case of fully supervised approaches. CRF propagates labels between neighboring pixels and refines the prediction from the model to cover better object extent and shape.

Although these approaches are able to roughly localize objects, they often fail to infer accurate pixel-wise labels as they tend to focus only on small discriminative parts (e.g., the head of an animal) instead of the whole body of an object. It is because their objective during training is to minimize a classification loss, which is easier to achieve by considering small parts that can be well distinguished from other categories. Indeed, estimating pixel-wise labels only from image-wise labels is a significantly ill-posed problem. To reduce the gap between coarse image-level labels and fine per-pixel labels, the approaches introduced in the next sections incorporate additional weak annotations together with image-level class labels, utilize stronger but still weaker annotations than pixel-level labels, or adopt additional data sources that are also weakly annotated.

Prior knowledge

One way to compensate for the lack of details in image-level class labels is to exploit extra prior knowledge about the segmentation target. Pathak et al. [36] proposed the employment of prior knowledge about object size, which roughly provides information about how much area of an image is occupied by the target object. In this approach, a user is asked to provide

one-bit information for each object class per image, which indicates whether the size of the object occupies more than 10% of the image area or not. This information is incorporated during training by enforcing that the output score map of the model satisfies the size constraint given by user.

Objectness, which is also known as *saliency*, is another form of the prior knowledge that can provide information about object extents [4], [20], [34], [38]. It is a real-valued score assigned to each pixel or local image region to indicate whether the pixel or region belongs to an actual object, regardless of the semantic class. Since it is class agnostic, the objectness typically covers a larger object area including nondiscriminative parts, thus is useful to compensate the limitation of weakly supervised approaches that favor only small discriminative

parts. For this reason, Pinheiro and Collobert [38] adopt an off-the-shelf algorithm that returns a set of region proposals with associated objectness scores [2]. The per-pixel objectness score is then computed by aggregating the scores associated with proposals and, in turn, is used to weight the class score on corresponding pixel locations. In addition to image-level labels, the objectness score has also been applied to different types of weak labels, such as the point [4] and bounding box [20], to impose

larger weights on potential object areas. On the other hand, Wei et al. [54] compute the saliency map on images associated with a single class and use the obtained saliency masks to initialize the network for pixel-wise classification. Similarly, Oh et al. [34] generate saliency masks using a model trained for foreground segmentation, and they assign class labels on the generated saliency mask by propagating the class label seeds obtained by the discriminative localization technique.

Point supervision

An instance-wise point, which roughly indicates the center location of an object, is the simplest form of weak annotations that provide object location information, since it can be obtained by a single user click per object. Bearman et al. propose in [4] to employ a combined loss for both classification and localization, where the latter is used to ensure that a model predicts correct labels on the pixel localized by the point annotation. Since the point supervision is extremely sparse, in this work, an additional prior knowledge on objectness is further employed to estimate foreground regions that well-cover the object.

Bounding box

Although the point supervision provides coarse locations of semantic classes, the information about areas covered by the classes is still missing. A bounding box annotation can offer such information by indicating a rectangular area that tightly covers the entire object region. Also, its annotation cost is still cheaper than that of pixel-level segmentation annotation.

Existing approaches [9], [20], [35] that have been proposed to infer pixel-wise labels given bounding boxes

The main challenge in weakly supervised semantic segmentation then is the incomplete annotations that miss accurate object boundary information required to learn segmentation.

formulate semantic segmentation as automatic foreground/background segmentation within each bounding box area. To this end, Papandreou et al. [35] incorporate techniques developed for interactive segmentation (i.e., GrabCut [43]) to estimate foreground pixels within the box, where pixels inside and outside the box are considered as initial seeds for the foreground and background, respectively. Then, a model for semantic segmentation is trained using the estimated segmentation masks as ground truth. On the other hand, instead of the direct estimation of pixel-wise labels from the bounding box, Dai et al. [9] exploit off-the-shelf region proposals [2]. As the adopted region proposal algorithm provides a candidate set of masks that potentially correspond to an object in a box, the problem is reduced to choosing the best region proposal among the ones sufficiently overlapped with each bounding box. To this end, an iterative refinement procedure similar to [35] is adopted for training, where the model is trained by pixel-wise annotations computed from the selected region proposals and the learned model is, in turn, used to refine the proposal selection. Khoreva et al. [20] improve the label prediction within the box by using the objectness prior [2] as the initial foreground seeds for GrabCut segmentation and applying a recursive refinement procedure as in [9].

Since a bounding box provides incomplete yet sufficiently strong supervision for object location and area missing in image-level class labels, all of the approaches based on bounding box annotations substantially improve the performance over the ones trained only with the image-level labels. Moreover, they are even competitive to the fully supervised counterpart.

Scribble

A scribble is a line in an arbitrary form obtained by a single user stroke, and as another form of weak annotation, it provides sparse information about object location and extent. One can consider a scribble as a middle ground of point- and box-level annotations since a point is a special case of a scribble (i.e., a scribble with zero length) and a scribble roughly indicates object area as a bounding box does. Scribbles provide not only a user-friendly way for annotation but also an easier way to localize objects in arbitrary shapes. Since the scribble covers only a partial area of a semantic entity, the inference of pixel-wise labels is reduced to propagating the annotated labels to unmarked pixels. Lin et al. [26] formulate the label propagation as an optimization problem based on a graphical model, where vertices of the graph are superpixels of each image. Training is performed by alternating label estimation and model parameter learning, where the model is trained under the supervision of superpixel labels and, in turn, used to update the labels as a part of the graph-based optimization procedure.

Microuser annotation

As described previously, there exists a tradeoff between annotation cost and the amount of supervision in the selection of

weak labels; complicated labels usually provide stronger supervision for segmentation while increasing the human annotation cost. To obtain cost-effective labels from a human, some approaches propose utilizing microuser annotation. In these approaches, a model presents to users multiple candidate masks inferred from an image and asks them to choose the best mask among the candidates. This process makes the annotation task intuitive and efficient because it needs a simple user verification by a single click to obtain dense segmentation masks. The success of these approaches is thus heavily dependent on generating diverse and high-quality segmentation masks.

Motivated by this, Saleh et al. [44] generate multiple foreground masks by inferring multiple CRF solutions that are diverse and have low energy at the same time. Kolesnikov and Lampert [21] compute candidate masks by clustering image regions into multiple groups, where each region is described by a feature vector computed by a DCNN. In both of the aforementioned approaches, users are asked to select the best mask among the predicted multiple diverse candidates, and the selected masks are considered as strong supervision for learning a semantic segmentation model.

Natural language description

A natural language description of an image can be used as an annotation since it provides comprehensive information of the image including object attributes, relations between objects, scene context, and so on. Also, such a description is readily available for a large number of images found on photo-sharing sites like Flickr.

Lin et al. [28] exploits natural language image description as weak annotation. Specifically, they propose performing semantic segmentation by aligning the semantic structures of image description and image regions. To this end, both an image and its description are parsed into tree structures through independent procedures; the tree for the description follows the grammatical structure of sentences, and that of the image, obtained by a recurrent neural network, defines a hierarchical structure of image segments discovered by a semantic segmentation model. The segmentation model is trained to align the two parsing trees.

Additional data source

Some of the aforementioned annotation types—point, scribble, and microuser annotation—are not readily available in existing large-scale data sets and demand a certain level of human intervention. Although such types of annotations are much easier to obtain than pixel-wise labels, their demands for human intervention is not desirable when considering that the main motivation of weakly supervised learning is to reduce human intervention required for training. To incorporate stronger supervision without extra human labeling effort, some approaches propose the exploitation of an additional source of data, which are freely available in other data sets or different data domains.

The objective of semantic segmentation is to infer semantic class labels of every pixel in an image.

Hong et al. [17] exploit segmentation annotations of semantic classes, irrelevant to those we want to segment, in a transfer learning framework. The motivation is that the knowledge required to estimate object shapes can be transferred from a group of categories to another one. For example, a person capable of segmenting a dog can generalize his or her knowledge to segment other animals, such as a cat and a horse, as they are similar in overall shape (e.g., four legs, a head, and tails). This idea is implemented by a deep encoder-decoder network, where the encoder network estimates coarse locations of objects and the decoder network generates segmentation masks corresponding to the object indicated by the result of the encoder network. It is shown empirically in [17] that segmentation knowledge of irrelevant classes can be successfully transferred to that of target classes for which only image-level class labels are available.

Other approaches [18], [49] in this category make use of web videos as additional data. Motion in video is a powerful cue to separate foregrounds from surrounding background since object and background typically exhibit distinctive motion patterns due to their different dynamics and three-dimensional positions. For this reason, the approaches take advantage of densely computed video motions, which are also called *optical flows*. They first conduct semantic segmentation on individual video frames using DCNN learned with weakly annotated images and enhance the quality of the segmentation results by taking dense motions into account. The video segmentation results are then used as synthesized supervision to train a model for semantic segmentation. The main challenge in this direction is to collect relevant video clips from web repositories (e.g., YouTube) with no human intervention. Hong et al. [18] tackle this issue by a fully automatic video retrieval algorithm that crawls videos from a web repository using the class labels as search key words and refines the search results by selecting only relevant intervals from the videos using a DCNN-based classifier learned from weak labels.

Semisupervised semantic segmentation

The problem of semisupervised semantic segmentation takes the middle ground between fully and weakly supervised semantic segmentation. In semisupervised learning of semantic segmentation, the training data set is composed of both weakly and fully annotated examples, which is different from the traditional semisupervised learning setting where training data consist of unlabeled and labeled examples. As both semi- and weakly supervised approaches share the same motivation—reducing the significant burden for full annotation—it is assumed that the number of fully annotated examples is limited to only a small portion of training data. Thus, the main challenge of semisupervised semantic segmentation is to train a model using a set of unbalanced and heterogeneous annotations. As in weakly supervised setting, various types of labels have been employed

as weak supervision, such as image-level class labels [16], [35], bounding boxes [9], [35], and scribbles [26].

Given these annotations, the simplest and most popular approach to semisupervised semantic segmentation is to directly apply the identical model used in the weakly supervised case to semisupervised semantic segmentation [9], [26], [35]. In these approaches, the model is trained with both of estimated and ground-truth pixel-wise labels associated with weakly and fully annotated data, respectively. Since the full annotations provide clean and reliable update signals that guide the learning process of the DCNN, the semisupervised approaches have, in general, shown better performance than their weakly supervised counterparts.

Hong et al. [16] design a more sophisticated DCNN architecture based on a decoupled deep encoder-decoder network, specialized for semisupervised semantic segmentation. The main idea is that semantic segmentation can be decoupled into two subproblems: classification and (class-agnostic) segmentation. Following this idea, the network architecture is accordingly decoupled to be a concatenation of two separate networks, one for classification and the other for segmentation, which are trained independently; the classification network is learned from many weak annotations (i.e., image-level class

labels) and the segmentation network from segmentation annotations for all semantic categories with no distinction between them. By decomposing the problem into the two subproblems requiring different degrees of supervision, the performance of semisupervised semantic segmentation can be significantly improved even with a very small number of full annotations.

Empirical analysis

This section provides empirical analysis of weakly and semisupervised semantic segmentation algorithms presented in the previous sections. There exist several benchmarks for evaluation of the algorithms. PASCAL VOC 2012 [11] is the most popular, where each pixel in image is annotated as one of the 20 object classes or background. This benchmark originally consisted of 1,464 images for training and 1,449 for validation, but most approaches tested on it are learned with 10,582 images augmented with the semantic boundaries data set [13]. Another popular one is MS-COCO [29], which is composed of 82,783 images for training and 40,504 for validation with annotations for 80 object categories including those of PASCAL VOC 2012. Cityscapes [8] is a benchmark for semantic segmentation of urban street scenes. It focuses on 30 semantic entities observed frequently from streets and provides 5,000 images with fine segmentation annotations and 20,000 images with coarse ones. ADE20K [59] is a recently released benchmark that provides comprehensive segmentation annotations for a large-scale image collection. It contains 20,210 training images and 2,000 validation images, which are annotated with 2,693 object categories and 476 part types.

To resolve the difficulties in training data collection and design more flexible and scalable models for semantic segmentation, approaches based on weakly supervised semantic segmentation have been proposed to utilize much weaker labels than pixel-wise ones.

In this article, we employ the PASCAL VOC 2012 benchmark for evaluation since it has the largest record of reported performance, thus allowing fair and comprehensive comparisons among various methods. The evaluation metric is mean intersection-over-union (mIoU) between ground-truth and predicted segmentation results. We present scores reported in the original papers, for both validation and test splits. For an approach with multiple variations in model architecture, only the score of the best model is presented for the sake of brevity. For an approach adopting various types of weak supervision (e.g., [4], [35], [36], and [38]), we present multiple scores corresponding to the individual supervision types. In addition to annotations for training, we also report the type of extra information adopted by each method if exists, as such information may introduce additional supervision that are not available from the PASCAL VOC 2012 training data.

Table 1 summarizes comparison results for the weakly supervised approaches. As discussed in the section “Weakly Supervised Semantic Segmentation,” they are categorized by

the type of supervision employed for training. Note that it is not appropriate to compare different types of weak annotation directly since the models employed in each method have different configurations and capacities. However, the general performance trend across various approaches given in the table clearly demonstrates the impact of supervision levels and benefits of using extra information.

Approaches based only on image-level class labels perform poorly in general, as shown in Table 1. As described in the section “Weakly Supervised Semantic Segmentation,” it is mainly because the discriminative learning objective employed in weakly supervised approaches tends to focus on small discriminative parts. The performance is improved by adopting additional cues such as discriminative localization and underlying low-level image structures, since they provide useful information to regularize the prediction during training. Also, it is clearly observed that prior knowledge such as objectness generally improves the performance, as it helps to estimate a better object extent by injecting class-agnostic objectness likelihood in a pixel level. Increasing the strength

Table 1. The comparison results of weakly supervised semantic segmentation algorithms on the PASCAL VOC 2012 data set.

Supervision	Method	Extra Information	mIoU (val)	mIoU (test)
Image-level label	MIL-FCN [37]	—	25.1	25.7
	WSSL [35]	—	38.2	39.6
	CCNN [36]	—	35.3	36.4
	AugFeed [40]	—	52.7	52.6
	WTP [4]	—	29.8	—
	MIL-SP [38]	Superpixel [12] [†]	36.6	35.8
	SPN [24]	Superpixel [60]	50.3	46.9
	SEC [22]	Localization [58] [†]	50.7	51.7
	DCSM [46]	Localization [47] [†]	44.1	45.1
Prior knowledge	CCNN [36]	Object size	45.1	42.4
	MIL-SP [38]	Objectness [2]	42.6	40.6
	AugFeed [40]	Objectness [2]	54.3	55.5
	STC [54]	Objectness [2]	49.8	51.2
	Saliency [34]	Saliency	55.7	56.7
Point supervision	WTP [4]	Point + objectness [1]	43.8	—
Bounding box	WSSL [35]	Bounding box	60.6	62.2
	Boxsup [9]	Bounding box + objectness [2]	62.0	64.6
	SDI [20]	Bounding box + objectness [2]	65.7	67.5
Scribble	ScribbleSup [26]	Scribble	63.1	—
Microannotation	CheckMask [44]	User feedback	51.5	52.9
	MicroAnno [21]	User feedback	51.9	53.2
Additional data	TransferNet [17]	Exclusive segmentation mask [29]	52.1	51.2
	MCNN [49]	Web videos [39]	38.1	39.8
	CrawlSeg [18]	Web videos (YouTube)	58.1	58.7

[†]Indicates extra information obtained without additional supervision.

of supervision in training annotations also improves the segmentation quality in general. Specifically, algorithms using bounding box annotations show noticeably improved performance compared to the ones based only on an image-level label. It is because the bounding box provides supervisory signals for object location and area at the same time, which are critical to train a model for semantic segmentation, and totally missing in image-level labels. Also, if used in the right way, an additional data source like web videos provides useful information for segmentation without additional human effort since the supervision for segmentation can be reliably synthesized and effectively transferred from the domain of the additional data source.

In addition to the weakly supervised approaches, we quantify and compare the results of semisupervised methods. As shown in Table 2, employing segmentation annotations for training effectively improves performance even when their number is very small. It is because the segmentation annotations provide strong and clear guidance to learn semantic segmentation models reliably from weak labels. Similar to the weakly supervised cases, the performance of semisupervised approaches depends on the supervision strength of the employed weak annotations. Also, increasing the amount of segmentation annotations naturally improves the performance.

Other applications

This survey article has focused on semantic segmentation algorithms for RGB images with various objects. However, weakly supervised semantic segmentation can be naturally applied to other tasks as well, and this section introduces a few examples.

Medical image analysis would be an important application of weakly supervised semantic segmentation. Semantic segmentation of medical images (e.g., segmentation of cancer lesions) plays an important role in disease recognition and diagnosis, but collecting segmentation labels for medical images is particularly expensive because annotators must be domain experts. Thus, weakly supervised approaches have attracted much attention even before the era of deep learning [55]. Recently Jia et al. [19] proposed a DCNN-based approach for histopathology cancer image segmentation. Similar to [36], a roughly estimated size of a cancerous region is used as weak supervision in addition to image-level class labels since it is less costly to obtain relaxed information compared to pixel-level class labels.

Weakly supervised semantic segmentation has been also applied for autonomous driving. Barnes et al. [3] present a weakly

supervised framework to learn a DCNN segmenting a path proposal and obstacles from a road scene. The path proposal means the pixel-level area of a route that the vehicle would take, and manual annotation of such an area is expensive as it is in the typical semantic segmentation setting. In [3], segmentation annotations of those entities are generated with no human intervention other than driver behavior: a large volume of video data is first recorded by a camera and a lidar sensor mounted on the vehicle, then path proposals are annotated by projecting the future path of the vehicle into each video frame and obstacle areas are derived from lidar scanning results. Finally, the generated annotations are used to train a DCNN for segmentation.

Summary and discussion

This article provided a comprehensive review of weakly supervised approaches for semantic segmentation based on DCNNs. Semantic segmentation aims to estimate pixel-level areas of semantic categories, and its results enable many interesting applications that require fine-grained interpretation of an image. Following their great success in other visual recognition tasks, DCNNs also have shown impressive performance on public benchmarks for semantic segmentation. However, even with DCNNs, we still have a long way to go in achieving semantic segmentation in an uncontrolled and realistic environment. Because of the data-hungry nature of DCNNs and the lack of segmentation annotations for training, fully supervised DCNNs can handle only a small number of semantic classes that are defined in the existing training data sets. Weakly supervised approaches tackle this issue by leveraging readily available or easily obtainable weak labels instead of segmentation masks. As summarized in the section “Empirical Analysis,” the records achieved by recent weakly supervised approaches are impressive, especially when considering the fact that no pixel-level supervision is provided to train them. However, there is still a certain gap between the performance of fully supervised approaches and that of weakly supervised ones.

Here we suggest a few directions worth investigating for further improvement of weakly supervised semantic segmentation. The first is to get help from unsupervised computer-vision techniques. Object shape information is essential to learn semantic segmentation but absent in weak annotations, and we believe that such missing information can be compensated by the unsupervised techniques. For example, superpixels and region proposals well preserve underlying image structures including object boundaries, so they allow us to bypass explicit shape estimation during inference

Table 2. The comparison results of semisupervised semantic segmentation algorithms on the PASCAL VOC 2012 data set.

Method	Weak Supervision	Number of Segmentation Annotations	mIoU (val)	mIoU (test)
WSSL [35]	Image-level label	0.5K (5% of the training set)	56.9	—
DecoupledNet [16]	Image-level label	0.5K (5% of the training set)	62.1	62.5
WSSL [35]	Bounding box	1.4K (13% of the training set)	65.1	66.6
Boxsup [9]	Bounding box	1.4K (13% of the training set)	63.5	66.2

through selecting those belonging to target semantic categories [9], [18], [24], [35]. Also, in [18] and [49], optical flows help synthesize more accurate segmentation annotations from videos by propagating class-localization information between consecutive frames along with motion. It will be an interesting approach to weakly supervised semantic segmentation to newly introduce other unsupervised techniques that can recoup the gap between the level of supervision and that of prediction.

Another direction is transfer learning. Existing benchmarks for semantic segmentation [11], [29] provide segmentation annotations, and it is obviously desirable to exploit the existing segmentation annotations although they are given for only a small number of semantic categories that may not be relevant to target categories we would like to segment. Transfer learning realizes this motivation by transferring segmentation knowledge learned for certain categories into that for the other categories. Then, from our point of view, the key determinants of success in this line of research are network architecture and learning strategy that enable DCNNs to learn segmentation knowledge that can be applied to arbitrary semantic categories out of the training data set. A pioneer study has been done by Hong et al. [17], but there is still much room for improvement in terms of both network architecture and learning strategy.

The aforementioned suggestions are mainly for weakly supervised learning of DCNNs for semantic segmentation, but we believe that some of the ideas and techniques can be applied to weakly supervised learning of other complicated visual recognition models for which annotated training examples are not sufficiently given.

Acknowledgments

This work was supported in part by the Institute for Information and Communications Technology Promotion grant (2016-0-00563, Research on Adaptive Machine Learning Technology Development for Intelligent Autonomous Digital Companion), National Research Foundation grant (NRF-2011-0031648, Global Frontier R&D Program on Human-Centered Interaction for Coexistence), and DGIST R&D Program (17-ST-02), which are funded by the Korean government. It is also partly supported by the Ministry of Culture, Sports, and Tourism of Korea, and Korea Creative Content Agency.

Authors

Seunghoon Hong (maga33@postech.ac.kr) received the B.S. and Ph.D. degrees from the Department of Computer Science and Engineering at POSTECH, Pohang, South Korea, in 2011 and 2017, respectively. He is currently a postdoctoral fellow in the Department of Electrical Engineering and Computer Science at the University of Michigan. His current research interests include computer vision and machine learning. He received the Microsoft Research Asia Fellowship in 2014.

It will be an interesting approach to weakly supervised semantic segmentation to newly introduce other unsupervised techniques that can recoup the gap between the level of supervision and that of prediction.

Suha Kwak (skwak@dgist.ac.kr) received the B.S. and Ph.D. degrees from the Department of Computer Science and Engineering at POSTECH, South Korea, in 2007 and 2014, respectively. He is an assistant professor in the Department of Information and Communication Engineering at Daegu Gyeongbuk Institute of Science and Technology (DGIST), South Korea. Before joining DGIST, he was a postdoctoral fellow of the WILLOW team at Inria Paris and École Normale Supérieure,

France. His research interests include computer vision and machine learning.

Bohyung Han (bhhan@postech.ac.kr) received the B.S. and M.S. degrees in computer engineering in 1997 and 2000, respectively, from Seoul National University, South Korea, and the Ph.D. degree in computer science in 2005 from the University of Maryland at College Park. He is currently an associate professor in the Department of Computer Science and Engineering at POSTECH, South Korea. He is an associate editor of *Computer Vision and Image Understanding* and *Machine Vision and Applications*. He served or will serve as an area chair of the International Conference on Computer Vision (ICCV) 2015/2017, IEEE Conference on Computer Vision and Pattern Recognition 2017, and Annual Conference on Neural Information Processing Systems 2015, and as a tutorial chair in ICCV 2019. His current research interests include computer vision and machine learning with an emphasis on deep learning.

References

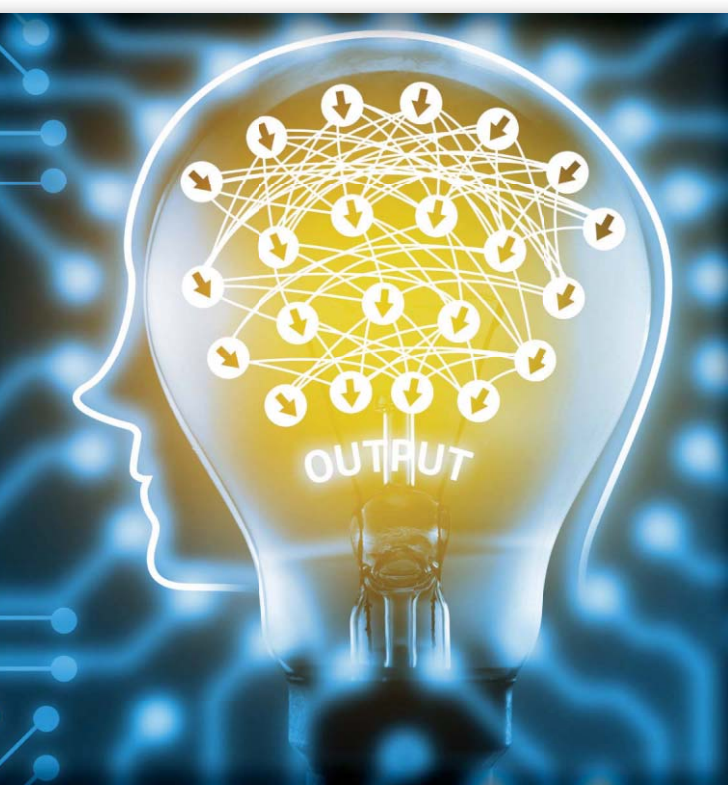
- [1] B. Alexe, T. Deselaers, and V. Ferrari, "Measuring the objectness of image windows," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 11, pp. 2189–2202 2012.
- [2] P. Arbeláez, J. Pont-Tuset, J. Barron, F. Marques, and J. Malik, "Multiscale combinatorial grouping," in *Proc. IEEE Conf. Vision and Pattern Recognition*, 2014, pp. 328–335.
- [3] D. Barnes, W. Maddern, and I. Posner, "Find your own way: Weakly-supervised segmentation of path proposals for urban autonomy," in *Proc. IEEE Int. Conf. Robotics and Automation*, 2017, pp. 203–210.
- [4] A. Bearman, O. Russakovsky, V. Ferrari, and L. Fei-Fei, "What's the point: Semantic segmentation with point supervision," in *Proc. European Conf. Computer Vision*, 2016, pp. 549–565.
- [5] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Semantic image segmentation with deep convolutional nets and fully connected CRFs," in *Proc. Int. Conf. Learning Representations*, 2015.
- [6] Y. Chen, L.-C. Yang, J. Wang, W. Xu, and A. L. Yuille, "Attention to scale: Scale-aware semantic image segmentation," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2016, pp. 3640–3649.
- [7] X. Chu, W. Ouyang, H. Li, and X. Wang, "Structured feature learning for pose estimation," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, June 2016, pp. 4715–4723.
- [8] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The cityscapes dataset for semantic urban scene understanding," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2016, pp. 3213–3223.
- [9] J. Dai, K. He, and J. Sun, "BoxSup: Exploiting bounding boxes to supervise convolutional networks for semantic segmentation," in *Proc. IEEE Int. Conf. Computer Vision*, 2015, pp. 1635–1643.
- [10] I. J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2009, pp. 248–255.

- [11] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman, "The Pascal visual object classes (VOC) challenge," *Int. J. Comput. Vision*, vol. 88, no. 2, pp. 303–338, 2010.
- [12] P. F. Felzenszwalb and D. P. Huttenlocher, "Efficient graph-based image segmentation," *Int. J. Comput. Vision*, vol. 59, no. 2, pp. 167–181, 2004.
- [13] B. Hariharan, P. Arbeláez, L. Bourdev, S. Maji, and J. Malik, "Semantic contours from inverse detectors," in *Proc. IEEE Int. Conf. Computer Vision*, 2011, pp. 991–998.
- [14] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," in *Proc. IEEE Int. Conf. Computer Vision*, 2015, pp. 1026–1034.
- [15] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2016, pp. 770–778.
- [16] S. Hong, H. Noh, and B. Han, "Decoupled deep neural network for semi-supervised semantic segmentation," in *Proc. Neural Information Processing Systems*, 2015, pp. 1495–1503.
- [17] S. Hong, J. Oh, B. Han, and H. Lee, "Learning transferrable knowledge for semantic segmentation with deep convolutional neural network," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2016, pp. 3204 – 3212.
- [18] S. Hong, D. Yeo, S. Kwak, H. Lee, and B. Han, "Weakly supervised semantic segmentation using web-crawled videos," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2017, pp. 7322–7330.
- [19] Z. Jia, X. Huang, E. I. Chang, and Y. Xu, "Constrained deep weak supervision for histopathology image segmentation," *arXiv Preprint*, arXiv:1701.00794, 2017.
- [20] A. Khoreva, R. Benenson, J. Hosang, M. Hein, and B. Schiele, "Exploiting saliency for object segmentation from image level labels," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2017, pp. 876–885.
- [21] A. Kolesnikov and C. H. Lampert, "Improving weakly-supervised object localization by micro-annotation," in *Proc. British Machine Vision Conf.*, 2016, pp. 92.1–92.12.
- [22] A. Kolesnikov and C. H. Lampert, "Seed, expand and constrain: Three principles for weakly-supervised image segmentation," in *Proc. European Conf. Computer Vision*, 2016, pp. 695–711.
- [23] P. Krähenbühl and V. Koltun, "Efficient inference in fully connected CRFs with Gaussian edge potentials," in *Proc. Neural Information Processing Systems*, 2011, pp. 109–117.
- [24] S. Kwak, S. Hong, and B. Han, "Weakly supervised semantic segmentation using superpixel pooling network," in *Proc. AAAI Conf. Artificial Intelligence*, 2017, pp. 4111–4117.
- [25] Y. Lecun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.
- [26] D. Lin, J. Dai, J. Jia, K. He, and J. Sun, "Scribblesup: Scribble-supervised convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2016.
- [27] G. Lin, C. Shen, A. van den Hengel, and I. Reid, "Efficient piecewise training of deep structured models for semantic segmentation," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2016, pp. 3194–3203.
- [28] L. Lin, G. Wang, R. Zhang, R. Zhang, X. Liang, and W. Zuo, "Deep structured scene parsing by learning with image description," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2016, pp. 2276–2284.
- [29] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft COCO: Common objects in context," in *Proc. European Conf. Computer Vision*, 2014, pp. 740–755.
- [30] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "SSD: Single shot multibox detector," in *Proc. European Conf. Computer Vision*, 2016, pp. 21–37.
- [31] Z. Liu, X. Li, P. Luo, C. C. Loy, and X. Tang, "Semantic image segmentation via deep parsing network," in *Proc. IEEE Int. Conf. Computer Vision*, 2015, pp. 1377–1385.
- [32] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2015, pp. 3431–3440.
- [33] H. Noh, S. Hong, and B. Han, "Learning deconvolution network for semantic segmentation," in *Proc. IEEE Int. Conf. Computer Vision*, 2015, pp. 1520–1528.
- [34] S. J. Oh, R. Benenson, A. Khoreva, Z. Akata, M. Fritz, and B. Schiele, "Exploiting saliency for object segmentation from image level labels," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2017, pp. 4410–4419.
- [35] G. Papandreou, L.-C. Chen, K. Murphy, and A. L. Yuille, "Weakly-and semi-supervised learning of a DCNN for semantic image segmentation," in *Proc. IEEE Int. Conf. Computer Vision*, 2015, pp. 1742–1750.
- [36] D. Pathak, P. Krähenbühl, and T. Darrell, "Constrained convolutional neural networks for weakly supervised segmentation," in *Proc. IEEE Int. Conf. Computer Vision*, 2015, pp. 1742–1750.
- [37] D. Pathak, E. Shelhamer, J. Long, and T. Darrell, "Fully convolutional multi-class multiple instance learning," in *Proc. Int. Conf. Learning Representations*, 2015.
- [38] P. O. Pinheiro and R. Collobert, "From image-level to pixel-level labeling with convolutional networks," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2015, pp. 1713–1721.
- [39] A. Prest, C. Leistner, J. Civera, C. Schmid, and V. Ferrari, "Learning object class detectors from weakly annotated video," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2012, pp. 3282–3289.
- [40] X. Qi, Z. Liu, J. Shi, H. Zhao, and J. Jia, "Augmented feedback in semantic segmentation under image level supervision," in *Proc. European Conf. Computer Vision*, 2016, pp. 90–105.
- [41] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, June 2016., pp. 779–788.
- [42] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Proc. Medical Image Computing and Computer-Assisted Intervention*, 2015, pp. 234–241.
- [43] C. Rother, V. Kolmogorov, and A. Blake "Grabcut": Interactive foreground extraction using iterated graph cuts," in *Proc. SIGGRAPH*, 2004, pp. 309–314.
- [44] F. Saleh, M. S. A. Akbarian, M. Salzmann, L. Petersson, S. Gould, and J. M. Alvarez, "Built-in foreground/background prior for weakly-supervised semantic segmentation," in *Proc. European Conf. Computer Vision*, 2016, pp. 413–432.
- [45] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, June 2015, pp. 815–823.
- [46] W. Shimoda and K. Yanai, "Distinct class-specific saliency maps for weakly supervised semantic segmentation," in *Proc. European Conf. Computer Vision*, 2016, pp. 218–234.
- [47] J. T. Springenberg, A. Dosovitskiy, T. Brox, and M. A. Riedmiller, "Striving for simplicity: The all convolutional net," in *Proc. Int. Conf. Learning Representations*, 2015.
- [48] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf, "Deepface: Closing the gap to human-level performance in face verification," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, June 2014, pp. 1701–1708.
- [49] P. Tokmakov, K. Alahari, and C. Schmid, "Learning semantic segmentation with weakly-annotated videos," in *Proc. European Conf. Computer Vision*, 2016, pp. 388–404.
- [50] A. Vezhnevets and J. M. Buhmann, "Towards weakly supervised semantic segmentation by means of multiple instance and multitask learning," in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, 2010, pp. 3249–3256.
- [51] A. Vezhnevets, V. Ferrari, and J. M. Buhmann, "Weakly supervised semantic segmentation with a multi-image model," in *Proc. IEEE Int. Conf. Computer Vision*, pp. 643–650, 2011.
- [52] A. Vezhnevets, V. Ferrari, and J. M. Buhmann, "Weakly supervised structured output learning for semantic segmentation," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2012, pp. 845–852.
- [53] S.-E. Wei, V. Ramakrishna, T. Kanade, and Y. Sheikh, "Convolutional pose machines," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, June 2016, pp. 4724–4732.
- [54] Y. Wei, X. Liang, Y. Chen, X. Shen, M. M. Cheng, J. Feng, Y. Zhao, and S. Yan, "STC: A simple to complex framework for weakly-supervised semantic segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, 2016.
- [55] Y. Xu, J.-Y. Zhu, E. I. Chang, M. Lai, and Z. Tu, "Weakly supervised histopathology cancer image segmentation and classification," *Med. Image Anal.*, vol. 18, no. 3, pp. 591–604, 2014.
- [56] L. Zhang, M. Song, Z. Liu, X. Liu, J. Bu, and C. Chen, "Probabilistic graphlet cut: Exploiting spatial structure cue for weakly supervised image segmentation," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2013, pp. 1908–1915.
- [57] S. Zheng, S. Jayasumana, B. Romera-Paredes, V. Vineet, Z. Su, D. Du, C. Huang, and P. Torr, "Conditional random fields as recurrent neural networks," in *Proc. IEEE Int. Conf. Computer Vision*, 2015, pp. 1529–1537.
- [58] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2016, pp. 2921–2929.
- [59] B. Zhou, H. Zhao, X. Puig, S. Fidler, A. Barriuso, and A. Torralba, "Scene parsing through ade20k dataset," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2017, pp. 633–641.
- [60] L. Zitnick and P. Dollár, "Edge boxes: Locating object proposals from edges," in *Proc. European Conf. Computer Vision*, 2014, pp. 391–405.

Alhussein Fawzi, Seyed-Mohsen Moosavi-Dezfooli,
and Pascal Frossard

The Robustness of Deep Networks

A geometrical perspective



©ISTOCKPHOTO.COM/ZAPP2PHOTO

Deep neural networks have recently shown impressive classification performance on a diverse set of visual tasks. When deployed in real-world (noise-prone) environments, it is equally important that these classifiers satisfy robustness guarantees: small perturbations applied to the samples should not yield significant loss to the performance of the predictor. The goal of this article is to discuss the robustness of deep networks to a diverse set of perturbations that may affect the samples in practice, including adversarial perturbations, random noise, and geometric transformations. This article further discusses the recent works that build on the robustness analysis to provide geometric insights on the classifier's decision surface, which help in developing a better understanding of deep networks. Finally, we present recent solutions that attempt to increase the robustness of deep networks. We hope this review article will contribute to shed light on the open research challenges in the robustness of deep networks and stir interest in the analysis of their fundamental properties.

Introduction

With the dramatic increase of digital data and the development of new computing architectures, deep learning has been developing rapidly as a predominant framework for data representation that can contribute in solving very diverse tasks. Despite this success, several fundamental properties of deep neural networks are still not understood and have been the subject of intense analysis in recent years. In particular, the robustness of deep networks to various forms of perturbations has received growing attention due to its importance when applied to visual data. That path of work has been mostly initiated by the illustration of the intriguing properties of deep networks in [1], which are shown to be particularly vulnerable to very small additive perturbations in the data, even if they achieve impressive performance on complex visual benchmarks [2]. An illustration of the vulnerability of deep networks to small additive perturbations can be seen in Figure 1. A dual phenomenon was observed in [3], where unrecognizable images to the human eye are classified with high confidence by deep neural

Digital Object Identifier 10.1109/MSP.2017.2740965
Date of publication: 13 November 2017

networks. The transfer of these deep networks to critical applications that possibly consist in classifying high-stake information is seriously challenged by the low robustness of deep networks. For example, in the context of self-driving vehicles, it is fundamental to accurately recognize cars, traffic signs, and pedestrians, when these are affected by clutter, occlusions, or even adversarial attacks. In medical imaging [4], it is also important to achieve high classification rates on potentially perturbed test data. The analysis of state-of-the-art deep classifiers' robustness to perturbation at test time is therefore an important step for validating the models' reliability to unexpected (possibly adversarial) nuisances that might occur when deployed in uncontrolled environments. In addition, a better understanding of the capabilities of deep networks in coping with data perturbation actually allows us to develop important insights that can contribute to developing yet better systems.

The fundamental challenges raised by the robustness of deep networks to perturbations have led to a large number of important works in recent years. These works study empirically and theoretically the robustness of deep networks to different types of perturbations, such as adversarial perturbations, additive random noise, structured transformations, or even universal perturbations. The robustness is usually measured as the sensitivity of the discrete classification function (i.e., the function that assigns a label to each image) to such perturbations. While robustness analysis is not a new problem, we provide an overview of the recent works that propose to assess the vulnerability of deep network architectures. In addition to quantifying the robustness of deep networks to various forms of perturbations, the analysis of robustness has further contributed to developing important insights on the geometry of the complex decision boundary of such classifiers, which remain hardly understood due to the very high dimensionality of the problems that they address. In fact, the robustness properties of a classifier are strongly tied to the geometry of the decision boundaries. For example, the high instability of deep neural networks to adversarial perturbations shows that data points reside extremely close to the classifier's decision boundary. The study of robustness is, therefore, not only interesting from the practical perspective of the system's reliability but has a more fundamental component that allows "understanding" of the geometric properties of classification regions and derives insights toward the improvement of current architectures.

This overview article has multiple goals. First, it provides an accessible review of the recent works in the analysis of the robustness of deep neural network classifiers to different forms of perturbations, with a particular emphasis on image analysis and visual understanding applications. Second, it presents connections between the robustness of deep networks and the geometry of the decision boundaries of such classifiers. Third, the article discusses ways to improve the robustness in deep networks architectures and finally highlights some of the important open problems.

Robustness of classifiers

In most classification settings, the proportion of misclassified samples in the test set is the main performance metric used

to evaluate classifiers. The empirical test error provides an estimate of the classifier's risk, defined as the probability of misclassification, when considering samples from the data distribution. Formally, let us define μ to be a distribution defined over images. The risk of a classifier f is equal to

$$R(f) = \mathbb{P}_{x \sim \mu}(f(x) \neq y(x)), \quad (1)$$

where x and $y(x)$ correspond, respectively, to the image and its associated label. While the risk captures the error of f on the data distribution μ , it does not capture the robustness to small arbitrary perturbations of data points. In visual classification tasks, it is desirable to learn classifiers that achieve robustness to small perturbations of the input; i.e., the application of a small perturbation to images (e.g., additive perturbations on the pixel values or geometric transformation of the image) should not alter the estimated label of the classifier.

Before going into more detail about robustness, we first define some notations. Let \mathcal{X} denote the ambient space where images live. We denote by \mathcal{R} the set of admissible perturbations. For example, when considering geometric perturbations, \mathcal{R} is set to be the group of geometric (e.g., affine) transformations under study. Alternatively, if we are to measure the robustness to arbitrary additive perturbations, we set $\mathcal{R} = \mathcal{X}$. For $r \in \mathcal{R}$, we define $T_r: \mathcal{X} \rightarrow \mathcal{X}$ to be the perturbation operator by r ; i.e., for a data point $x \in \mathcal{X}$, $T_r(x)$ denotes the image x perturbed by r . Armed with these notations, we define the minimal perturbation changing the label of the classifier, at x , as follows:

$$r^*(x) = \operatorname{argmin}_{r \in \mathcal{R}} \|r\|_{\mathcal{R}} \text{ subject to } f(T_r(x)) \neq f(x), \quad (2)$$

where $\|\cdot\|_{\mathcal{R}}$ is a metric on \mathcal{R} . For notation simplicity, we omit the dependence of $r^*(x)$ on f , \mathcal{R} , δ , and operator T . Moreover, when the image x is clear from the context, we will use r^* to refer to $r^*(x)$. See Figure 2 for an illustration. The pointwise robustness of f at x is then measured by $\|r^*(x)\|_{\mathcal{R}}$. Note that larger values of $\|\cdot\|_{\mathcal{R}}$ indicate a higher robustness at x . While this definition of robustness considers the smallest perturbation $r^*(x)$ (with respect to the metric $\|\cdot\|_{\mathcal{R}}$) that causes the classifier f to

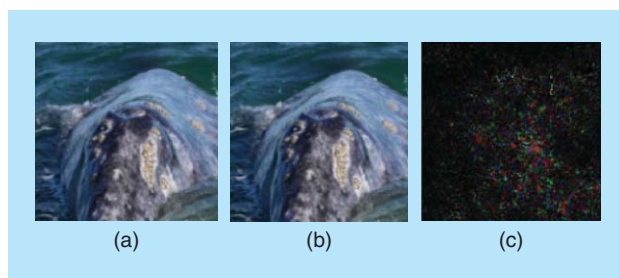


FIGURE 1. An example of an adversarial perturbations in state-of-the-art neural networks. (a) The original image that is classified as a "whale," (b) the perturbed image classified as a "turtle," and (c) the corresponding adversarial perturbation that has been added to the original image to fool a state-of-the-art image classifier [5].

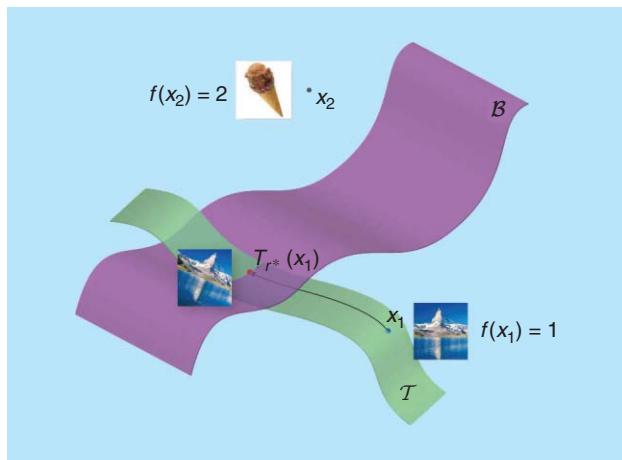


FIGURE 2. Here, \mathcal{B} denotes the decision boundary of the classifier between classes 1 and 2, and \mathcal{T} denotes the set of perturbed versions of x_1 (i.e., $\mathcal{T} = \{T_r(x_1): r \in \mathcal{R}\}$), where we recall that \mathcal{R} denotes the set of admissible perturbations. The pointwise robustness at x_1 is defined as the smallest perturbation in \mathcal{R} that causes x_1 to change class.

change the label at x , other works have instead adopted slightly different definitions, where a “sufficiently small” perturbation is sought (instead of the minimal one) [7]–[9]. To measure the global robustness of a classifier f , one can compute the expectation of $\|r^*(x)\|_{\mathcal{R}}$ over the data distribution [1], [10]. That is, the global robustness $\rho(f)$ is defined as follows:

$$\rho(f) = \mathbb{E}_{x \sim \mu} (\|r^*(x)\|_{\mathcal{R}}). \quad (3)$$

It is important to note that in our robustness setting, the perturbed point $T_r(x)$ need not belong to the support of the data distribution. Hence, while the focus of the risk in (1) is the accuracy on typical images (sampled from μ), the focus of the robustness computed from (2) is instead on the distance to the “closest” image (potentially outside the support of μ) that changes the label of the classifier. The risk and robustness hence capture two fundamentally different properties of the classifier, as illustrated in “Robustness and Risk: A Toy Example.”

Robustness and Risk: A Toy Example

To illustrate the general concepts of robustness and risk of classifiers, we consider the simple binary classification task illustrated in Figure S1, where the goal is to discriminate between images representing vertical and horizontal stripes. In addition to the orientation of the stripe that separates the two classes, a very small positive bias is added to pixels of first-class images and subtracted from the pixels of the images in the second class. This bias is chosen to be very small, in such a way that it is imperceptible to humans; see Figure S2 for example images of class 1 and 2 with the pixel values, where a denotes the bias.

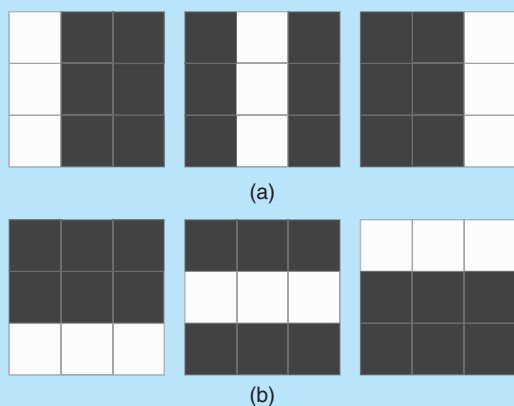


FIGURE S1. (a) The images belonging to class 1 (vertical stripe and positive bias) and (b) the images belonging to class 2 (horizontal stripe and negative bias).

It is easy to see that a linear classifier can perfectly separate the two classes, thus achieving zero risk (i.e., $R(f) = 0$). Note, however, that such a classifier only achieves zero risk because it captures the bias but fails to distinguish between the images based on the orientation of the stripe. Hence, despite being zero risk, this classifier is highly unstable to additive perturbation, as it suffices to perturb the bias of the image (i.e., by adding a very small value to all pixels) to cause misclassification. On the other hand, a more complex classifier that captures the orientation of the stripe will be robust to small perturbations (while equally achieving zero risk), as changing the label would require changing the direction of the stripe, which is the most visual (and natural) concept that separates the two classes.

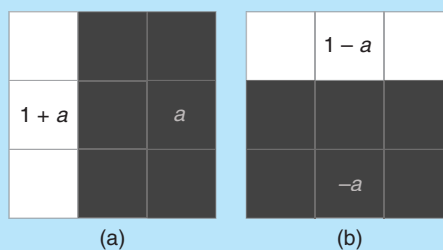


FIGURE S2. (a) An example image of class 1. White pixels have value $1 + a$, and black pixels have value a . (b) An example image of class -1 . White pixels have value $1 - a$, and black pixels have value $-a$. The bias a is set to be very small, in such a way that it is imperceptible.

Observe that classification robustness is strongly related to support vector machine (SVM) classifiers, whose goal is to maximize the robustness, defined as the margin between support vectors. Importantly, the max-margin classifier in a given family of classifiers might, however, still not achieve robustness (in the sense of high $\rho(f)$). An illustration is provided in “Robustness and Risk: A Toy Example,” where a no zero-risk linear classifier—in particular, the max-margin classifier—achieves robustness to perturbations. Our focus in this article is turned toward assessing the robustness of the family of deep neural network classifiers that are used in many visual recognition tasks.

Perturbation forms

Robustness to additive perturbations

We first start by considering the case where the perturbation operator is simply additive; i.e., $T_r(x) = x + r$. In this case, the magnitude of the perturbation can be measured with the ℓ_p norm of the minimal perturbation that is necessary to change the label of a classifier. According to (2), the robustness to additive perturbations of a data point x is defined as

$$\min_{r \in \mathcal{R}} \|r\|_p \text{ subject to } f(x+r) \neq f(x). \quad (4)$$

Depending on the conditions that one sets on the set \mathcal{R} that supports the perturbations, the additive model leads to different forms of robustness.

Adversarial perturbations

We first consider the case where the additive perturbations are unconstrained (i.e., $\mathcal{R} = \mathcal{X}$). The perturbation obtained by solving (4) is often referred to as an adversarial perturbation, as it corresponds to the perturbation that an adversary (having full knowledge of the model) would apply to change the label of the classifier, while causing minimal changes to the original image.

The optimization problem in (4) is nonconvex, as the constraint involves the (potentially highly complex) classification function f . Different techniques exist to approximate adversarial perturbations. In the following, we briefly mention some of the existing algorithms for computing adversarial perturbations:

- Regularized variant [1]: The method in [1] computes adversarial perturbations by solving a regularized variant of the problem in (4), given by

$$\min_c \|r\|_p + J(x+r, \tilde{y}, \theta), \quad (5)$$

where \tilde{y} is a target label of the perturbed sample, J is a loss function, c is a regularization parameter, and θ is the model parameters. In the original formulation [1], an additional constraint is added to guarantee $x+r \in [0, 1]$, which is omitted in (5) for simplicity. To solve the optimization problem in (5), a line search is performed over c to find the maximum $c > 0$ for which the minimizer of (5) satisfies $f(x+r) = \tilde{y}$. While leading to very accurate estimates, this approach can be costly to compute on high-dimensional and large-scale data sets. More-

over, it computes targeted adversarial perturbations, where the target label is known.

- Fast gradient sign (FGS) [11]: This solution estimates an untargeted adversarial perturbation by going in the direction of the sign of gradient of the loss function:

$$\in \text{sign}(\nabla_x J(x, y(x), \theta)),$$

where J , the loss function, is used to train the neural network and θ denotes the model parameters. While efficient, this one-step algorithm provides a coarse approximation to the solution of the optimization problem in (4) for $p = \infty$.

- DeepFool [5]: This algorithm minimizes (4) through an iterative procedure, where each iteration involves the linearization of the constraint. The linearized (constrained) problem is solved in closed form at each iteration, and the current estimate is updated; the optimization procedure terminates when the current estimate of the perturbation fools the classifier. In practice, DeepFool provides a tradeoff between the accuracy and efficiency of the two previous approaches [5].

In addition to the aforementioned optimization methods, several other approaches have recently been proposed to compute adversarial perturbations, see, e.g., [9], [12], and [13]. Different from the previously mentioned gradient-based techniques, the recent work in [14] learns a network (the adversarial transformation network) to efficiently generate a set of perturbations with a large diversity, without requiring the computation of the gradients.

Using the aforementioned optimization techniques, one can compute the robustness of classifiers to additive adversarial perturbations. Quite surprisingly, deep networks are extremely vulnerable to such additive perturbations; i.e., small and even imperceptible adversarial perturbations can be computed to fool them with high probability. For example, the average perturbations required to fool the CaffeNet [15] and GoogleNet [16] architectures on the ILSVRC 2012 task [17] are 100 times smaller than the typical norm of natural images [5] when using the ℓ_2 norm. The high instability of deep neural networks to adversarial perturbations, which was first highlighted in [1], shows that these networks rely heavily on proxy concepts to classify objects, as opposed to strong visual concepts typically used by humans to distinguish between objects.

To illustrate this idea, we consider once again the toy classification example (see “Robustness and Risk: A Toy Example”), where the goal is to classify images based on the orientation of the stripe. In this example, linear classifiers could achieve a perfect recognition rate by exploiting the imperceptibly small bias that separates the two classes. While this proxy concept achieves zero risk, it is not robust to perturbations: one could design an additive perturbation that is as simple as a minor variation of the bias, which is sufficient to induce data misclassification. On the same line of thought, the high instability of classifiers to additive perturbations observed in [1] suggests that deep neural networks potentially capture one of the proxy concepts that separate the different classes. Through a quantitative analysis of polynomial

classifiers, [10] suggests that higher-degree classifiers tend to be more robust to perturbations, as they capture the “stronger” (and more visual) concept that separates the classes (e.g., the orientation of the stripe in Figure S1 in “Robustness and Risk: A Toy Example”). For neural networks, however, the relation between the flexibility of the architecture (e.g., depth and breadth) and adversarial robustness is not well understood and remains an open problem.

Random noise

In the random noise regime, data points are perturbed by noise having a random direction in the input space. Unlike the adversarial case, the computation of random noise does not require knowledge of the classifier; it is therefore crucial for state-of-the-art classifiers to be robust to this noise regime. We measure the pointwise robustness to random noise by setting \mathcal{R} to be a direction sampled uniformly at random from the ℓ_2 unit sphere \mathbb{S}^{d-1} in \mathcal{X} (where d denotes the dimension of \mathcal{X}). Therefore, (4) becomes

$$r_v^*(x) = \operatorname{argmin}_{r \in \{\alpha v : \alpha \in \mathbb{R}\}} \|r\|_2 \text{ subject to } f(x+r) \neq f(x), \quad (6)$$

where v is a direction sampled uniformly at random from the unit sphere \mathbb{S}^{d-1} . The pointwise robustness is then defined as the ℓ_2 norm of the perturbation, i.e., $\|r_v^*(x)\|_2$.

The robustness of classifiers to random noise has previously been studied empirically in [1] and theoretically in [10] and [18]. Empirical investigation suggests that state-of-the-art classifiers are much more robust to random noise than to adversarial perturbations, i.e., the norm of the noise $r_v^*(x)$ required to change the label of the classifier can be several orders of magnitudes larger than that of the adversarial perturbation. This result is confirmed theoretically, as linear classifiers in [10] and nonlinear classifiers in [18] are shown to have a robustness to random noise that behaves as

$$\|r_v^*(x)\|_2 = \Theta(\sqrt{d} \|r_{\text{adv}}^*(x)\|_2)$$

with high probability, where $\|r_{\text{adv}}^*(x)\|_2$ denotes the robustness to adversarial perturbations [(4) with $\mathcal{R} = \mathcal{X}$]. In other words, this result shows that, in high-dimensional classification settings (i.e., large d), classifiers can be robust to random noise, even if the pointwise adversarial robustness of the classifier is very small.

Semirandom noise

Finally, the semirandom noise regime generalizes this additive noise model to random subspaces \mathcal{S} of dimension $m \leq d$. Specifically, in this perturbation regime, an adversarial perturbation is sought within a random subspace \mathcal{S} of dimension m . That is, the semirandom noise is defined as follows:

$$r_{\mathcal{S}}^*(x) = \operatorname{argmin}_{r \in \mathcal{S}} \|r\|_2 \text{ subject to } f(x+r) \neq f(x). \quad (7)$$

With the dramatic increase of digital data and the development of new computing architectures, deep learning has been developing rapidly as a predominant framework for data representation in solving very diverse tasks.

Note that, when $m = 1$, this semirandom noise regime precisely coincides with the random noise regime, whereas $m = d$ corresponds to the adversarial perturbation regime defined previously. For this generalized noise regime, a precise relation between the robustness to semirandom and adversarial perturbation exists [18], as it is shown that

$$\|r_{\mathcal{S}}^*(x)\|_2 = \Theta\left(\sqrt{\frac{d}{m}} \|r_{\text{adv}}^*(x)\|_2\right).$$

This result shows in particular that, even when the dimension m is chosen as a small fraction of d , it is still possible to find

small perturbations that cause data misclassification. In other words, classifiers are not robust to semirandom noise that is only mildly adversarial and overwhelmingly random [18]. This implies that deep networks can be fooled by very diverse small perturbations, as these can be found along random subspaces of dimension $m \ll d$.

Robustness to structured transformations

In visual tasks, it is not only crucial to have classifiers that are robust against additive perturbations as described previously. It is also equally important to achieve invariance to structured nuisance variables such as illumination changes, occlusions, or standard local geometric transformations of the image. Specifically, when images undergo such structured deformations, it is desirable that the estimated label remains the same.

One of the main strengths of deep neural network classifiers with respect to traditional shallow classifiers is that the former achieve higher levels of invariance [19] to transformations. To verify this claim, several empirical works have been introduced. In [6], a formal method is proposed that leverages the generalized robustness definition of (2) to measure the robustness of classifiers to arbitrary transformation groups. The robustness to structured transformations is precisely measured by setting the admissible perturbation space \mathcal{R} to be the set of transformations (e.g., translations, rotations, dilation) and the perturbation operator T of (2) to be the warping operator transforming the coordinates of the image. In addition, $\|\cdot\|_{\mathcal{R}}$ is set to measure the change in appearance between the original and transformed images. Specifically, $\|\cdot\|_{\mathcal{R}}$ is defined to be the length of the shortest path on the nonlinear manifold of transformed images $\mathcal{T} = \{T_r(x) : r \in \mathcal{R}\}$. Using this approach, it is possible to quantify the amount of change that the image should undergo to cause the classifier to make the wrong decision. Despite improving the invariance over shallow networks, the method in [6] shows that deep classifiers are still not robust to sufficiently small deformations on simple visual classification tasks. In [20], the authors assess the robustness of face recognition deep networks to physically realizable structured perturbations. In particular, wearing eyeglass frames is shown to cause state-of-the-art face-recognition algorithms to misclassify. In [7], the robustness to other

forms of complex perturbations is tested, and state-of-the-art deep networks are shown once again to be unstable to these perturbations. An empirical analysis of the ability of current convolutional neural networks (CNNs) to manage location and scale variability is proposed in [21]. It is shown, in particular, that CNNs are not very effective in factoring out location and scale variability, despite the popular belief that the convolutional architecture and the local spatial pooling provides invariance to such representations. The aforementioned works show that, just as state-of-the-art deep neural networks have been observed to be unstable to additive unstructured perturbations, such modern classifiers are not robust to perturbations even when severely restricting the set of possible transformations of the image.

Universal additive perturbations

All of the previous definitions capture different forms of robustness, but they all rely on the computation of data-specific perturbations. Specifically, they consider the necessary change that should be applied to specific samples to change the decision of the classifier. More generally, one might be interested to understand if classifiers are also vulnerable to generic (data and network agnostic) perturbations. The analysis of the robustness to such perturbations is interesting from several perspectives: 1) these perturbations might not require the precise knowledge of the classifier under test, 2) they might cap-

ture important security and reliability properties of classifiers, and 3) they show important properties on the geometry of the decision boundary of the classifier.

In [22], deep networks are shown to be surprisingly vulnerable to universal (image-agnostic) perturbations. Specifically, a universal perturbation v can be defined as the minimal perturbation that fools a large fraction of the data points sampled from the data distribution μ , i.e.,

$$v = \underset{r}{\operatorname{argmin}} \|r\|_p \text{ subject to } \mathbb{P}_{x \sim \mu}(f(x+r) \neq f(x)) \geq 1 - \epsilon, \quad (8)$$

where ϵ controls the fooling rate of the universal perturbation. Unlike adversarial perturbations that target to fool a specific data point, universal perturbations attempt to fool most images sampled from the natural images distribution μ . Specifically, by adding this single (image-agnostic) perturbation to a natural image, the label estimated by the deep neural network will be changed with high probability. In [22], an algorithm is provided to compute such universal perturbations; these perturbations are further shown to be quasi-imperceptible while fooling state-of-the-art deep networks on unseen natural images with probability edging 80%. Specifically, the ℓ_p norm of these perturbations is at least one order of magnitude smaller than the norm of natural images but causes most perturbed images to be misclassified. Figure 3 illustrates examples of scaled universal

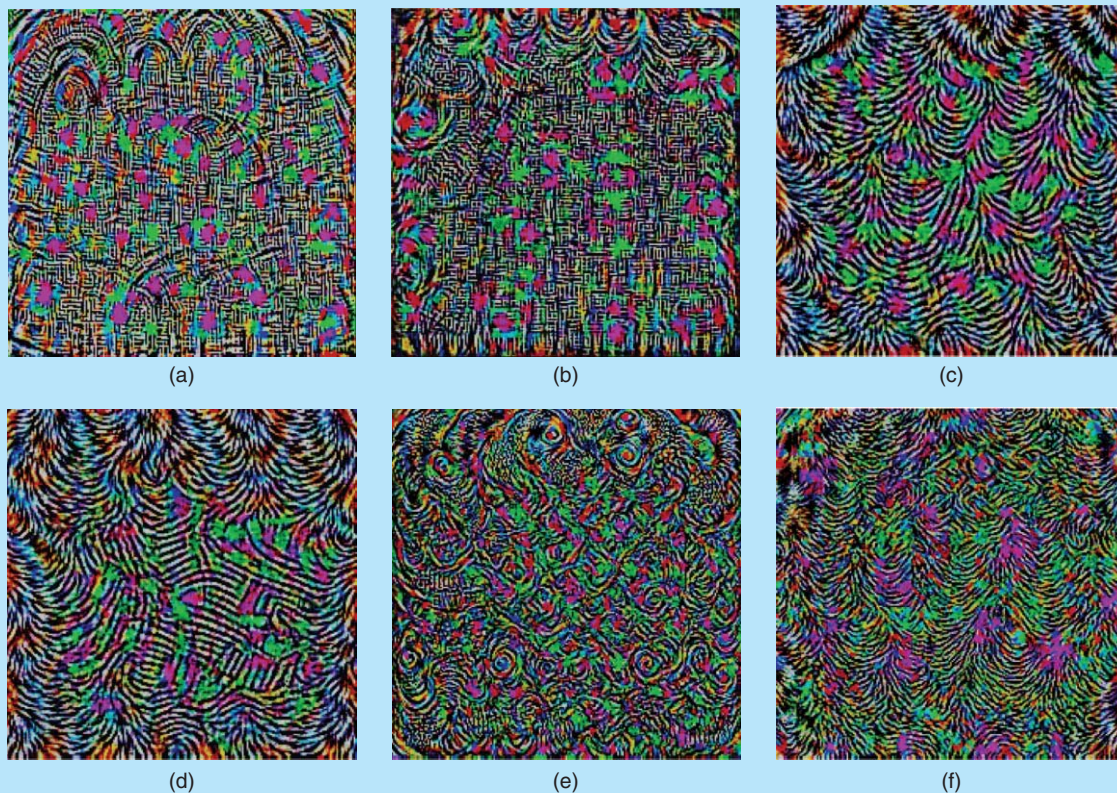


FIGURE 3. Universal perturbations computed for different deep neural network architectures. The pixel values are scaled for visibility. (a) CaffeNet, (b) VGG-F, (c) VGG-16, (d) VGG-19, (e) GoogLeNet, and (f) ResNet-152.

perturbations computed for different deep neural networks, and Figure 4 illustrates examples of perturbed images. When added to the original images, a universal perturbation is quasi-imperceptible but causes most images to be misclassified. Note that adversarial perturbations computed using the algorithms described in the section “Adversarial Perturbations” are not universal across data points, as shown in [22]. That is, adversarial perturbations only generalize mildly to unseen data points, for a fixed norm comparable to that of universal perturbations.

Universal perturbations are further shown in [22] to transfer well across different architectures; a perturbation computed for a given network is also very likely to fool another network on most natural images. In that sense, such perturbations are doubly universal, as they generalize well across images and architectures. Note that this property is shared with adversarial perturbations, as the latter perturbations have been shown to transfer well across different models (with potentially different architectures) [1], [23]. The existence of general-purpose perturbations can be very problematic from a safety perspective, as an attacker might need very

little information about the actual model to craft successful perturbations [24].

Figure 5 illustrates a summary of the different types of perturbations considered in this section on a sample image. As can be seen, the classifier is not robust to slight perturbations of the image (for most additive perturbations) and natural geometric transformations of the image.

Geometric insights from robustness

The study of robustness allows us to derive insights about the classifiers and, more precisely, about the geometry of the classification function acting on the high-dimensional input space. We recall that $f: \mathcal{X} \rightarrow \{1, \dots, C\}$ denotes our C -class classifier, and we denote by g_1, \dots, g_C the C probabilities associated to each class by the classifier. Specifically, for a given $x \in \mathcal{X}$, $f(x)$ is assigned to the class having a maximal score; i.e., $f(x) = \operatorname{argmax}_i \{g_i(x)\}$. For deep neural networks, the functions g_i represent the outputs of the last layer in the network (generally the softmax layer). Note that the classifier f can be seen as a mapping that partitions the input space \mathcal{X} into classification regions, each of which has a constant



FIGURE 4. Examples of natural images perturbed with the universal perturbation and their corresponding estimated labels with GoogLeNet. (a)–(h) Images belonging to the ILSVRC 2012 validation set. (i)–(l) Personal images captured by a mobile phone camera. (Figure used courtesy of [22].)



FIGURE 5. (a) The original image. The remaining images are minimally perturbed images (along with the corresponding estimated label) that misclassify the CaffeNet deep neural network. (b) Adversarial perturbation, (c) random noise, (d) semirandom noise with $m = 1,000$, (e) universal perturbation, (f) affine transformation. (Figure used courtesy of [17].)

estimated label (i.e., $f(x)$ is constant for each such region). The decision boundary \mathcal{B} of the classifier is defined as the union of the boundaries of such classification regions (see Figure 2).

Adversarial perturbations

We first focus on additive adversarial perturbations and highlight their relation with the geometry of the decision boundary. This link relies on the simple observation shown in “Geometric Properties of Adversarial Perturbations.” The two geometric properties are illustrated in Figure 6. Note that these geometric properties are specific to the ℓ_2 norm. The high instability of classifiers to adversarial perturbations, which we highlighted in the previous section, shows that natural images lie very closely to the classifier’s decision boundary. While this result is key to understanding the geometry of the data points with regard to the classifier’s decision boundary, it does not provide any insights on the shape of the decision boundary. A local geometric description of the decision boundary (in the vicinity of x) is rather captured by the direction of $r_{\text{adv}}^*(x)$, due to the orthogonality property of adversarial perturbations (highlighted in “Geometric Properties of Adversarial Perturbations”). In [18] and [25], these geometric properties of adversarial perturbations are leveraged to visualize typical cross sections of the decision boundary at the vicinity of the data points. Specifically, a two-dimensional normal section of the decision boundary is illustrated, where the sectioning plane is spanned by the adversarial perturbation (normal to the decision boundary) and a random vector in the tangent space. Examples of normal sections of decision boundaries are illustrated in Figure 7.

Observe that the decision boundaries of state-of-the-art deep neural networks have a very low curvature on these two-dimensional cross sections (note the difference between the x and y axis). In other words, these plots suggest that the decision boundary at the vicinity of x can be locally well

Geometric Properties of Adversarial Perturbations

Observation

Let $x \in \mathcal{X}$ and $r_{\text{adv}}^*(x)$ be the adversarial perturbation, defined as the minimizer of (4), with $p = 2$ and $\mathcal{R} = \mathcal{X}$. Then, we have the following:

- 1) $\|r_{\text{adv}}^*(x)\|_2$ measures the Euclidean distance from x to the closest point on the decision boundary \mathcal{B} .
- 2) The vector $r_{\text{adv}}^*(x)$ is orthogonal to the decision boundary of the classifier, at $x + r_{\text{adv}}^*(x)$.

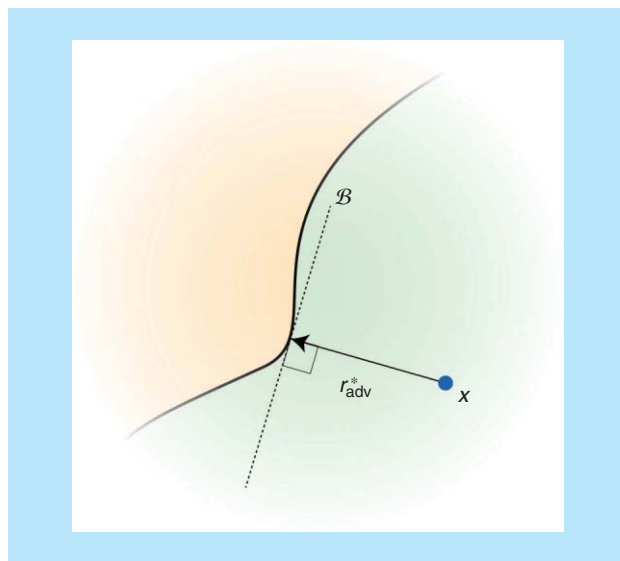


FIGURE 6. r_{adv}^* denotes the adversarial perturbation of x (with $p = 2$). Note that r_{adv}^* is orthogonal to the decision boundary \mathcal{B} and $\|r_{\text{adv}}^*\|_2 = \text{dist}(x, \mathcal{B})$.

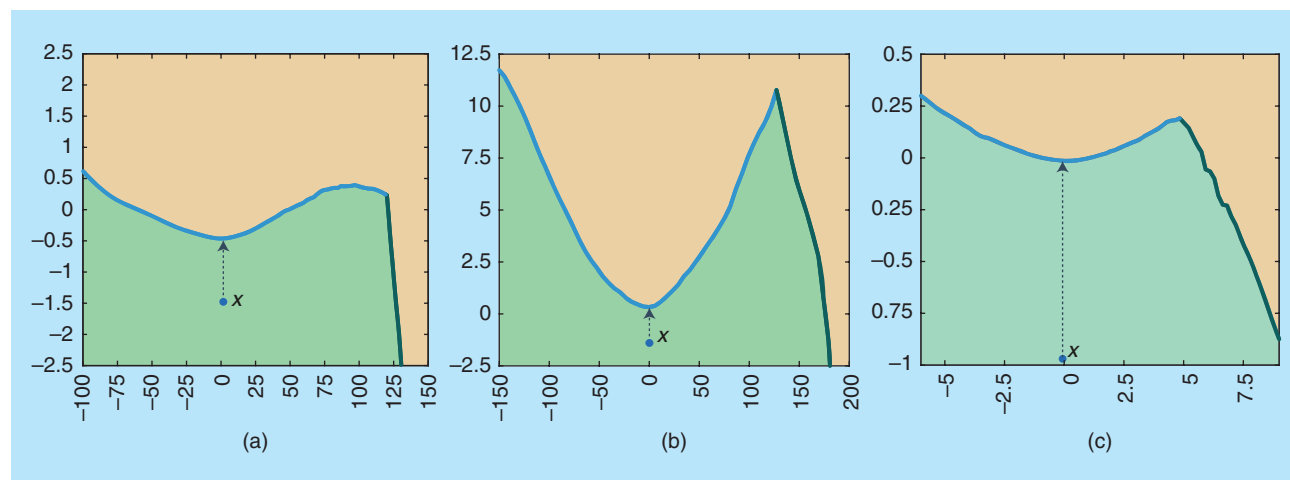


FIGURE 7. The two-dimensional normal cross sections of the decision boundaries for three different classifiers near randomly chosen samples. The section is spanned by the adversarial perturbation of the data point x (vertical axis) and a random vector in the tangent space to the decision boundary (horizontal axis). The green region is the classification region of x . The decision boundaries with different classes are illustrated in different colors. Note the difference in range between the x and y axes. (a) VGG-F (ImageNet), (b) LeNet (CIFAR), (c) LeNet (MNIST). (Figure used with permission from [18].)

approximated by a hyperplane passing through $x + r_{adv}^*(x)$ with the normal vector $r_{adv}^*(x)$. In [11], it is hypothesized that state-of-the-art classifiers are “too linear,” leading to decision boundaries with very small curvature and further explaining the high instability of such classifiers to adversarial perturbations. To motivate the linearity hypothesis of deep networks, the success of the FGS method (which is exact for linear classifiers) in finding adversarial perturbations is invoked. However, some recent works challenge this linearity hypothesis; for example, in [26], the authors show that there exist adversarial perturbations that cannot be explained with this hypothesis, and, in [27], the authors provide a new explanation based on the tilting of the decision boundary with respect to the data manifold. We stress here that the low curvature of the decision boundary does not, in general, imply that the function learned by the deep neural network (as a function of the input image) is linear, or even approximately linear. Figure 8 shows illustrative examples of highly nonlinear functions resulting in flat decision boundaries. Moreover, it should be noted that, while the decision boundary of deep networks is very flat on random two-dimensional cross sections, these boundaries are not flat

on all cross sections. That is, there exist directions in which the boundary is very curved. Figure 9 provides some illustrations of such cross sections, where the decision boundary has large curvature and therefore significantly departs from the first-order linear approximation, suggested by the flatness of the decision boundary on random sections in Figure 7. Hence, these visualizations of the decision boundary strongly suggest that the curvature along a small set of directions can be very large and that the curvature is relatively small along random directions in the input space. Using a numerical computation of the curvature, the sparsity of the curvature profile is empirically verified in [28] for deep neural networks, and the directions where the decision boundary is curved are further shown to play a major role in explaining the robustness properties of classifiers. In [29], the authors provide a complementary analysis on the curvature of the decision boundaries induced by deep networks and show that the first principal curvatures increase exponentially with the depth of a random neural network. The analyses of [28] and [29] hence suggest that the curvature profile of deep networks is highly sparse (i.e., the decision boundaries are almost flat along most directions) but can have a very large curvature along a few directions.

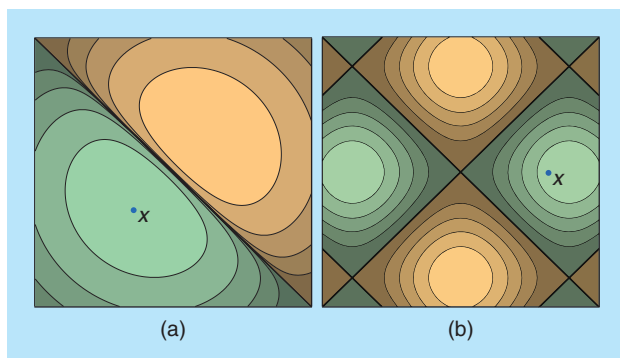


FIGURE 8. The contours of two highly nonlinear functions (a) and (b) with flat boundaries. Specifically, the contours in the green and yellow regions represent the different (positive and negative) level sets of $g(x)$ [where $g(x) = g_1(x) - g_2(x)$, the difference between class 1 and class 2 score]. The decision boundary is defined as the region of the space where $g(x) = 0$ and is indicated with a solid black line. Note that, although g is a highly nonlinear function in these examples, the decision boundaries are flat.

Universal perturbations

The vulnerability of deep neural networks to universal (image-agnostic) perturbations studied in [22] sheds light on another aspect of the decision boundary: the correlations between different regions of the decision boundary, in the vicinity of different natural images. In fact, if the orientations of the decision boundary in the neighborhood of different data points were uncorrelated, the best universal perturbation would correspond to a random perturbation. This is refuted in [22], as the norm of the random perturbation required to fool 90% of the images is ten times larger than the norm of universal perturbations. Such correlations in the decision boundary are quantified in [22], as it is shown empirically that normal vectors to the decision boundary at the vicinity of different data points (or, equivalently, adversarial perturbations due to the orthogonality property in “Geometric Properties of Adversarial Perturbations”) approximately span a low-dimensional

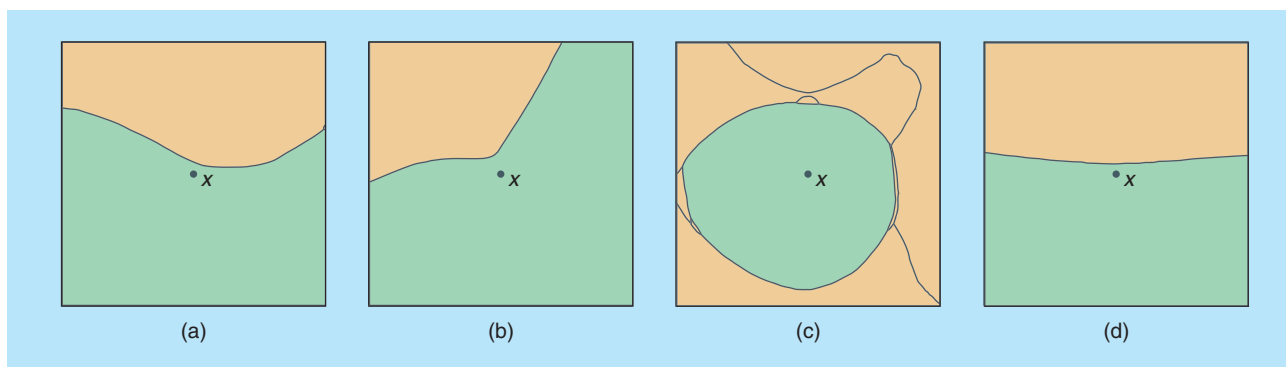


FIGURE 9. Cross sections of the decision boundary in the vicinity of data point x . (a), (b), and (c) show decision boundaries with high curvature, while (d) shows the decision boundary along a random normal section (with very small curvature). The correct class and the neighboring classes are colored in green and orange, respectively. The boundaries between different classes are shown in solid black lines. The x and y axes have the same scale.

subspace. It is conjectured that the existence of universal perturbations fooling classifiers for most natural images is partly due to the existence of such a low-dimensional subspace that captures the correlations among different regions of the decision boundary. In fact, this subspace “collects” normals to the decision boundary in different regions, and perturbations belonging to this subspace are therefore likely to fool other data points. This observation implies that the decision boundaries created by deep neural networks are not sufficiently “diverse,” despite the very large number of parameters in modern deep neural networks.

A more thorough analysis is provided in [30], where universal perturbations are shown to be tightly related to the curvature of the decision boundary in the vicinity of data points. Specifically, the existence of universal perturbations is attributed to the existence of common directions where the decision boundary is positively curved in the vicinity of most natural images. Figure 10 intuitively illustrates the link between positive curvature and vulnerability to perturbations; the required perturbation to change the label (along a fixed direction v) of the classifier is smaller if the decision boundary is positively curved, than if the decision boundary is flat (or negatively curved).

With this geometric perspective, universal perturbations correspond exactly to directions where the decision boundary is positively curved in the vicinity of most natural images. As shown in [30], this geometric explanation of universal perturbations suggests a new algorithm to compute such perturbations as well as to explain several properties, such as the diversity and transferability of universal perturbations.

Classification regions

The robustness of classifiers is not only related to the geometry of the decision boundary, but it is also strongly tied to the classification regions in the input space \mathcal{X} . The classification region associated to class $c \in \{1, \dots, C\}$ corresponds to the set of points $x \in \mathcal{X}$ such that $f(x) = c$. The study of universal perturbations in [22] has shown the existence of dominant labels, with universal perturbations mostly fooling natural images into such labels. The existence of such domi-

nant classes is attributed to the large volumes of classification regions corresponding to dominant labels in the input space \mathcal{X} : in fact, images sampled uniformly at random from the Euclidean sphere $\alpha\mathbb{S}^{d-1}$ of the input space \mathcal{X} (where the radius α is set to reflect the typical norm of natural images) are classified as one of these dominant labels. Hence, such dominant labels represent high-volume “oceans” in the image space; universal perturbations therefore tend to fool images into such target labels, as these generally result in smaller fooling perturbations. It should be noted that these dominant labels are classifier specific and are not a result of the visual properties of the images in the class.

To further understand the geometrical properties of classification regions, we note that, just like natural images, random images are strongly vulnerable to adversarial perturbations. That is, the norm of the smallest adversarial perturbation needed to change the label of a random image (sampled from \mathcal{X}) is several orders of magnitude smaller than the norm of the image itself. This observation suggests that classification regions are “hollow” and that most of their mass occurs at the boundaries. In [28], further topological properties of classification regions are observed; in particular, these regions are shown empirically to be connected.

In other words, each classification region in the input space \mathcal{X} is made up of a single connected (possibly complex) region, rather than several disconnected regions.

We have discussed in this section that the properties and optimization methods derived to analyze the robustness properties of classifiers allow us to derive insights on the geometry of the classifier. In particular, through visualizations, we have seen that the decision boundaries on normal random sections have very low curvature, while being very curved along a few directions of the input space. Moreover, the high vulnerability of state-of-the-art deep networks to universal perturbations suggests that the decision boundaries of such networks do not have sufficient diversity. To improve the robustness to such perturbations, it is therefore key to “diversify” the decision boundaries of the network and leverage the large number of parameters that define the neural network.

The study of robustness allows us to derive insights about the classifiers and, more precisely, about the geometry of the classification function acting on the high-dimensional input space.

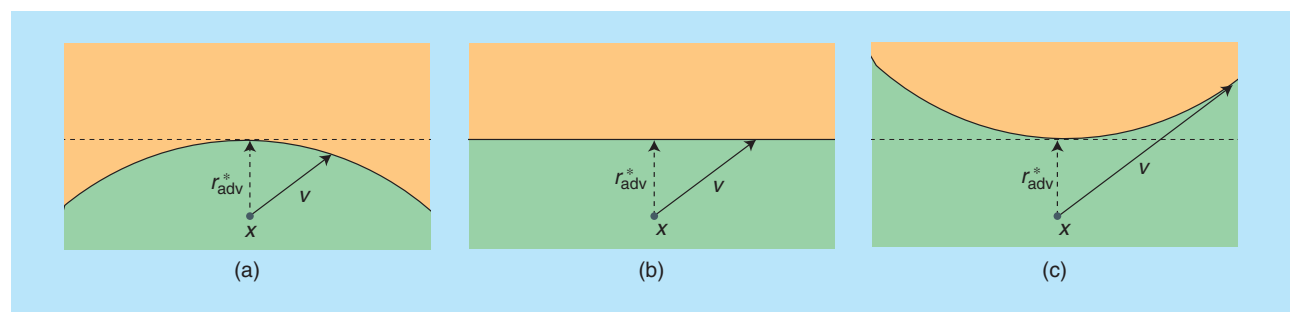


FIGURE 10. The link between robustness and curvature of the decision boundary. When the decision boundary is (a) positively curved, small universal perturbations are more likely to fool the classifier. (b) and (c) illustrate the case of a flat and negatively curved decision boundary, respectively.

Improving robustness

An important objective of the analysis of robustness is to contribute to the design of better and more reliable systems. We next summarize some of the recent attempts that have been made to render systems more robust to different forms of perturbations.

Improving the robustness to adversarial perturbations

We first describe the methods that have been proposed to construct deep networks with better robustness to adversarial perturbations, following the papers [1], [9] that originally highlighted the vulnerability of these classifiers. The straightforward approach, which consists of adding perturbed images to the training set and fine-tuning the network, has been shown to be mildly effective against newly computed adversarial perturbations [5]. To further improve the robustness, it is natural to consider the Jacobian matrix $\partial g/\partial x$ of the model (with g the last layer of the neural network) and ensure that all of the elements in the matrix are sufficiently small. Following this idea, the authors of [31] consider a modified objective function, where a term is added to penalize the Jacobians of the function computed by each layer with respect to the previous layer. This has the effect of learning smooth functions with respect to the input and thus learn more robust classifiers. In [32], a robust optimization formulation is considered for training deep neural networks. Specifically, a minimization-maximization approach is proposed, where the loss is minimized over worst-case examples, rather than only on the original data. That is, the following minimization-maximization training procedure is used to train the network:

$$\min_{\theta} \sum_{i=1}^N \max_{r \in \mathcal{U}} J(x_i + r_i, y_i, \theta), \quad (9)$$

where θ , N , and \mathcal{U} denote, respectively, the parameters of the network, the number of training points, and the set of plausible perturbations; and y_i denotes the label of x_i . The set \mathcal{U} is generally set to be the ℓ_2 or ℓ_∞ ball centered at zero and of sufficiently small radius. Unfortunately, this optimization problem in (9) is difficult to solve efficiently. To circumvent this difficulty, [32] proposes an alternating iterative method where a single step of gradient ascent and descent is performed at each iteration. Note that the construction of robust classifiers using min-max robust optimization methods has been an active area of research, especially in the context of SVM classifiers [33]. In particular, for certain sets \mathcal{U} , the objective function of various learning tasks can be written as a convex optimization function as shown in [34]–[37], which makes the task of finding a robust classifier feasible. In a very recent work inspired by biophysical principles of neural circuits, Nayeibi and Ganguli consider a regularizer to push activations of the network in the saturating regime of the nonlinearity

(i.e., the region where the nonlinear activation function is flat) [47]. The networks learned using this approach are shown to significantly improve in terms of robustness on a simple digit recognition classification task, without losing significantly in terms of accuracy. In [38], the authors propose to improve the robustness by using distillation, a technique first introduced in [39] for transferring knowledge from larger architectures to smaller ones. However, [40] shows that, when using more elaborate algorithms to compute perturbations, this approach fails to improve the robustness. In [41], a regularization scheme is introduced for improving the network's sensitivity to perturbations by constraining the Lipschitz constant of the network. In [42], an information-theoretic loss function is used to train

stochastic neural networks; the resulting classifiers are shown to be more robust to adversarial perturbations than their deterministic counterpart. The increased robustness is intuitively due to the randomness of the neural network, which maps an input to a distribution of features; attacking the network with a small designed perturbation therefore becomes harder than for deterministic neural networks.

While all of these methods are shown to yield some improvements on

the robustness of deep neural networks, the design of robust visual classifiers on challenging classification tasks (e.g., ImageNet) is still an open problem. Moreover, while the previously mentioned methods provide empirical results showing the improvement in robustness with respect to one or a subset of adversarial generation techniques, it is necessary in many applications to design robust networks against all adversarial attacks. To do so, we believe it is crucial to derive formal certificates on the robustness of newly proposed networks, as it is practically impossible to test against all possible attacks, and we see this as an important future work in this area.

Although there is currently no method to effectively (and provably) combat adversarial perturbations on large-scale data sets, several studies [42]–[44] have recently considered the related problem of detectability of adversarial perturbations. The detectability property is essential in real-world applications, as it allows the possibility to raise an exception when tampered images are detected. In [42], the authors propose to augment the network with a detector network, which detects original images from perturbed ones. Using the optimization methods in the section “Adversarial Perturbations,” the authors conclude that the network successfully learns to distinguish between perturbed samples and original samples. Moreover, the overall network (i.e., the network and detector) is shown to be more robust to adversarial perturbations tailored for this architecture. In [43], the Bayesian uncertainty estimates in the subspace of learned representations are used to discriminate perturbed images from clean samples. Finally, as shown in [44], side

The importance of analyzing the vulnerability of deep neural networks to perturbations therefore goes beyond the practical security implications, as it further reveals crucial geometric properties of deep networks.

information such as depth maps can be exploited to detect adversarial samples.

Improving the robustness to geometric perturbations

Just as in the case of adversarial perturbations, one popular way of building more invariant representations to geometric perturbations is through virtual jittering (or data augmentation), where training data are transformed and fed back to the training set. One of the drawbacks of this approach is, however, that the training can become intractable, as the size of the training set becomes substantially larger than the original data set. In another effort to improve the invariance properties of deep CNNs, the authors in [45] proposed a new module, the spatial transformer, that geometrically transforms the filter maps. Similarly to other modules in the network, spatial transformer modules are trained in a purely supervised fashion. Using spatial transformer networks, the performance of classifiers improves significantly, especially when images have noise and clutter, as these modules automatically learn to localize and unwarped corrupted images. To build robust deep representations, [46] considers instead a new architecture with fixed filter weights. Specifically, a similar structure to CNNs (i.e., cascade of filtering, nonlinearity, and pooling operations) is considered with the additional requirement of stability of the representation to local deformations, while retaining maximum information about the original data. The scattering network is proposed, where successive filtering with wavelets and pointwise nonlinearities is applied and further shown to satisfy the stability constraints. Note that the approach used to build this scattering network significantly differs from traditional CNNs, as no learning of the filters is involved. It should further be noted that while scattering transforms guarantee that representations built by deep neural networks are robust to small changes in the input, this does not imply that the overall classification pipeline (feature representation and discrete classification) is robust to small perturbations in the input, in the sense of (2). We believe that building deep architectures with provable guarantees on the robustness of the overall classification function is a fundamental open problem in the area.

Summary and open problems

The robustness of deep neural networks to perturbations is a fundamental requirement in a large number of practical applications involving critical prediction problems. We discussed in this article the robustness of deep networks to different forms of perturbations: adversarial perturbations, random noise, universal perturbations, and geometric transformations. We further highlighted close connections between the robustness to additive perturbations and geometric properties of the classifier's decision boundary (such as the curvature).

One of the main strengths of deep neural network classifiers with respect to traditional shallow classifiers is that the former achieve higher levels of invariance to transformations.

The importance of analyzing the vulnerability of deep neural networks to perturbations therefore goes beyond the practical security implications, as it further reveals crucial geometric properties of deep networks. We hope that this close relation between robustness and geometry will continue to be leveraged to design more robust systems.

Despite the recent and insightful advances in the analysis of the vulnerability of deep neural networks, several challenges remain:

- It is known that deep networks are vulnerable to universal perturbations due to the existence of correlations between different parts of the decision boundary. Yet, little is known about the elementary operations in the architecture (or learned weights) of a deep network that cause the classifier to be sensitive to such directions.
 - Similarly, the causes underlying the transferability of adversarial perturbations across different architectures are still not understood formally.
 - While the classifier's decision boundary has been shown to have a very small curvature when sectioned by random normal planes, it is still unclear whether this property of the decision boundary is due to the optimization method (i.e., stochastic gradient descent) or rather to the use of piecewise linear activation functions.
 - While natural images have been shown to lie very close to the decision boundary, it is still unclear whether there exist points that lie far away from the decision boundary.
- Finally, one of the main goals of the analysis of robustness is to propose architectures with increased robustness to additive and structured perturbations. This is probably one of the fundamental problems that needs special attention from the community in the years to come.

Authors

Alhussein Fawzi (fawzi@cs.ucla.edu) received the M.S. and Ph.D. degrees in electrical engineering from the Swiss Federal Institute of Technology, Lausanne, in 2012 and 2016, respectively. He is now a postdoctoral researcher in the Computer Science Department at the University of California, Los Angeles. He received the IBM Ph.D. fellowship in 2013 and 2015. His research interests include signal processing, machine learning, and computer vision.

Seyed-Mohsen Moosavi-Dezfooli (seyed.moosavi@epfl.ch) received the B.S. degree in electrical engineering from Amirkabir University of Technology (Tehran Polytechnic), Iran, in 2012 and the M.S. degree in communication systems from the École Polytechnique Fédérale de Lausanne (EPFL), Switzerland, in 2014. Currently, he is a Ph.D. degree student in the Signal Processing Laboratory 4 at EPFL under the supervision of Prof. Pascal Frossard. Previously, he was a research assistant in the Audiovisual Communications Laboratory at EPFL. During the spring and the summer of

2014, he was a research intern with ABB Corporate Research, Baden-Daettwil. His research interests include signal processing, machine learning, and computer vision.

Pascal Frossard (pascal.frossard@epfl.ch) received the M.S. and Ph.D. degrees in electrical engineering from the École Polytechnique Fédérale de Lausanne (EPFL), Switzerland, in 1997 and 2000, respectively. From 2001 to 2003, he was a member of the research staff with the IBM T.J. Watson Research Center, Yorktown Heights, New York, where he was involved in media coding and streaming technologies. Since 2003, he has been a faculty member at EPFL, where he is currently the head of the Signal Processing Laboratory. His research interests include signal processing on graphs and networks, image representation and coding, visual information analysis, and machine learning.

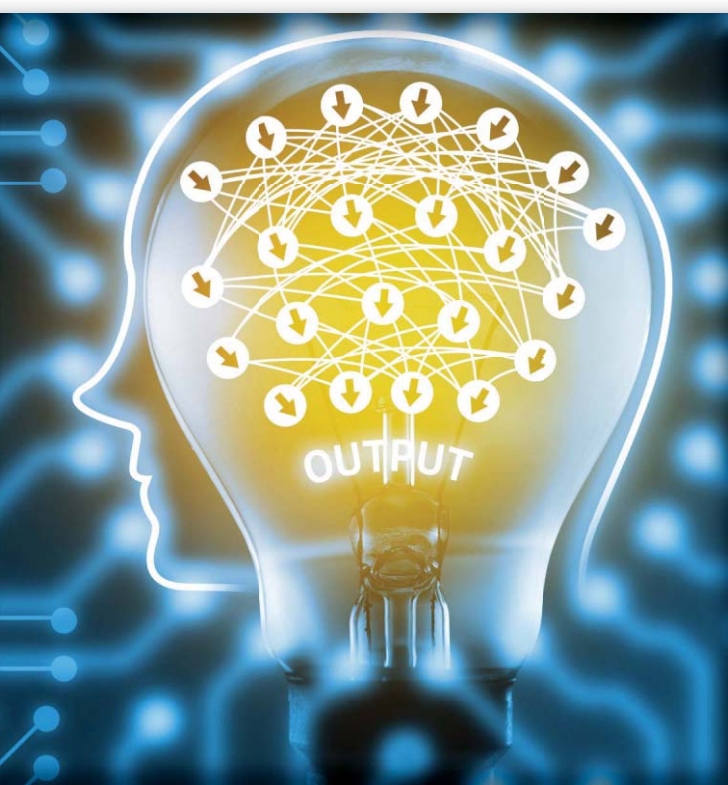
References

- [1] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus, "Intriguing properties of neural networks," in *Proc. Int. Conf. Learning Representations*, 2014.
- [2] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proc. Advances in Neural Information Processing Systems*, 2012, pp. 1097–1105.
- [3] A. Nguyen, J. Yosinski, and J. Clune, "Deep neural networks are easily fooled: High confidence predictions for unrecognizable images," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2015, pp. 427–436.
- [4] G. Litjens, T. Kooi, B. Ehteshami Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. W. M. van der Laak, B. van Ginneken and C. I. Sánchez, "A survey on deep learning in medical image analysis," *Med. Image Anal.*, vol. 42, pp. 60–88, 2017.
- [5] S.-M. Moosavi-Dezfooli, A. Fawzi, and P. Frossard, "Deepfool: A simple and accurate method to fool deep neural networks," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2016, pp. 2574–2582.
- [6] A. Fawzi and P. Frossard, "Manitest: Are classifiers really invariant?" in *Proc. British Machine Vision Conf.*, 2015, pp. 106.1–106.13.
- [7] A. Fawzi and P. Frossard, "Measuring the effect of nuisance variables on classifiers," in *Proc. British Machine Vision Conf.*, 2016, pp. 137.1–137.12.
- [8] A. Fawzi, H. Samulowitz, D. Turaga, and P. Frossard, "Adaptive data augmentation for image classification," in *Proc. Int. Conf. Image Processing*, 2016, pp. 3688–3692.
- [9] B. Biggio, I. Corona, D. Maiorca, B. Nelson, N. Srndic, P. Laskov, G. Giacinto, and F. Roli, "Evasion attacks against machine learning at test time," in *Proc. Joint European Conf. Machine Learning and Knowledge Discovery in Databases*, 2013, pp. 387–402.
- [10] A. Fawzi, O. Fawzi, and P. Frossard, "Analysis of classifiers' robustness to adversarial perturbations," *Machine Learning*, Aug. 2017. [Online]. Available: <https://doi.org/10.1007/s10994-017-5663-3>
- [11] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," in *Proc. Int. Conf. Learning Representations*, 2015.
- [12] A. Rozsa, E. M. Rudd, and T. E. Boult, "Adversarial diversity and hard positive generation," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition Workshops*, 2016, pp. 25–32.
- [13] N. Carlini and D. Wagner, "Towards evaluating the robustness of neural networks," *arXiv Preprint*, arXiv:1608.04644, 2016.
- [14] S. Baluja and I. Fischer, "Adversarial transformation networks: Learning to generate adversarial examples," *arXiv Preprint*, arXiv:1703.09387, 2017.
- [15] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell, "Caffe: Convolutional architecture for fast feature embedding," in *Proc. ACM Int. Conf. Multimedia*, 2014, pp. 675–678.
- [16] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2015.
- [17] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. Berg, and L. Fei-Fei, "Imagenet large scale visual recognition challenge," *Int. J. Computer Vision*, vol. 115, no. 3, pp. 211–252, 2015.
- [18] A. Fawzi, S. Moosavi-Dezfooli, and P. Frossard, "Robustness of classifiers: from adversarial to random noise," in *Proc. Neural Information Processing Systems Conf.*, 2016, pp. 1632–1640.
- [19] H. Larochelle, D. Erhan, A. Courville, J. Bergstra, and Y. Bengio, "An empirical evaluation of deep architectures on problems with many factors of variation," in *ACM Int. Conf. Machine Learning*, 2007, pp. 473–480.
- [20] M. Sharif, S. Bhagavatula, L. Bauer, and M. K. Reiter, "Accessorize to a crime: Real and stealthy attacks on state-of-the-art face recognition," in *Proc. 2016 ACM SIGSAC Conf. Computer and Communications Security*, 2016, pp. 1528–1540.
- [21] N. Karianakis, J. Dong, and S. Soatto, "An empirical evaluation of current convolutional architectures ability to manage nuisance location and scale variability," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2016, pp. 4442–4451.
- [22] S.-M. Moosavi-Dezfooli, A. Fawzi, O. Fawzi, and P. Frossard, "Universal adversarial perturbations," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2017.
- [23] Y. Liu, X. Chen, C. Liu, and D. Song, "Delving into transferable adversarial examples and black-box attacks," *arXiv Preprint*, arXiv:1611.02770, 2016.
- [24] N. Papernot, P. McDaniel, I. Goodfellow, S. Jha, Z. Berkay Celik, and A. Swami, "Practical black-box attacks against deep learning systems using adversarial examples," *arXiv Preprint*, arXiv:1602.02697, 2016.
- [25] D. Warde-Farley, I. Goodfellow, T. Hazan, G. Papandreou, and D. Tarlow, "Adversarial perturbations of deep neural networks," in *Perturbations, Optimization, and Statistics*. Cambridge, MA: MIT Press, 2016.
- [26] S. Sabour, Y. Cao, F. Faghri, and D. J. Fleet, "Adversarial manipulation of deep representations," in *Proc. Int. Conf. Learning Representations*, 2016.
- [27] T. Tanay and L. Griffin, "A boundary tilting perspective on the phenomenon of adversarial examples," *arXiv Preprint*, arXiv:1608.07690, 2016.
- [28] A. Fawzi, S.-M. Moosavi-Dezfooli, P. Frossard, and S. Soatto, "Classification regions of deep neural networks," *arXiv Preprint*, arXiv:1705.09552, 2017.
- [29] B. Poole, S. Lahiri, M. Raghu, J. Sohl-Dickstein, and S. Ganguli, "Exponential expressivity in deep neural networks through transient chaos," in *Proc. Advances in Neural Information Processing Systems Conf.*, 2016, pp. 3360–3368.
- [30] S.-M. Moosavi-Dezfooli, A. Fawzi, O. Fawzi, P. Frossard, and S. Soatto, "Analysis of universal adversarial perturbations," *arXiv Preprint*, arXiv:1705.09554, 2017.
- [31] S. Gu and L. Rigazio, "Towards deep neural network architectures robust to adversarial examples," *arXiv Preprint*, arXiv:1412.5068, 2014.
- [32] U. Shaham, Y. Yamada, and S. Negahban, "Understanding adversarial training: Increasing local stability of neural nets through robust optimization," *arXiv Preprint*, arXiv:1511.05432, 2015.
- [33] C. Caramanis, S. Mannor, and H. Xu, "Robust optimization in machine learning," in *Optimization for Machine Learning*, S. Suvrit, N. Sebastian, and W. J. Stephen, Eds., Cambridge, MA: MIT Press, ch. 14, 2012.
- [34] H. Xu, C. Caramanis, and S. Mannor, "Robustness and regularization of support vector machines," *J. Machine Learning Res.*, vol. 10, pp. 1485–1510, July 2009.
- [35] G. Lanckriet, L. E. Ghaoui, C. Bhattacharyya, and M. I. Jordan, "A robust min-max approach to classification," *J. Machine Learning Res.*, vol. 3, pp. 555–582, Dec. 2003.
- [36] C. Bhattacharyya, "Robust classification of noisy data using second order cone programming approach," in *Proc. Intelligent Sensing and Information Processing Conf.*, 2004, pp. 433–438.
- [37] T. B. Trafalis and R. C. Gilbert, "Robust support vector machines for classification and computational issues," *Optim. Methods Software*, vol. 22, no. 1, pp. 187–198, 2007.
- [38] N. Papernot, P. McDaniel, X. Wu, S. Jha, and A. Swami, "Distillation as a defense to adversarial perturbations against deep neural networks," in *Proc. 2016 IEEE Symp. Security and Privacy*, 2016, pp. 582–597.
- [39] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," *arXiv Preprint*, arXiv:1503.02531, 2015.
- [40] N. Carlini and D. Wagner, "Defensive distillation is not robust to adversarial examples," *arXiv Preprint*, arXiv:1607.04311, 2016.
- [41] A. A. Alemi, I. Fischer, J. V. Dillon, and K. Murphy, "Deep variational information bottleneck," *arXiv Preprint*, arXiv:1612.00410, 2016.
- [42] J. H. Metzen, T. Genewein, V. Fischer, and B. Bischoff, "On detecting adversarial perturbations," *arXiv Preprint*, arXiv:1702.04267, 2017.
- [43] R. Feinman, R. R. Curtin, S. Shintre, and A. B. Gardner, "Detecting adversarial samples from artifacts," *arXiv Preprint*, arXiv:1703.00410, 2017.
- [44] J. Lu, T. Issararone, and D. Forsyth, "SafetyNet: Detecting and rejecting adversarial examples robustly," *arXiv Preprint*, arXiv:1704.00103, 2017.
- [45] M. Jaderberg, K. Simonyan, and A. Zisserman, "Spatial transformer networks," in *Proc. Advances in Neural Information Processing Systems Conf.*, 2015, pp. 2017–2025.
- [46] A. Nayeibi and S. Ganguli, "Biologically inspired protection of deep networks from adversarial attacks," *arXiv Preprint*, arXiv:1703.09202, 2017.
- [47] M. Cisse, A. Courville, P. Bojanowski, and E. Grave, Y. Dauphin, and N. Usunier, "Parseval networks: Improving robustness to adversarial examples," in *Proc. Int. Conf. Machine Learning*, 2017, pp. 854–863.

Damien Teney, Qi Wu, and Anton van den Hengel

Visual Question Answering

A tutorial



©ISTOCKPHOTO.COM/ZAPP2PHOTO

The task of visual question answering (VQA) is receiving increasing interest from researchers in both the computer vision and natural language processing fields. Tremendous advances have been seen in the field of computer vision due to the success of deep learning, in particular on low- and midlevel tasks, such as image segmentation or object recognition. These advances have fueled researchers' confidence for tackling more complex tasks that combine vision with language and high-level reasoning. VQA is a prime example of this trend. This article presents the ongoing work in the field and the current approaches to VQA based on deep learning. VQA constitutes a test for deep visual understanding and a benchmark for general artificial intelligence (AI). While the field of VQA has seen recent successes, it remains a largely unsolved task.

Introduction

VQA involves an image and a related text question, to which the machine must determine the correct answer. This task spans the fields of computer vision and natural language processing, since it requires both the comprehension of the question and parsing the visual elements of the image. VQA is a practical setting to evaluate deep visual understanding, itself considered the overarching goal of the field of computer vision. Deep visual understanding can be defined as the ability of algorithm to extract high-level information from images and to perform reasoning based on that information. In this regard, VQA is an alternative to other tasks proposed to evaluate this capability. Examples include the visual Turing test [23], the task of image captioning [20], [73], and recent works on visual dialogs [18].

A second parallel motivation for the study of VQA is its utility in its own right. A system capable of answering questions about images has direct practical applications, such as a personal assistant, or in robotics as aids for the visually impaired. Note, however, that current VQA data sets do not directly address this setting, because questions are typically collected in a nongoal-oriented setting. Realistic, motivated

Digital Object Identifier 10.1109/MSP.2017.2739826
Date of publication: 13 November 2017

questions would likely require information not present in the image and involve rare words and concepts. In comparison, most questions in current data sets are purely visual (e.g., about counts or colors) and centered on common concepts. For example, in one of the most popular data sets [5], a mere 1,000 different answers can correctly answer more than 90% of questions.

The recent interest in VQA [5], [45], [81] originates from the latest advances in computer vision on low- and mid-level tasks. This encouraged further research on higher-level tasks, and the combination of vision with other modalities, particularly language. Historically, one of the earliest integrations of computer vision with language was the SHRDLU system dating back to 1972 [78], which allowed the use of language to instruct a computer to move objects in a simulated “blocks world.” Other attempts at creating conversational robotic agents [15], [47], [59] were also grounded in the visual world. However, these early works were often limited to specific domains and/or simple language. Deep learning has now been applied to virtually every problem imaginable in computer vision, and convolutional neural networks (CNNs) are approaching human performance in tasks such as image segmentation [39] or object recognition [19], [24]. The success of deep learning on perceptual tasks drove an increasing enthusiasm for high-level tasks. VQA particularly embodies this confidence in achieving high-level image understanding.

Task definition and data sets

An instance of VQA consists of an image and a related question given in plain text (see examples in Figure 1). The task for the machine is to determine the correct answer, which is, in current data sets, typically a few words or a short phrase. Two practical variants are usually considered, an open-ended and a multiple-choice setting [5], [92]. In the latter, a set of candidate answers are proposed. This makes the evaluation of a generated answer easier than in the open-ended setting, where the comparison between the machine’s output and a ground truth (i.e., human provided) answer faces issues with synonyms and paraphrasing.

In comparison to classical tasks of computer vision such as object recognition or image segmentation, instances of VQA cover a wide range of complexity. Indeed, the question itself can take an arbitrary form, and so can the set of operations required to answer it. In this sense, VQA more closely reflects the challenges of general image understanding. VQA is also related to the task of textual question answering [10], [14], [88], in which the answer is to be found in a textual narrative (i.e., reading comprehension) or in large knowledge bases (KBs) (i.e., information retrieval). Textual QA has been studied for a long time in the natural language processing (NLP) community, and VQA is basically its extension to a visual input. The additional challenge of a visual input is significant because images are simply much higher dimensional than text. Images capture the richness of the real world in a noisy manner, whereas natural language already represents a certain level of abstraction. For example, compare the phrase “a red hat” with the multitude of its representations that one could picture, e.g., with many different styles and details that cannot be described in a short phrase.

While, to some extent, the processing of language is possible with discrete- and rule-based approaches, such as syntactic parsers and regular expression matching, the complexity of images renders such engineered methods intractable. Modern computer vision is based on statistical learning, and recent works combining vision and language (including image captioning and VQA) similarly evolved from machine-learning techniques. Finally, both language and vision are inherently compositional in their structure. This constitutes both a challenge and an opportunity when considering the generalization capabilities of learned models (see the section “Compositional Models”).

Let us mention the relation of VQA with the task of automatic image captioning [20], [73], [79], i.e., generating a textual description of a given image. It has also attracted significant interest in the past few years and can be compared to VQA as they both combine vision and language. The two tasks are complementary as they evaluate different capabilities. Captioning requires mostly descriptive capabilities that involve almost purely visual information. VQA, in

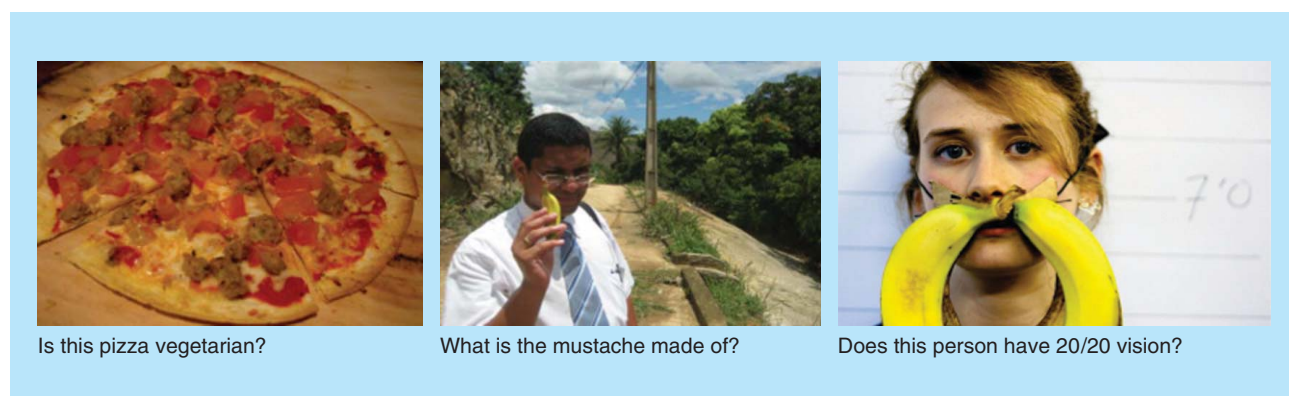


FIGURE 1. The task of VQA is a significant step toward general AI and a departure from low- and mid-level tasks in classical computer vision. It requires relating visual concepts with elements of language, common-sense, and general knowledge. (Photos are examples from a major public data set [5].)

comparison, often requires reasoning with common sense and with other information not present in the given image. In this respect, VQA constitutes an AI-complete task [5] since it requires multimodal knowledge beyond specific domains. This reinforces the motivation for research on VQA, as it provides a proxy to evaluate progress toward general AI, with systems capable of advanced reasoning combined with deep image and language understanding.

Data sets for training and evaluating VQA

We now examine data sets that have been specifically compiled for research on VQA. These data sets contain, at a minimum, triples made each of an image, a question, and its correct answer.

Some early data sets were generated semiautomatically (e.g., from image captions [45]) but modern data sets were created manually through crowdsourcing [5], [35]. The creation of these sets of questions with ground-truth answers is very time-consuming, and today's largest data sets of several hundreds of thousands of instances [35] represent a major effort. Those data sets are designed for both evaluating and training VQA systems in a supervised setting, and the latter demands such large amounts of data. As will be discussed in the section "Directions of Current and Future Research," this very need for large amounts of data is a significant limit of current approaches.

For the purpose of standardized comparisons and benchmarking of different algorithms, data sets are split into predetermined sets of instances for training, validation, and testing. Benchmarks typically do not provide the ground-truth answers of the test set. The evaluation is performed by an automatic online service that compares the provided answers (inferred by the algorithm to be evaluated) and the private ground truth [5]. This method typically restricts the number and frequency of submissions so as to prevent cheating or unintentional overfitting of the test set.

Existing data sets vary mainly along three dimensions 1) their size, i.e., the number and variety concepts represented in the images and questions; 2) the amount of required reasoning, e.g., whether the detection of a single object is sufficient or whether inference is required over multiple facts or concepts; and 3) how much information beyond what is present in the input image is necessary to infer an answer, e.g., common sense or subject-specific information. Most data sets lean toward visual-level questions and require little external knowledge beyond common sense. These characteristics reflect the fact that current state-of-the-art methods still struggle with simple visual questions.

The first VQA data set designed as a benchmark was Data Set for Question Answering on Real World (DAQUAR) for images [45]. The most popular modern data sets [5], [35], [92] use images sourced from Microsoft Common Objects in Context (COCO), [40] a data set initially devised for image recognition, which is itself composed of images from Flickr. Those images constitute a very diverse collection of photographs.

Despite undeniable advantages, VQA data sets of clipart images have seen little use compared to their counterparts of real images.

VQA-real

The most widely used data set is currently the one proposed by a team of researchers from Virginia Tech and is commonly referred to as *VQA* [5]. It comprises two parts, one using natural images named *VQA-real*, and a second one with clipart images named *VQA-abstract* (discussed at the end of this section). *VQA-real* comprises 123,287 training and 81,434 test images, respectively, sourced from COCO [40]. Human annotators were encouraged to provide interesting and diverse questions and short, concise answers (typically two to three words). The data set allows evaluation both in an open-ended and in a multiple-choice setting, the latter providing 17 additional (incorrect) candidate answers for each question. Overall, the data set contains 614,163 questions. According to an analysis performed by polling annotators, most subjects (at least six out of ten) estimated that some common sense was required for 18% of the questions, and adult-level knowledge was necessary for only 5.5% of the questions. These figures show that purely visual information is likely sufficient to answer most questions.

A recent, updated version of this data set, known as *VQA v2.0*, includes two images with each question that lead to different answers [25]. This aims at addressing issues of data set biases.

Visual genome and visual7W

The Visual Genome QA data set [35] is currently the largest one designed for VQA, with 1.7 million question/answer pairs. It is built with images from the Visual Genome project [35], which includes structured annotations of scene contents in the form of scene graphs. Those scene graphs describe the visual elements of each image with their attributes and the relationships between them. Human subjects provided questions that must start with one of the seven "Ws"—i.e., who, what, where, when, why, how, and which. The diversity of answers in the Visual Genome is larger than in *VQA-real* [5]. The 1,000 most-frequently given answers in the data set correspond only to the correct answers of 64% of all questions. In *VQA-real*, the corresponding top 1,000 answers cover more than 90% of questions. The Visual7w [92] data set is a subset of the Visual Genome that allows evaluation in a multiple-choice setting, as each question is provided with four plausible but incorrect candidate answers.

Zero-shot VQA

A special version of the Visual7W data set was proposed in [70]. The authors redefined the training and test splits such that every test instance includes one or several words that were not present in any training example. For example, a test question "How many zebras are in the image?" might arise even though the word *zebra* was never used in the training set. The evaluation of an algorithm with this data set emphasizes its capabilities for generalization beyond training examples and for using sources of information other than VQA-specific data sets. Another similar study appeared in [54].



FIGURE 2. Examples from the test splits of different VQA data sets. For the zero-shot VQA data set, the highlighted words are unknown words, i.e., not present in training examples.

Clipart images

Data sets for VQA have also been proposed with synthetic clipart images (referred to as *abstract scenes* in [5]). These images were created manually with cartoon representations of characters and objects from a predefined set. The motivation is to enable research on VQA in a controlled setting, where the computer vision part of the problem is eased by the restricted set of visual elements. Such data allows focusing on the high-level semantics of the scenes rather than on visual recognition. For this purpose, the images are provided with structured descriptions, in the form of XML files that list the objects present in the scene with their visual properties (e.g., position, scale, etc.). VQA methods can use these descriptions to completely bypass the visual parsing of the images.

Using synthetic images gives great control over the elements actually depicted, and this allowed the creation of a data set of balanced binary questions [90]. That data set contains only binary (yes/no) questions and each question appears twice in the data set, with two different images that give rise to opposite answers. This removes conditional biases that are common in other data sets, for example, a predominance of “yes” answers to questions of the form “Is there ... in the image?” Those biases otherwise allow to blindly guess correct answers, which hinders a meaningful evaluation of VQA systems. Despite undeniable advantages, VQA data sets of clipart images have seen little use [5], [69], [90] compared to their counterparts of real images.

Video-based QA

In addition to the studies on image QA mentioned previously, there have been a few works on VQA with videos. Zhu et al. [91] assembled a data set of over 100,000 videos and 400,000 questions, using existing collections of videos from different domains, from cooking scenarios to movies and web videos. Tapaswi et al. [67] proposed a setting named MovieQA, where questions have to be answered using multiple sources of information including he full-length movies, but also sub-

titles, scripts, and plot summaries. Zeng et al. [89] proposed the generation of questions from video descriptions.

Evaluation

VQA systems are evaluated by inferring the answers on the test split of a given data set. Recent data sets [92] recommend the multiple-choice setting, since there is only one correct answer among the multiple choices. The evaluation is thus straightforward, as one can simply measure the mean accuracy over test questions. In an open-ended setting, several answers could be equally valid, because of synonyms and paraphrasing. This makes a fair evaluation nontrivial. The usual workaround is to restrict answers, at the time of the creation of the data sets, to short phrases, typically one to three words. This restriction limits ambiguities by forcing questions and answers to be more specific, and allows evaluation by exact string-matching. Most data sets partition the test questions into subsets depending on the type of answer (e.g., yes/no, number, etc.) such that performance can be reported on each subset (see Table 1).

Deep neural networks for VQA

The common approach to VQA is to train a deep neural network with supervision which maps the given image and question to a relative scoring of candidate answers. The main idea is to learn a joint embedding of the visual and textual inputs. First, the image and the question are processed independently to obtain separate vector representations (see Figure 3). Those features are then are mapped with learned functions to a joint space, then combined and fed to an output stage. We examine each of those elements next. The section “Advanced Techniques” will then look at those techniques that build onto this model.

Image encoding

On the computer vision side, the input image x^I is processed with a deep convolutional neural network (CNN) to extract image features described as a vector y^I . This large fixed-size

Table 1. A selection of results on the VQA-real data set (test-std split) in both the open-ended and multiple-choice settings. Performance has incrementally improved over the past few years. The highest accuracies per column are in boldface.

Method	Yes/No	VQA-Real Open Ended			Multiple Choice
		Numbers	Other	All	All
Baseline: Deeper LSTM Q norm. I [42]	80.6	36.5	43.7	58.2	63.1
Neural modules networks [4]	81.2	37.7	44	58.7	—
Stacked attention networks [87]	—	—	—	58.9	—
Dynamic memory networks (DMNs+) [83]	—	—	—	60.4	—
DualNet [60]	81.9	37.8	49.7	61.7	66.7
Hierarchical coattention (HiCoAtt) [43]	—	—	—	62.1	66.1
VQA-machine [74]	81.4	38.2	53.2	63.3	67.8
MLB [34]	84	37.9	54.8	65.1	68.9
MCB ensemble 7 models [21]	83.2	39.5	58	66.5	70.1

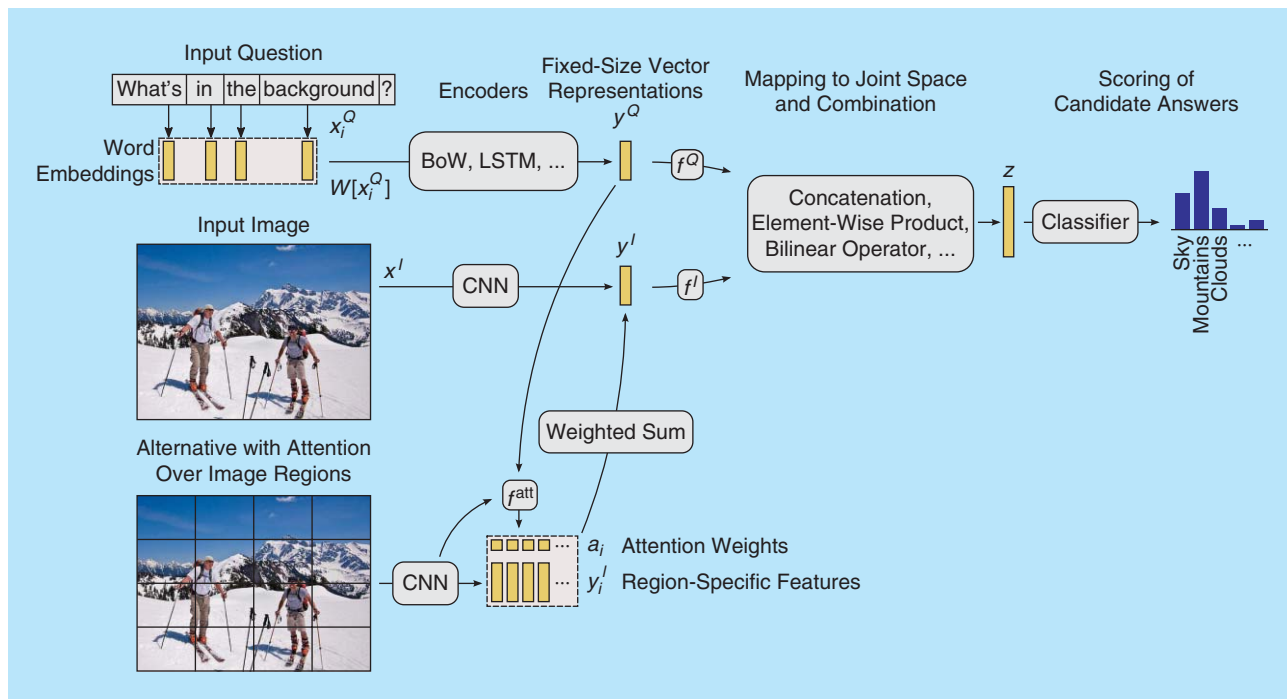


FIGURE 3. The common approach to VQA is to train a deep neural network for classification over a large set of candidate answers (see the section “Deep Neural Networks for VQA”). The input question and image are encoded into fixed-size feature vectors (orange bars), using the word *embeddings* and a CNN, respectively. The resulting representations are mapped into a joint space, then combined and passed on to the classifier. It assigns scores to a large set of candidate answers. The top-ranking candidate is returned as the final answer. An attention mechanism (see the section “Attention Mechanisms”) can improve this model and allows the model to focus on relevant parts of the image. In that case, the CNN extracts region-specific image features and aggregates them using scalar weights (orange squares).

vector encodes the contents of the image. This CNN is typically a standard network architecture that has been pretrained to perform image recognition [36]. The motivation for a pretrained network is to take advantage of the vast amounts of training data available for image recognition, relative to the amounts of data annotated for VQA. The pretrained network is used as a generic feature extractor, by discarding the final classification layers, and using the features produced within the CNN prior to this classification [55]. In comparison to classical handcrafted image features such as scale-invariant feature transform (commonly known as *SIFT*) [41] or histogram of oriented gradients (commonly known as *HOG*) [16], CNN features provide higher-level representations of the contents of the image, and are naturally produced as a fixed-size vector. The size of this vector is typically in the order of 1,024 or 2,048.

Question encoding

On the language side, the input question is also processed to obtain a fixed-size representation of its contents. Initially, the i th word of the question is represented by an index x_i^Q in the input vocabulary. Each word is then turned into a vector. This uses a mapping implemented as a lookup table $W[\cdot]$ that associates the index of any word of the input vocabulary to a learned vector. An alternative implementation initially represents each word with a one-hot vector (a vector of all zeros, except for a one at the location of the word index in the vocabulary), which is then multiplied with a dense weight

matrix that contains the embeddings of all words. The vectors of all words $W[x_1^Q], W[x_2^Q], \dots, W[x_N^Q]$ are then collapsed into a single vector. A simple option for this purpose is to make a bag-of-words (BoW), which corresponds to simply averaging the word vectors, i.e., $y^Q = (1/N) \sum_i W[x_i^Q]$. Another popular option is to feed the word vectors into a recurrent neural network (RNN) such as a long short-term memory (LSTM). An RNN processes words sequentially and can capture the sequential relationships between them. In comparison, a BoW does not account for word order, and, for example, would produce a same representation for “this man eats a hot dog” and “a hot man eats this dog.”

Combination of image and question features

The feature vectors y^I and y^Q represent the image and the questions, respectively. They are each passed through a learned function before being combined. The intuition here is to map the features to a joint space, in which distances between both modalities become comparable. The learned functions $f^I(\cdot)$ and $f^Q(\cdot)$ are typically implemented as additional layers of the neural network, e.g., $f(y) = \text{ReLU}(Wy + b)$, where W and b are learned weights and biases, and ReLU is a rectified linear unit that serves as a nonlinearity. The mapped features are then combined before being fed to the output stage. A simple option for this combination is to simply concatenate them as $z = [f^I(y^I); f^Q(y^Q)]$. Alternatively, it is popular to include multiplicative interactions within the neural network

to increase its capacity and use $z = f^l(y^l) \cdot f^q(y^q)$, where \cdot is the Hadamard (element-wise) product.

Output

The output stage of a VQA system can be seen either as a generation or as a classification task. The generation of a free-form answer has the advantage of being able to compose complex sentences. In practice however, such a model is difficult to learn [22], [46], [80]. Current data sets are limited to short answers, and a practical alternative is to rather learn a classifier over candidate answers [22], [44], [46], [57]. For this purpose, a large set of candidate answers is predetermined from the most common ones in the training set (typically in the order of 2,000). This inevitably leaves out some infrequent words, but such a set is typically sufficient to answer correctly more than 90% of test questions [5]. This is a nonlimiting issue since this figure is well above the accuracy of current systems. The combined features z are passed to a classifier over those candidate answers (a linear layer followed by a softmax [21] or sigmoid transformation [30]). The classifier assigns score to each candidate answer, and the top-ranked one is returned as the final output. In a multiple-choice setting, only the scores assigned to proposed choices are considered. For training the model, the classifier is followed by a cross-entropy loss, and the whole network is trained end-to-end by backpropagation to minimize this loss over the set of training examples.

Variations

A vast array of variations on the method presented previously have been proposed in the literature. Here are some examples.

- Encoding the question and the image with a single recurrent neural network (an LSTM) by passing the image features together with each word embedding [22] or only once prior to the question words [46], [57].
- Encoding the question with a bidirectional RNN, i.e., two LSTMs that process the words in forward and backward order, respectively. This aims at capturing the language structure with more uniform importance on the beginning and the end of the question [57].
- Adding additional multiplicative interactions within the network and between the features of the image and of the question. For example in [51], the authors present their “DPPnet” model as a way of dynamically adapting the computations applied on the image features based on the question (one branch of the network computes weights that are then multiplied with the inputs in another branch). Such interpretations are typical of deep-learning models, but have little concrete support. Performance benefits usually stem simply from the additional capacity of the network.
- Alternative schemes for combining image and question representations, such as element-wise sums and products [33], bilinear operations [30] such as multimodal compact bilinear pooling (MCB) [21], etc.
- Gradual increases in performance of the state of the art is also explained by increasingly better pretrained CNNs to

provide image features, and by the application of general enhancements for neural network architectures, such as highway networks and residual networks [33], dropout, batch normalization, etc.

Advanced techniques

In this section, we review popular improvements to the general approach described so far.

Attention mechanisms

One of the most effective improvements to the joint embedding model is to use visual attention. Humans have the ability to quickly understand visual representations by attending to regions of the image instead of processing the entire scene at once [58]. Mimicking human attention in deep neural networks has been applied with success to machine translation [8], reading comprehension [63], textual question answering [84], object recognition [64] and image captioning [86], and is also used in most modern VQA models (e.g., in [43] and [87]).

The main idea behind attention mechanisms is to allow the model to focus on certain regions of the image. The technique involves 1) using region-specific image features and 2) including multiplicative interactions within the neural network. The aforementioned basic VQA model described uses a CNN to extract a global feature vector y^l that describes the whole image. This can contain irrelevant or noisy information. Instead, we now extract local features $\{y_i^l\}_i$ for different regions $i = 1 \dots M$ of the image. Those features are obtained from an earlier layer in the pretrained CNN, prior to the last spatial pooling. The network computes a scalar attention weight a_i for each region using both the region and the question features, i.e., $a_i = f^{\text{att}}(y_i^l, y^q)$. The function $f^{\text{att}}(\cdot)$ is learned and implemented as additional layers of the network. The attention weights can be interpreted as the relevance of a given region, and the image is finally represented by a weighted sum of the region features, i.e., $y^l = \sum_i a_i y_i^l$.

The attention weights computed for a given question/image can be visualized in the form of “attention maps” for purposes of introspection into the VQA model. Each a_i corresponds to a specific region of the input image, and those values are overlaid onto the image canvas (see Figure 4).

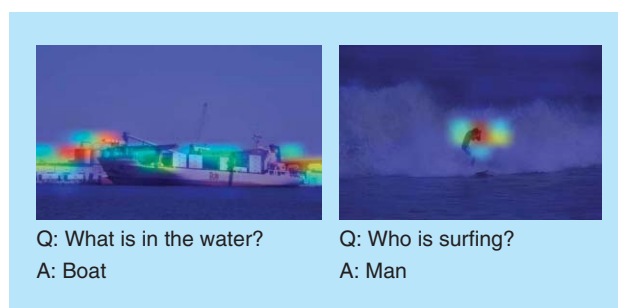


FIGURE 4. Attention weights are often visualized as spatial maps overlaid on the input image (warmer colors correspond to higher weights). They are interpreted as the importance given by the model to different regions of the image (examples used with permission from [74]).

They are interpreted as the importance given by the model to each image region.

The use of an attention mechanism has shown to be very beneficial and is now common practice. Variations on this principle have been proposed. For example, [85] and [87] use multiple rounds of visual attention to allow focusing on several regions. In [85], a two-step process performs a word-guided attention, then a question-guided one. In [65], the authors define image regions with object proposals and then select the regions most to the question and to given answer choices. In [43], the authors propose a “hierarchical coattention” (HieCoAtt) that performs a question-guided attention on the image and an image-guided attention on the question.

The overall idea of attention in neural networks was initially motivated by an analogy to the human visual system. Even though the model is capable of modeling a behavior similar to human attention, this only constitutes an interpretation. In a neural network-trained end to end, nothing enforces the attention mechanism to actually reflect human-like behavior. In a recent study [17], Das et al. compared the attention used by human subjects presented with VQA problems, and VQA models with attention [43], [87]. Their conclusion was a systematically low correlation.

Pretraining language representations

As described in the section “Deep Neural Networks for VQA,” the first step for encoding the question is to map words to vector representations called word *embeddings*. Each word of the input vocabulary (i.e., any word appearing in the training set) is associated with its own embedding, and those embeddings are normally learned alongside the other parameters of the network via backpropagation. Two potential issues can arise, however. First, word occurrences in any data set typically follow a long-tailed distribution, meaning that a majority of words occur infrequently. It is thus difficult to learn stable and meaningful embeddings for those rare words. Second, the long-tail property, at its extreme, means that it words commonly appear in test questions that were not seen in any training example. Embeddings for those words cannot be learned from those examples, and they are typically associated with an special vector (of zeros or of a special “unknown” token), and their meaning is practically discarded from the questions.

A solution to these issues is to pretrain word embeddings on a larger auxiliary data set. This practice is known in the field of natural language processing and has shown benefit in many tasks besides VQA. Popular methods for pretraining word embeddings include Global Vectors for Word Representation [53] (GloVe) and word2vec [48], which we outline next. The general principle is to use a large, auxiliary training set of unannotated text, such as news articles and *Wikipedia* pages. Those methods require no specific annotations. That data can thus be much larger than the training set used for VQA and involve a much larger vocabulary.

The idea in the skip-gram model of *word2vec* is to train a model which, using the representation (i.e., an embedding) of a given word, is predictive of the context, i.e., the neighboring words in which it frequently appears [49]. As a consequence, words that are interchangeable or appear in similar contexts become associated with similar embeddings. Distances between embeddings thus naturally capture semantic relatedness between the words they represent.

More precisely, the skip-gram model seeks to maximize the ability to predict, from each word embedding, the occurrences of other words in a small surrounding window. The objective function to be maximized is

$$J = \frac{1}{N} \sum_i \frac{1}{|\Omega(i)|} \sum_{j \in \Omega(i)} \log p(x_j | x_i), \quad (1)$$

where i indexes the N -ordered words in the training corpus, x_i is the index in the vocabulary of word i , $\Omega(i)$ is a context window of fixed size around word i in the corpus [49]. The conditional probability $\log p(x_j | x_i)$ is modeled as a compatibility measure between embeddings such as a dot product followed by a sigmoid, i.e.,

$$p(x_j | x_i) = 1/(1 + e^{-W[x_i] W[x_j]}), \quad (2)$$

where $W[\cdot]$ is a lookup table containing the embeddings of all words in the vocabulary, reusing the notation of the section “Deep Neural Networks for VQA.” After

the training, the context-prediction part of the model is discarded, and the embeddings associated with the words are retained (i.e., the table $W[\cdot]$) and used as word embeddings in the downstream application such as VQA. The embeddings can be used as “frozen weights,” i.e., static representations associated with the words, or they can serve as initial values to be subsequently fine-tuned, i.e., optimized with a lower learning rate relative to the other network parameters.

Using pretrained embeddings helps the generalization capabilities of a VQA model. Since semantically similar words are mapped to close points in the word embedding space, the processing by the subsequent layers of the network can more easily 1) interpolate across concepts and 2) generalize to words absent from training questions but for which embeddings were pretrained.

Memory-augmented neural networks

An active research area is the design of deep neural networks that include an internal memory [13], [52], [66], [77]. Memory-augmented networks have shown success on tasks such as textual question answering [28], reading comprehension [37], and VQA [83]. The general idea of memory-augmented networks is to maintain an internal representation of the input data, on which multiple read and write operations can be applied. The composition of multiple operations can potentially execute complex chains of inference on the data. A “controller” part of the network is responsible for

One of the most effective improvements to the joint embedding model is to use visual attention.

controlling those operations. The mechanism is comparable to multiple rounds of an attention mechanism, in that it also enables the modeling of interactions between specific section of the input data.

The variant proposed in [37] and [83], named *dynamic memory networks (DMNs)*, was successfully applied to VQA. It is built around four modules (see Figure 5). The input module transforms the input data into a set of discrete vectors called *facts*. A question module computes a vector representation of the question, using a gated recurrent unit [(GRU), a variant of LSTM]. An episodic memory module retrieves the facts required to answer the question. A key element is to allow the episodic memory module to perform multiple passes over the facts to allow transitive reasoning. An attention mechanism selects the relevant facts and an update mechanism iteratively generates new memory representations from the current state and the retrieved facts. The initial state is set as the representation produced by the question module. Finally, the answer module uses the final state of the memory and the question to predict the final output, using a classic classifier over candidate answers.

Run time retrieval of additional information

Interfacing a VQA method with external sources of information allows one to separate the reasoning from the representation of prior knowledge in a scalable manner. One limitation of the basic joint embedding approach is to attempt to capture all of the information of training examples within the parameters of a neural network. This cannot scale arbitrarily, however. On one hand, any network has a finite capacity and, on the other hand, training examples also provide finite information. Several works explored the idea of connecting a VQA system with external sources of information that can be virtually infinite (e.g., web searches) or extensible without needing to retrain the VQA model (e.g., structured KBs).

In [75] and [82], the authors train a model to interface with a KB. Such KBs, like DBpedia [7] and Freebase [12], are databases compiled with facts ranging from common sense to encyclopedic knowledge. Such nonvisual information can be helpful for VQA. For example, the question “How many mammals appear in this image?” requires understanding the word “mammal” and which animals belong to this category. The VQA system of [75] and [82] is trained to map the input question/image to queries to be executed on KBs. The queries retrieve information relevant to the concepts involved in the question and/or image, which is fed as an additional input to the output stage of the system. The overall principle has shown limited benefits on existing VQA data sets, since most questions do not require such specific, nonvisual information. The idea remains a promising direction for developing scalable VQA systems.

In [70], the authors propose the retrieval of visual information from web searches in the form of exemplar images of question words. Rare and novel words, for example, the name of an uncommon animal or of an up-and-coming celebrity, are not likely to appear or be even known during training. The retrieval of images from the web allows the method to expand its domain of applicability as needed. The implementation of [70] simply retrieves the top five images from Google for every word of the question, from which CNN features are extracted and fed alongside the input question/image to the VQA system. This mechanism, however crude, showed an advantage to questions involving unknown words (i.e., “zero-shot VQA”) while leaving substantial room for future developments; see the section “Issues with Unknown and Novel Words.”

Directions of current and future research

Most modern methods for VQA have been evaluated on the data set of Antol. et al., which has served as the de facto standard benchmark. State-of-the-art methods have consistently improved performance on this data set over the past few years, from an

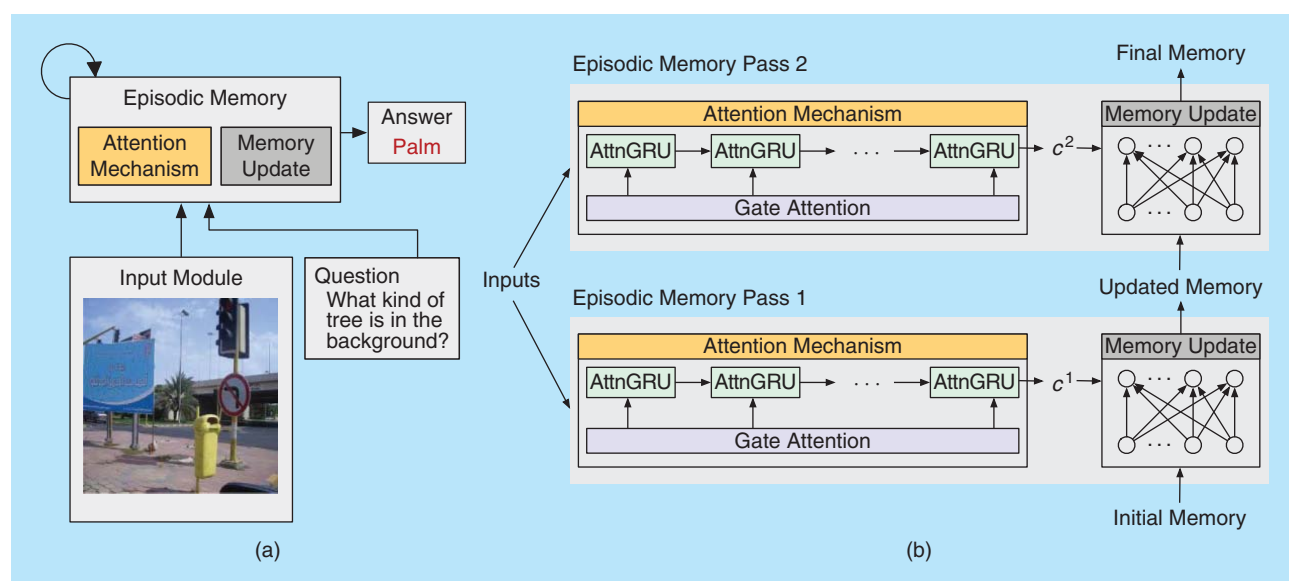


FIGURE 5. DMNs for VQA. (a) The overview and (b) details of the episodic memory module with two passes. (Figure adapted with permission from [83].)

accuracy of about 58% to over 70% today (see Tables 1 and 2 for a selection of results). These improvements have been incremental and have now seemed to plateau. In the following, we examine how current evaluations can mask some inherent issues of today's approaches and examine promising directions to bring future breakthroughs.

Issues of data set biases

Several studies have recently pointed out a fundamental issue with VQA data sets [25], [30], [90]. The text questions alone often provide strong cues that can be sufficient to answer them correctly, with no regards to the contents of the input image. These cues can be obvious. For example, questions starting with "Do you see a ..." can be correctly answered with a "yes" almost nine times out of ten [25]. These cue can also stem from an imbalance among possible answers. For example, questions starting with "How many ..." often have a correct answer of "one" or "two" but rarely "17." This issue can also be more subtle and manifest in the form of conditional biases. For example, we could imagine that questions starting with "What is the color ..." can often be answered correctly with "gray" if it also contains the word "car" and "red" if it contains the word "flower." Biases conditioned on image contents are also likely and yet more subtle. Biases are inherent to the real world, and it is desirable for a VQA model to capture and exploit them to some extent. However, today's methods have been shown to overly rely on data set biases and essentially be reduced to rote-learning of training questions. This is counterproductive to the objective of evaluating visual understanding. A blinded VQA model (i.e., not being shown the input image, and only guessing from the question) still achieves an accuracy of 56% versus 65% in the nonblinded case [30].

The issue of data set biases has been recognized. Attempts at addressing it include balanced data sets. Zhang et al. [90] first proposed a data set of clipart images where each binary question is accompanied by two different images that elicit "yes" and "no" answers, respectively. Goyal et al. applied the idea to real images, associating two images with each question that lead to different answers (see example in Figure 2). An appropriate performance metric in this case is to measure accuracy on pairs

of scenes. Blind models in this case would obtain an accuracy of 0%, and random guessing 25%. The use of balanced data sets encourages VQA models, to a larger extent, to utilize visual information instead of relying on language cues and data set biases. It is expected that future evaluations of algorithms on those data sets will be more representative of actual progress on visual understanding.

Issues with unknown and novel words

A VQA method to be used in a real-world setting, e.g., in robotics or as personal AI assistants, must be applicable to open, unrestricted domains. The current paradigm of training VQA systems with supervision, i.e., with data sets of questions and their ground-truth answers, can only cover a limited set of objects and concepts. Although VQA data sets have grown in size, no finite set of exemplars will ever cover the diversity of objects, actions, relations, etc. in the real world, for which an ideal VQA system should be prepared. A secondary issue with the current approach is the incentive for published methods to perform well on benchmark data sets. These benchmarks do not encourage addressing rare words and concepts, but rather focus on the concepts most frequent in the data set. Current methods are therefore designed to best learn—and often overfit—data set biases.

Recent works have argued for addressing a setting named *zero-shot VQA* [54], [70], where questions (or the proposed multiple-choice answers) specifically involve words that have not been seen in any training question. For example, a question "How many zebras are in the image?" may arise, even though no zebra was involved in the training set. This setting requires strong generalization capabilities. For example, a related training question "How many giraffes are in the image?" should be taken as an opportunity to learn to count, although not giraffes specifically. In parallel of works on VQA, the learning of high-level reasoning is addressed in the more abstract setting of program induction (see, e.g., [56]). We expect that VQA will ultimately require similar principled approaches, such as differentiable computing [26], [50], rather than brute-force learning from limited sets of examples.

Table 2. A selection of results on the newer VQA v2 data set (test-std split; open-ended questions). Baseline methods score lower on this harder data set, but the state of the art now reaches more than 70% of accuracy on open-ended questions. The highest accuracies per column are in boldface.

Method	VQA v2 Open Ended			
	Yes/No	Numbers	Other	All
Baseline: deeper LSTM Q norm. I [42]	73.46	35.18	41.83	54.22
MCB [21]	78.82	38.28	53.36	62.27
UPMC-LIP6 [9]	82.07	41.06	57.12	65.71
Athena [1]	82.50	44.19	59.97	67.59
LV-NUS [1]	81.89	46.29	58.30	66.77
HDU-USYD-UNCC [1]	86.65	51.13	61.75	70.92
Tips and Tricks VQA [2], [68]	86.60	48.64	61.15	70.34

External knowledge

The setting of the previously mentioned zero-shot VQA exposes the need for VQA systems to apply to concepts not present in training question/answers. This motivates the use of other kinds of data for training, and for retrieving additional information as needed at test time. This requires the system not only to capture actual information from training examples, but to learn to retrieve and use novel information, i.e., learn to learn. That capability of metalearning receives increased attention [11], [61], [72]. In the context of VQA, [70] showed the benefit of retrieving on-the-fly, exemplar images of unknown words from an online search engine. In [75] and [76], the authors showed the benefit of answering questions requiring background knowledge of retrieving additional information from a structured KB. The extension of these ideas is a promising research direction.

Modular approaches

Most current VQA models use a monolithic neural network and end-to-end supervision to learn the representations of data, the reasoning process, and to capture background knowledge from training examples. Alternatively, modular approaches have been explored [74], [80] with the goal of explicitly factoring the overall process of VQA into distinct subtasks. The principle of modularity allows one to decouple subtasks to some extent, and to use intermediate supervision and leverage several types of training data, as opposed to only “end-to-end” question/answer pairs. The use of pretrained word embeddings (see the section “Pretraining Language Representations”) is a very successful example of this general principle. Word embeddings are pretrained to capture language-based semantic similarities, and, in a similar spirit, other representations could be pretrained from auxiliary data to capture visual similarities [38] and other kinds of background information [71].

Modular systems for VQA also allow decoupling, to some degree, the visual perception from the high-level reasoning. For example, Wang et al. [74] proposed a VQA model on top of a collection of computer vision algorithms that detect visual elements such as objects, persons, and relations between them. Thereby, the VQA model only has to reason over this explicit high-level representation of the contents of the image.

Compositional models

The compositional nature of images and language lends itself to learning similarly compositional models [6]. The approach aims at addressing the challenge of generalization, i.e., applying the learned model to novel compositions of words and visual elements. Compositional models were proposed by Hendricks et al. on the task of image captioning [27]. Andreas et al. [4], [3], [29] were the first to propose a compositional architecture for VQA, named *neural module networks*. In their approach, the input question is processed

to determine the set of operations required to answer the question. A deep neural network is assembled with trained modules, each corresponding to one of those operations. A custom network is thus tailored specifically to each question, and finally applied on the image to infer the answer.

A data set of synthetic images named *CLEVR* (which stands for *compositional language and elementary visual reasoning*) [31] was specifically designed to evaluate generalization to novel combinations in VQA. It contains photorealistic images of shapes of various colors and materials. The data set also contains annotations describing the kind of reasoning that each question requires (i.e., as functional “programs”). The data set spurred a series of works on compositional models [29], [32].

The extra annotations facilitate the training of compositional models by serving as an intermediate supervision signal. This supervision correspond to an arrangement of modules or operations to be executed for each question. All of the aforementioned works demonstrated unique capabilities on synthetic data sets. However, it is still unclear how to best apply them to real images and how to train them only using end-to-end supervision, i.e., only knowing the final answer.

An alternative approach that addresses compositionality is the relational networks [62]. The idea is to consider the input as a set of objects, such as the locations in a CNN feature map, and to learn a common predictor that is applied to pairwise combinations of those objects. The predictor basically learns the relations between parts of the input. This proved effective on the CLEVR data set without the need for the intermediate supervision mentioned previously.

Conclusions

This article presented a review of the state of the art on visual question answering. We reviewed popular approaches based on deep learning, which treat the task as a classification problem over a set of candidate answers. We described the common joint embedding model, and additional improvements that build up on this concept, such as attention mechanisms. Despite shortcomings of current practices for both training and evaluating VQA systems, we identified a number of promising research avenues that could potentially bring future breakthroughs for both VQA and for the general objective of visual scene understanding.

Acknowledgment

All correspondence regarding this article should be addressed to Qi Wu at qi.wu01@adelaide.edu.au.

Authors

Damien Teney (contact@damienteney.info) obtained his B. Sc. degree in 2007, his M.Sc. degree in 2009, and his Ph.D. degree in 2013, all in computer science from the University of Liege, Belgium. He is a postdoctoral researcher at the

Most current VQA models use a monolithic neural network and end-to-end supervision to learn the representations of data, the reasoning process, and to capture background knowledge from training examples.

Australian Centre for Visual Technologies of the University of Adelaide, where he works on computer vision and machine learning. He was previously affiliated with Carnegie Mellon University, Pittsburgh, Pennsylvania; the University of Bath, United Kingdom; and the University of Innsbruck, Austria.

Qi Wu (qi.wu01@adelaide.edu.au) received a bachelor's degree in mathematical sciences from China Jiliang University, Hangzhou, and a master's degree in computer science and a Ph.D. degree in computer vision from the University of Bath, United Kingdom, in 2012 and 2015, respectively. He is a postdoctoral researcher at the Australian Centre for Robotic Vision of the University of Adelaide. His research interests include cross-depiction object detection and classification, attributes learning, neural networks, and image captioning.

Anton van den Hengel (anton.vandenhengel@adelaide.edu.au) received his bachelor's degree in mathematical science in 1991, his bachelor of laws degree in 1993, his master's degree in computer science in 1994, and his Ph.D. degree in computer vision in 2000, all from the University of Adelaide, Australia, where he is a professor and the founding director of the Australian Centre for Visual Technologies.

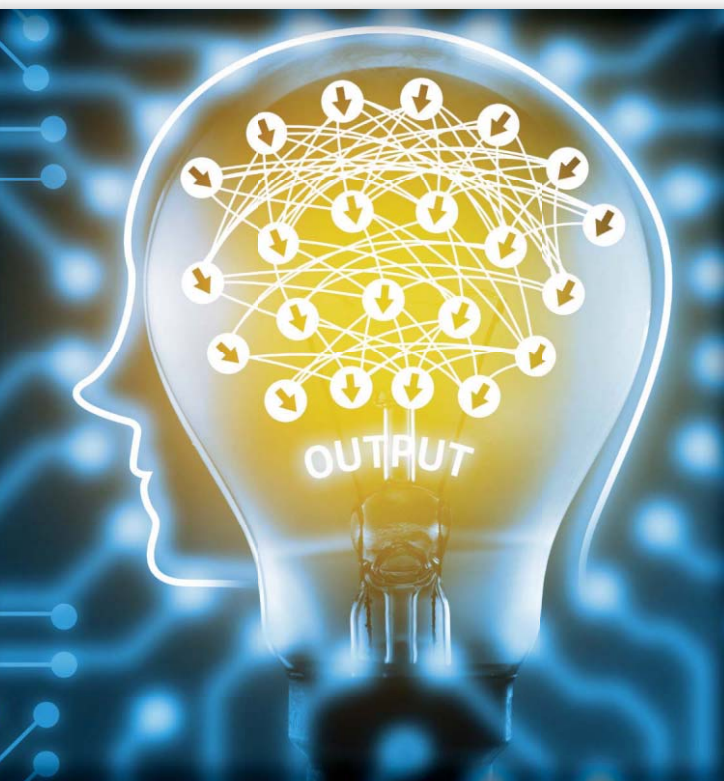
References

- [1] VQA challenge leaderboard. [Online]. Available: <http://visualqa.org/> <http://eva.lai.cloudev.org>
- [2] P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, and L. Zhang, "Bottom-up and top-down attention for image captioning and VQA," *arXiv Preprint*, arXiv:1707.07998, 2017.
- [3] J. Andreas, M. Rohrbach, T. Darrell, and D. Klein, "Learning to compose neural networks for question answering," in *Proc. Annu. Conf. North American Chapter Assoc. Computational Linguistics*, San Diego, CA, 2016, pp. 1545–1554.
- [4] J. Andreas, M. Rohrbach, T. Darrell, and D. Klein, "Neural module networks," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2016, pp. 39–48.
- [5] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. L. Zitnick, and D. Parikh, "VQA: Visual question answering," in *Proc. IEEE Int. Conf. Computer Vision*, 2015, pp. 2425–2433.
- [6] Y. Atzmon, J. Berant, V. Kezami, A. Globerson, and G. Chechik, "Learning to generalize to new compositions in image understanding," *arXiv Preprint*, arXiv:1608.07639, 2016.
- [7] S. Auer, C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak, and Z. Ives, *DBpedia: A Nucleus for a Web of Open Data*. New York: Springer, 2007.
- [8] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," in *Proc. Int. Conf. Learning Representation (ICLR)*, San Diego, CA, 2015.
- [9] H. Ben-younes, R. Cadène, M. Cord, and N. Thome, "MUTAN: multimodal tucker fusion for visual question answering," *arXiv Preprint*, arXiv:1705.06676, 2017.
- [10] J. Berant, A. Chou, R. Frostig, and P. Liang, "Semantic parsing on freebase from question-answer pairs," in *Proc. Conf. Empirical Methods Natural Language Processing*, 2013, pp. 1533–1544.
- [11] L. Bertinetto, J. F. Henriques, J. Valmadre, P. H. S. Torr, and A. Vedaldi, "Learning feed-forward one-shot learners," in *Proc. Neural Information Processing Systems (NIPS)*, 2016, pp. 523–531.
- [12] K. Bollacker, C. Evans, P. Paritosh, T. Sturge, and J. Taylor, "Freebase: A collaboratively created graph database for structuring human knowledge," in *Proc. ACM SIGMOD Int. Conf. Management of Data*, 2008, pp. 1247–1250.
- [13] A. Bordes, N. Usunier, S. Chopra, and J. Weston, "Large-scale simple question answering with memory networks," *arXiv Preprint*, arXiv:1506.02075, 2015.
- [14] Q. Cai and A. Yates, "Large-scale semantic parsing via schema matching and lexicon extension," in *Proc. Conf. Association Computational Linguistics*, 2013, pp. 423–433.
- [15] R. Cantrell, M. Scheutz, P. Schermerhorn, and X. Wu, "Robust spoken instruction understanding for hri," in *Proc. 5th ACM/IEEE Int. Conf. Human-Robot Interaction*, 2010, pp. 275–282.
- [16] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2005, vol. 1, pp. 886–893.
- [17] A. Das, H. Agrawal, C. L. Zitnick, D. Parikh, and D. Batra, "Human attention in visual question answering: Do humans and deep networks look at the same regions?" in *Proc. Conf. Empirical Methods Natural Language Processing*, 2016, pp. 932–937.
- [18] A. Das, S. Kottur, K. Gupta, A. Singh, D. Yadav, J. M. Moura, D. Parikh, and D. Batra, "Visual dialog," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2017.
- [19] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2009, pp. 248–255.
- [20] H. Fang, S. Gupta, F. Iandola, R. Srivastava, L. Deng, P. Dollár, J. Gao, X. He, M. Mitchell, and J. Platt, "From captions to visual concepts and back," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2015, pp. 1473–1482.
- [21] A. Fukui, D. H. Park, D. Yang, A. Rohrbach, T. Darrell, and M. Rohrbach, "Multimodal compact bilinear pooling for visual question answering and visual grounding," in *Proc. Conf. Empirical Methods Natural Language Processing (EMNLP)*, 2016, pp. 457–468.
- [22] H. Gao, J. Mao, J. Zhou, Z. Huang, L. Wang, and W. Xu, "Are you talking to a machine? Data set and methods for multilingual image question answering," in *Proc. Advances in Neural Information Processing Systems*, 2015, pp. 2296–2304.
- [23] D. Geman, S. Geman, N. Hallonquist, and L. Younes, "Visual turing test for computer vision systems," *Proc. Natl. Acad. Sci.*, vol. 112, no. 12, pp. 3618–3623, 2015.
- [24] R. Girshick, "Fast R-CNN," in *Proc. IEEE Int. Conf. Computer Vision*, 2015, pp. 1440–1448.
- [25] Y. Goyal, T. Khot, D. Summers-Stay, D. Batra, and D. Parikh, "Making the V in VQA matter: Elevating the role of image understanding in visual question answering," in *Proc. IEEE Conf. Comp. Vis. Patt. Recogn. (CVPR)*, 2017.
- [26] A. Graves, G. Wayne, and I. Danihelka, "Neural turing machines," *arXiv Preprint*, arXiv:1410.5401, 2014.
- [27] L. A. Hendricks, S. Venugopalan, M. Rohrbach, R. J. Mooney, K. Saenko, and T. Darrell, "Deep compositional captioning: Describing novel object categories without paired training data," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2015, pp. 1–10.
- [28] F. Hill, A. Bordes, S. Chopra, and J. Weston, "The goldilocks principle: Reading children's books with explicit memory representations," *arXiv Preprint*, arXiv:1511.02301, 2015.
- [29] R. Hu, J. Andreas, M. Rohrbach, T. Darrell, and K. Saenko, "Learning to reason: End-to-end module networks for visual question answering," *arXiv Preprint*, arXiv:1704.05526, 2017.
- [30] A. Jabri, A. Joulin, and L. van der Maaten, "Revisiting visual question answering baselines," in *Proc. European Conf. Computer Vision (ECCV)* 2016, pp. 727–739.
- [31] J. Johnson, B. Hariharan, L. van der Maaten, L. Fei-Fei, C. L. Zitnick, and R. B. Girshick, "CLEVR: A diagnostic data set for compositional language and elementary visual reasoning," in *Proc. Conf. Computer Vision and Pattern Recognition (CVPR)*, 2017.
- [32] J. Johnson, B. Hariharan, L. van der Maaten, J. Hoffman, F. Li, C. L. Zitnick, and R. B. Girshick, "Inferring and executing programs for visual reasoning," *CoRR*, 2017. [Online]. Available: <http://arxiv.org/abs/1705.03633>
- [33] J.-H. Kim, S.-W. Lee, D.-H. Kwak, M.-O. Heo, J. Kim, J.-W. Ha, and B.-T. Zhang, "Multimodal residual learning for visual QA," in *Proc. Advances Neural Information Processing Systems (NIPS)*, 2016, pp. 361–369.
- [34] J.-H. Kim, K.-W. On, J. Kim, J.-W. Ha, and B.-T. Zhang, "Hadamard product for low-rank bilinear pooling," *arXiv Preprint*, arXiv:1610.04325, 2016.
- [35] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L.-J. Li, D. A. Shamma, M. Bernstein, and L. Fei-Fei, "Visual genome: Connecting language and vision using crowdsourced dense image annotations," *arXiv Preprint*, arXiv:1602.07332, 2016.
- [36] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proc. Advances in Neural Information Processing Systems*, 2012, pp. 1106–1114.
- [37] A. Kumar, O. Irsoy, J. Su, J. Bradbury, R. English, B. Pierce, P. Ondruska, I. Gulrajani, and R. Socher, "Ask me anything: Dynamic memory networks for natural language processing," in *Proc. Int. Conf. Machine Learning*, 2016, pp. 1378–1387.
- [38] A. Lazaridou, N. T. Pham, and M. Baroni, "Combining language and vision with a multimodal skip-gram model," in *Proc. Conf. North American Chapter Assoc. Computational Linguistics–Human Language Technologies (HLT-NAACL)*, 2015, pp. 153–163.

- [39] G. Lin, A. Milan, C. Shen, and I. Reid, "RefineNet: Multi-path refinement networks for high-resolution semantic segmentation," in *Proc. Conf. Computer Vision and Pattern Recognition (CVPR)*, July 2017.
- [40] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft COCO: Common objects in context," in *Proc. European Conf. Computer Vision*, 2014, pp. 740–755.
- [41] D. G. Lowe, "Object recognition from local scale-invariant features," in *Proc. IEEE Int. Conf. Computer Vision*, 1999, vol. 2, pp. 1150–1157.
- [42] J. Lu, X. Lin, D. Batra, and D. Parikh. (2015). Deeper lstm and normalized CNN visual question answering model [Online]. Available: https://github.com/VT-vision-lab/VQA_LSTM_CNN
- [43] J. Lu, J. Yang, D. Batra, and D. Parikh, "Hierarchical question-image co-attention for visual question answering," in *Proc. Advances Neural Information Processing Systems*, 2016, pp. 289–297.
- [44] L. Ma, Z. Lu, and H. Li, "Learning to answer questions from image using convolutional neural network" in *Proc. 30th AAAI Conference on Artificial Intelligence*, 2016, pp. 3567–3573.
- [45] M. Malinowski and M. Fritz, "A multi-world approach to question answering about real-world scenes based on uncertain input," in *Proc. Advances Neural Information Processing Systems*, 2014, pp. 1682–1690.
- [46] M. Malinowski, M. Rohrbach, and M. Fritz, "Ask your neurons: A neural-based approach to answering questions about images," in *Proc. IEEE Int. Conf. Computer Vision*, 2015, pp. 1–9.
- [47] C. Matuszek, N. FitzGerald, L. Zettlemoyer, L. Bo, and D. Fox, "A joint model of language and perception for grounded attribute learning," in *Proc. Int. Conf. Machine Learning*, 2012, pp. 1671–1678.
- [48] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *arXiv Preprint*, arXiv:1301.3781, 2013.
- [49] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Proc. Advances in Neural Information Processing Systems*, 2013, pp. 3111–3119.
- [50] K. W. Murray and J. Krishnamurthy, "Probabilistic neural programs," *arXiv Preprint*, arXiv:1612.00712, 2016.
- [51] H. Noh, P. H. Seo, and B. Han, "Image question answering using convolutional neural network with dynamic parameter prediction," in *Proc. IEEE Conf. Computer Vision Pattern Recognition*, 2016, pp. 30–38.
- [52] B. Peng, Z. Lu, H. Li, and K. Wong, "Toward neural network-based reasoning," *arXiv Preprint*, arXiv:1508.05508, 2015.
- [53] J. Pennington, R. Socher, and C. Manning, "Glove: global vectors for word representation," in *Proc. Conf. Empirical Methods Natural Language Processing*, 2014, pp. 1532–1543.
- [54] S. K. Ramakrishnan, A. Pal, G. Sharma, and A. Mittal, "An empirical evaluation of visual question answering for novel objects," *arXiv Preprint*, arXiv:1704.02516, 2017.
- [55] A. S. Razavian, H. Azizpour, J. Sullivan, and S. Carlsson, "Cnn features off-the-shelf: An astounding baseline for recognition," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition Workshops*, 2014, pp. 806–813.
- [56] S. E. Reed and N. de Freitas, "Neural programmer-interpreters," in *Proc. Int. Conf. Learning Representations*, 2016.
- [57] M. Ren, R. Kiros, and R. Zemel, "Image question answering: a visual semantic embedding model and a new data set," in *Proc. Advances Neural Information Processing Systems*, 2015.
- [58] R. A. Rensink, "The dynamic representation of scenes," *Visual Cognition*, vol. 7, no. 1–3, pp. 17–42, 2000.
- [59] D. Roy, K.-Y. Hsiao, and N. Mavridis, "Conversational robots: building blocks for grounding word meaning," in *Proc. HLT-NAACL Workshop on Learning Word Meaning Non-Linguistic Data*, 2003, pp. 70–77.
- [60] K. Saito, A. Shin, Y. Ushiku, and T. Harada, "Dualnet: Domain-invariant network for visual question answering," *arXiv Preprint*, arXiv:1606.06108, 2016.
- [61] A. Santoro, S. Bartunov, M. Botvinick, D. Wierstra, and T. P. Lillicrap, "Meta-learning with memory-augmented neural networks," in *Proc. Int. Conf. Machine Learning*, 2016, vol. 48, pp. 1842–1850.
- [62] A. Santoro, D. Raposo, D. G. Barrett, M. Malinowski, R. Pascanu, P. Battaglia, and T. Lillicrap, "A simple neural network module for relational reasoning," *arXiv Preprint*, arXiv:1706.01427, 2017.
- [63] M. J. Seo, A. Kembhavi, A. Farhadi, and H. Hajishirzi, "Bidirectional attention flow for machine comprehension," *arXiv Preprint*, arXiv:1611.01603, 2016.
- [64] P. Sermanet, A. Frome, and E. Real, "Attention for fine-grained categorization," *arXiv Preprint*, arXiv:1412.7054, 2014.
- [65] K. J. Shih, S. Singh, and D. Hoiem, "Where to look: Focus regions for visual question answering," in *Proc. IEEE Conf. Computer Vision Pattern Recognition*, 2016, pp. 4613–4621.
- [66] S. Sukhbaatar, A. Szlam, J. Weston, and R. Fergus, "Weakly supervised memory networks," *arXiv Preprint*, arXiv:1503.08895, 2015.
- [67] M. Tapaswi, Y. Zhu, R. Stiefelwagen, A. Torralba, R. Urteasun, and S. Fidler, "Movieqa: Understanding stories in movies through question-answering," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2016, pp. 4631–4640.
- [68] D. Teney, P. Anderson, X. He, and A. van den Hengel, "Tips and tricks for visual question answering: Learnings from the 2017 challenge," *arXiv Preprint*, arXiv:1708.02711, 2017.
- [69] D. Teney, L. Liu, and A. van den Hengel, "Graph-structured representations for visual question answering," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2017.
- [70] D. Teney and A. van den Hengel, "Zero-shot visual question answering," *arXiv Preprint*, arXiv: 1611.05546, 2016.
- [71] I. Vendrov, R. Kiros, S. Fidler, and R. Urteasun, "Order-embeddings of images and language," in *Proc. Int. Conf. Learning Representations*, 2016.
- [72] O. Vinyals, C. Blundell, T. Lillicrap, K. Kavukcuoglu, and D. Wierstra, "Matching networks for one shot learning," in *Proc. Neural Information Processing System (NIPS)*, 2016, pp. 3630–3638.
- [73] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: A neural image caption generator," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2014, pp. 3156–3164.
- [74] P. Wang, Q. Wu, C. Shen, and A. v d. Hengel, "The VQA-machine: Learning how to use existing vision algorithms to answer new questions," *arXiv Preprint*, arXiv:1612.05386, 2016.
- [75] P. Wang, Q. Wu, C. Shen, A. v d. Hengel, and A. Dick, "Explicit knowledge-based reasoning for visual question answering," *arXiv Preprint*:1511.02570, 2015.
- [76] P. Wang, Q. Wu, C. Shen, A. v d. Hengel, and A. Dick, "FVQA: Fact-based visual question answering," *arXiv Preprint*, arXiv:1606.05433, 2016.
- [77] J. Weston, S. Chopra, and A. Bordes, "Memory networks," *arXiv Preprint*, arXiv:11410.3916, 2015.
- [78] T. Winograd, "Understanding natural language," *Cognit. Psychol.*, vol. 3, no. 1, pp. 1–191, 1972.
- [79] Q. Wu, C. Shen, A. v. d. Hengel, L. Liu, and A. Dick, "What value do explicit high level concepts have in vision to language problems?" in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2016, pp. 203–212.
- [80] Q. Wu, C. Shen, A. v d. Hengel, P. Wang, and A. Dick, "Image captioning and visual question answering based on attributes and their related external knowledge," *arXiv Preprint*, arXiv:1603.02814, 2016.
- [81] Q. Wu, D. Teney, P. Wang, C. Shen, A. Dick, and A. van den Hengel, "Visual question answering: a survey of methods and data sets," *Computer Vision and Image Understanding*, to be published.
- [82] Q. Wu, P. Wang, C. Shen, A. Dick, and A. v. d. Hengel, "Ask me anything: Free-form visual question answering based on knowledge from external sources," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2016, pp. 4622–4630.
- [83] C. Xiong, S. Merity, and R. Socher, "Dynamic memory networks for visual and textual question answering," in *Proc. Int. Conf. Machine Learning*, 2016, pp. 2397–2406.
- [84] C. Xiong, V. Zhong, and R. Socher, "Dynamic coattention networks for question answering," *arXiv Preprint*, arXiv:1611.01604, 2016.
- [85] H. Xu and K. Saenko, "Ask, attend and answer: exploring question-guided spatial attention for visual question answering," *arXiv Preprint*, arXiv:1511.05234, 2015.
- [86] K. Xu, J. Ba, R. Kiros, A. Courville, R. Salakhutdinov, R. Zemel, and Y. Bengio, "Show, attend and tell: neural image caption generation with visual attention," in *Proc. Int. Conf. Machine Learning*, 2015, pp. 2048–2057.
- [87] Z. Yang, X. He, J. Gao, L. Deng, and A. Smola, "Stacked attention networks for image question answering," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2016, pp. 21–29.
- [88] X. Yao and B. Van Durme, "Information extraction over structured data: Question answering with freebase," in *Proc. Conf. Association Computational Linguistics*, 2014, pp. 956–966.
- [89] K.-H. Zeng, T.-H. Chen, C.-Y. Chuang, Y.-H. Liao, J. C. Nibbles, and M. Sun, "Leveraging video descriptions to learn video question answering," in *Proc. Conf. Artificial Intelligence AAAI*, 2017, pp. 4334–4340.
- [90] P. Zhang, Y. Goyal, D. Summers-Stay, D. Batra, and D. Parikh, "Yin and yang: Balancing and answering binary visual questions," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2016, pp. 5014–5022.
- [91] L. Zhu, Z. Xu, Y. Yang, and A. G. Hauptmann, "Uncovering temporal context for video question and answering," *arXiv Preprint*, arXiv:1511.04670, 2015.
- [92] Y. Zhu, O. Groth, M. Bernstein, and L. Fei-Fei, "Visual7W: Grounded question answering in images," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2016, pp. 4995–5004.

Deep Metric Learning for Visual Understanding

An overview of recent advances



©ISTOCKPHOTO.COM/ZAPP2PHOTO

Metric learning aims to learn a distance function to measure the similarity of samples, which plays an important role in many visual understanding applications. Generally, the optimal similarity functions for different visual understanding tasks are task specific because the distributions for data used in different tasks are usually different. It is generally believed that learning a metric from training data can obtain more encouraging performances than handcrafted metrics [1]–[3], e.g., the Euclidean and cosine distances. A variety of metric learning methods have been proposed in the literature [2]–[5], and many of them have been successfully employed in visual understanding tasks such as face recognition [6], [7], image classification [2], [3], visual search [8], [9], visual tracking [10], [11], person reidentification [12], cross-modal matching [13], image set classification [14], and image-based geolocation [15]–[17].

Metric learning techniques are usually classified into two categories: unsupervised [4] and supervised [4]. Unsupervised metric learning attempts to learn a low-dimensional subspace to preserve the useful geometrical information of the samples. Supervised metric learning, which is the mainstream metric learning technique and the focus in this article, seeks an appropriate metric by formulating an optimization objective function to exploit supervised information of the training samples, where the objective functions are designed for different specific tasks. However, most conventional metric learning methods usually learn a linear mapping to project samples into a new feature space, which suffer from the nonlinear relationship of data points in metric learning. While the kernel trick can be adopted to address this nonlinearity problem, this type of method suffers from the scalability problem because the kernel trick has two major issues: 1) choosing a kernel is typically difficult and quite empirical and 2) the expression power of kernel functions is often not flexible enough to capture the nonlinearity in the data. Motivated by the fact that deep learning is an effective solution to model the nonlinearity of samples, several deep metric learning (DML) methods [6]–[10], [12], [14], [18]–[34] have been proposed in recent years. The key idea

Digital Object Identifier 10.1109/MSP.2017.2732900
Date of publication: 13 November 2017

of DML is to explicitly learn a set of hierarchical nonlinear transformations to map data points into other feature space for comparing or matching by exploiting the architecture of neural networks in deep learning, which unifies feature learning and metric learning into a joint learning framework. The goal of this article is to provide an overview of recent advances in DML techniques and their various applications in different visual understanding tasks.

Mathematical background

To have a deep understanding of the concept of metric learning, we briefly introduce some necessary mathematical background. This section simply introduces the basic definitions of a metric space and how to find a well-defined metric (or pseudo-metric) over the original inputs by finding a mapping into a Euclidean space.

Definition 1

A metric over a set \mathcal{X} is a mapping $d: \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}_+$ and this mapping d satisfies the following properties (axioms) for all $\mathbf{x}, \mathbf{y}, \mathbf{z} \in \mathcal{X}$:

- 1) $d(\mathbf{x}, \mathbf{y}) \geq 0$
- 2) $d(\mathbf{x}, \mathbf{y}) = d(\mathbf{y}, \mathbf{x})$
- 3) $d(\mathbf{x}, \mathbf{z}) \leq d(\mathbf{x}, \mathbf{y}) + d(\mathbf{y}, \mathbf{z})$
- 4) $d(\mathbf{x}, \mathbf{x}) = 0$
- 5) $d(\mathbf{x}, \mathbf{y}) = 0 \iff \mathbf{x} = \mathbf{y}$.

In Definition 1, axiom 1) is called the *nonnegativity axiom*, axiom 2) is known as the *symmetry axiom*, axiom 3) is called the *triangle inequality axiom*, axiom 4) is referred to as the *identity axiom*, and axiom 5) is known as the *identity of indiscernibles axiom*. A pair (\mathcal{X}, d) , in which \mathcal{X} is a set and d is a metric, is called a *metric space*.

Definition 2

A pseudo-metric over a set \mathcal{X} is a mapping $d: \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}_+$ satisfying the following properties (axioms) for all $\mathbf{x}, \mathbf{y}, \mathbf{z} \in \mathcal{X}$:

- 1) $d(\mathbf{x}, \mathbf{y}) \geq 0$
- 2) $d(\mathbf{x}, \mathbf{y}) = d(\mathbf{y}, \mathbf{x})$
- 3) $d(\mathbf{x}, \mathbf{z}) \leq d(\mathbf{x}, \mathbf{y}) + d(\mathbf{y}, \mathbf{z})$
- 4) $d(\mathbf{x}, \mathbf{x}) = 0$.

A pair (\mathcal{X}, d) , in which \mathcal{X} is a set and d is a pseudo-metric, is called a *pseudo-metric space*. We find that the pseudo-metric doesn't need to satisfy the identity of indiscernibles axiom of the metric. In metric learning, we may consider the pseudo-metrics sometimes instead of metrics and refer to them as *metrics*.

The Euclidean distance is a widely used metric, which is usually adopted to measure the dissimilarity of data points. Give two data points \mathbf{x} and \mathbf{y} , the Euclidean distance between \mathbf{x} and \mathbf{y} is defined as

$$d(\mathbf{x}, \mathbf{y}) = \|\mathbf{x} - \mathbf{y}\|_2 = \sqrt{(\mathbf{x} - \mathbf{y})^T (\mathbf{x} - \mathbf{y})}, \quad (1)$$

in which a large distance means the dissimilarity of \mathbf{x} and \mathbf{y} , and a small distance denotes the similarity of \mathbf{x} and \mathbf{y} .

The main objective of metric learning is to learn a metric over the input data points. One widely used method to learn a metric

is to first map the input data points of the original space into a Euclidean metric space and then compute the Euclidean distance after the mapping. The following lemma declares this method.

Lemma 1

Let $\mathcal{X} = \{\mathbf{x}, \mathbf{y}, \mathbf{z}, \dots\}$ be a set, $f: \mathcal{X} \rightarrow \mathbb{R}^n$ be any well-defined mapping, and $d: \mathbb{R}^n \times \mathbb{R}^n \rightarrow \mathbb{R}_+$ be the Euclidean metric over \mathbb{R}^n , then $d_f: \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}_+$ defined by $d_f(\mathbf{x}, \mathbf{y}) = d(f(\mathbf{x}), f(\mathbf{y})) = \|f(\mathbf{x}) - f(\mathbf{y})\|_2$ is a well-defined pseudo-metric over \mathcal{X} .

As Definition 2 (pseudo-metric) keeps for all data points $f(\mathbf{x}), f(\mathbf{y}), f(\mathbf{z})$ and it is independent of the selection of mapping f , Lemma 1 is verified.

With Lemma 1, metric learning is the procedure of learning the mapping function f . In addition, from the perspective of feature representation, the goal of metric learning can be obviously interpreted as finding a new feature representation $\mathbf{h} = f(\mathbf{x})$ of the data point \mathbf{x} to better suit the Euclidean space. Thus, the objective of metric learning is to find mapping f under various loss functions and constraints.

An illustration

To simply illustrate how metric learning works, we conducted an experiment on the MNIST data set [36]. We sampled 150 samples from three classes of handwritten digits: four, seven, and nine, where each class contains 50 samples. Each digit sample is a 28×28 grayscale image, and we lexicographically converted it into a 784-dimensional feature vector. We employed the linear discriminant analysis (LDA) as a metric learning method to project data points from the original space to the transformed space. Figure 1 shows an example of how metric learning works on this real-world data set. As seen, samples from different classes are mixed in the original space, and they are well separated in the transformed space.

In this article, we focus on DML, which explicitly learns a nonlinear mapping f to map data points into a new feature space by exploiting the architecture of deep neural networks, in which the nonlinear mapping f is parameterized by the weights and biases of deep neural network.

DML

In this section, we introduce the basic concepts of DML, and discuss the similarities and differences among the existing DML methods.

Basic concepts

From Lemma 1, DML is to explicitly learn a nonlinear mapping f to map data points into a new feature space by exploiting the architecture of deep neural networks, in which the nonlinear mapping f is parameterized by the weights and biases of deep neural network.

Given a simple neural network architecture as shown in Figure 2, for an input $\mathbf{x} \in \mathbb{R}^{r^{(0)}}$, its output of the first layer is $\mathbf{h}^{(1)} = \varphi(\mathbf{W}^{(1)} \mathbf{x} + \mathbf{b}^{(1)}) \in \mathbb{R}^{r^{(1)}}$, and its output of the m th layer is $\mathbf{h}^{(m)} = \varphi(\mathbf{W}^{(m)} \mathbf{h}^{(m-1)} + \mathbf{b}^{(m)}) \in \mathbb{R}^{r^{(m)}}$, $1 \leq m \leq M$, $\mathbf{h}^{(0)} = \mathbf{x}$, where matrix $\mathbf{W}^{(m)} \in \mathbb{R}^{r^{(m)} \times r^{(m-1)}}$ and vector $\mathbf{b}^{(m)} \in \mathbb{R}^{r^{(m)}}$ are weights and biases of this neural network, M is the total number

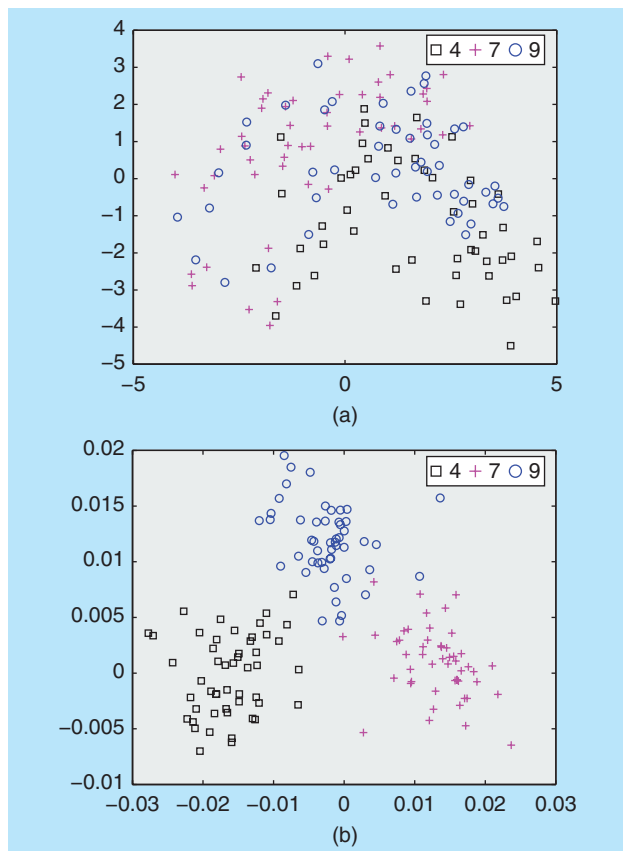


FIGURE 1. An example on the MNIST data set to illustrate how metric learning works. For ease of visualization, these samples are embedded into the two-dimensional feature spaces (a) and (b) by principal component analysis and LDA, respectively.

of layers, $r^{(m)}$ is the number of neural units in the m th layer, $\varphi: \mathbb{R} \rightarrow \mathbb{R}$ is a nonlinear activation function (e.g., sigmoid and tanh). In this way, the output of this neural network at the most top layer can be represented as:

$$\begin{aligned} f(\mathbf{x}) &= \mathbf{h}^{(M)} = \varphi(\mathbf{W}^{(M)}\mathbf{h}^{(M-1)} + \mathbf{b}^{(M)}) \in \mathbb{R}^{r^{(M)}} \\ \mathbf{h}^{(1)} &= \varphi(\mathbf{W}^{(1)}\mathbf{x} + \mathbf{b}^{(1)}), \end{aligned} \quad (2)$$

where the mapping $f: \mathbb{R}^{r^{(0)}} \rightarrow \mathbb{R}^{r^{(M)}}$ is a parametric nonlinear function which is determined by a set of parameters $\{\mathbf{W}^{(m)}, \mathbf{b}^{(m)}\}_{m=1}^M$.

Let f be the mapping function of a neural network. For an input \mathbf{x} , $f(\mathbf{x})$ is its output through this neural network. According to Lemma 1, the distance of data points \mathbf{x}_i and \mathbf{x}_j in the deep metric space is to calculate the Euclidean distance between $f(\mathbf{x}_i)$ and $f(\mathbf{x}_j)$ as:

$$d_f(\mathbf{x}_i, \mathbf{x}_j) = d(f(\mathbf{x}_i), f(\mathbf{x}_j)) = \|f(\mathbf{x}_i) - f(\mathbf{x}_j)\|_2. \quad (3)$$

The goal of DML is to learn the mapping f under certain constraints, where f is parameterized by the weights and biases of the neural network.

Figure 3 shows another widely used architecture of neural network, called a *convolutional neural network (CNN)*, which

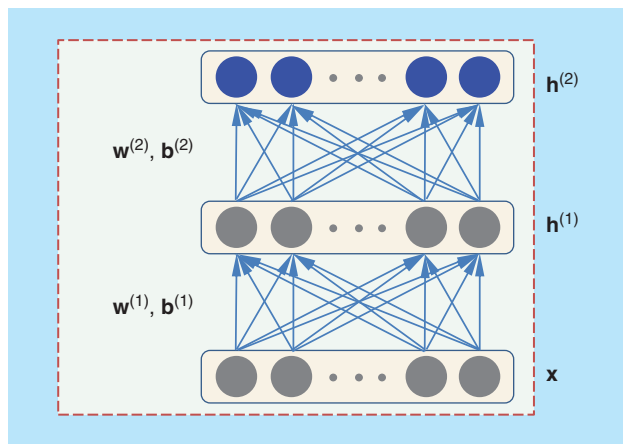


FIGURE 2. A simple illustration of a feed-forward neural network architecture used in many DML methods [23]. The input to the network is \mathbf{x} , and the output of the hidden layer and the top layer is $\mathbf{h}^{(1)}$ and $\mathbf{h}^{(2)}$, respectively, in which $\mathbf{W}^{(m)}$ and $\mathbf{b}^{(m)}$ are weights and biases of this neural network, $1 \leq m \leq 2$.

has been employed by many DML algorithms recently. Generally, CNNs comprise several convolutional layers, subsampling layers, and fully connected layers. Specifically, the feed-forward neural network in Figure 2 is the fully connected part of CNN architecture in Figure 3.

DML via Siamese networks

Typically, there are two main types of neural networks used in DML methods: Siamese networks and triplet networks. Figure 4 shows the diagrams of Siamese networks and triplet networks for DML. For a pair of data points $(\mathbf{x}_i, \mathbf{x}_j)$, we say they are a *similar pair* (or *positive pair*) if \mathbf{x}_i and \mathbf{x}_j are semantically similar, and they are called a *dissimilar pair* (or *negative pair*) if they are semantically dissimilar. Let $\mathcal{S} = \{(i, j)\}$ be an index set consisting of similar pairs, and $\mathcal{D} = \{(i, j)\}$ be an index set consisting of dissimilar pairs, respectively. The Siamese networks-based DML framework is trained by minimizing a contrastive loss function:

$$\begin{aligned} L(\{\mathbf{W}^{(m)}, \mathbf{b}^{(m)}\}_{m=1}^M) &= \sum_{(i,j) \in \mathcal{S}} h(d_f(\mathbf{x}_i, \mathbf{x}_j) - \tau_1) \\ &+ \sum_{(i,j) \in \mathcal{D}} h(\tau_2 - d_f(\mathbf{x}_i, \mathbf{x}_j)), \end{aligned} \quad (4)$$

where $h(x) = \max(0, x)$ is the hinge loss function, and τ_1 and τ_2 are two positive thresholds, $\tau_1 < \tau_2$. By minimizing this contrastive loss function, we expect the distance $d_f(\mathbf{x}_i, \mathbf{x}_j)$ for a positive pair to be less than a smaller parameter τ_1 and that of a negative pair to be larger than a larger parameter τ_2 . Figure 5 shows the key idea of such DML methods.

Dimensionality reduction by learning an invariant mapping (DrLIM) [18], [19] is an important work on DML via Siamese networks for face verification. DrLIM exploited discriminative information from neighborhood relationships of samples to learn the mapping function. There are four characteristics in their method: 1) it only needs neighborhood relationships between training samples; 2) it learns distance functions that are robust to nonlinear transformations of the input signals; 3) the learned function can handle the unseen classes problem so that the new

coming testing samples can also be used with the learned metric; and 4) the mappings generated by the function is smooth and coherent in the output space.

Cai et al. [20] introduced a deep nonlinear metric learning (DNLML) method by using a deep independent subspace analysis (ISA) network, called DNLML-ISA for face verification. ISA is an unsupervised learning algorithm and a two-layer neural network, where different active functions in the first and second layers were used, respectively. Specifically, DNLML-ISA employed the ISA network to transform features from the original space to another feature subspace. To identify discriminative features, DNLML-ISA combined the side information constraints for metric learning with ISA, and stacked the ISA networks into a deep architecture. Since DNLML-ISA is trained layer by layer, it cannot use the backpropagation algorithm to update the model and also cannot fully exploit the discriminative information.

Hu et al. [6] introduced a discriminative DML (DDML) method for face verification. Unlike the stacked model used in DNLML-ISA, DDML employed a fully connected deep neural network to learn multiple nonlinear transformations to map face samples into a discriminative distance space, where the similarity of each positive pair is enlarged and that of each negative pair is reduced, respectively. The denoising auto-encoder was used as the initialization of the parameters of each layer and then the backpropagation was used to update the model. The key advantage of DDML is that it can be trained on a small size of training data set and without using the extensive outside labeled data.

Taigman et al. [21] introduced a DeepFace method by employing an end-to-end metric learning method with the Siamese network for face recognition. Unlike DDML, where only the metrics were learned at the fully connected layers, DeepFace performed discriminative learning with the convolutional, pooling, and fully connected layers so that more labeled training samples were used to train the model. Finally, the parameters of the Siamese network were trained by the standard cross-entropy loss and backpropagation method.

Sun et al. [7] used carefully designed deep convolutional networks (deep ConvNets) by making use of both the verification and identification information to learn the deep identification-verification features (DeepID2) [7] for face verification. Specifically, their method extracted deep features with two signals: the first is the identification signal, which was achieved by following the DeepID2 layer with an n -way softmax layer. The network was trained by minimizing the cross-entropy identification loss. The

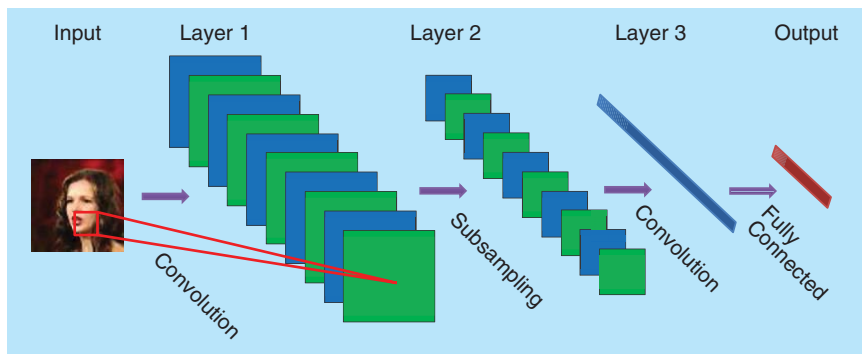


FIGURE 3. An illustration of a CNN architecture. This CNN comprises two convolutional layers C_1 and C_3 , a subsampling layer S_2 , and a fully connected layer F_3 .

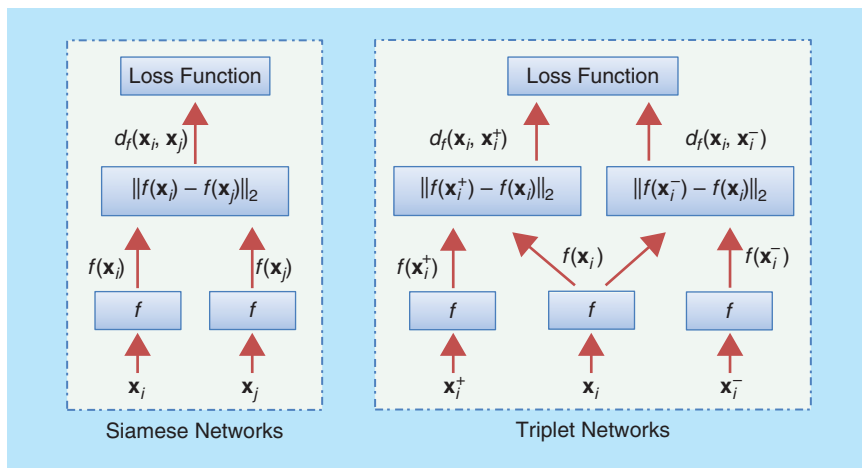


FIGURE 4. Diagrams of Siamese networks and triplet networks for DML. Siamese networks are composed of two same neural networks f with shared parameters, where $(\mathbf{x}_i, \mathbf{x}_j)$ is a similar/dissimilar pair. Triplet networks consist of three same neural networks f with shared parameters, where $(\mathbf{x}_i, \mathbf{x}_i^+, \mathbf{x}_i^-)$ is a triplet, \mathbf{x}_i is a reference, \mathbf{x}_i^+ and \mathbf{x}_i^- are similar and dissimilar examples to \mathbf{x}_i .

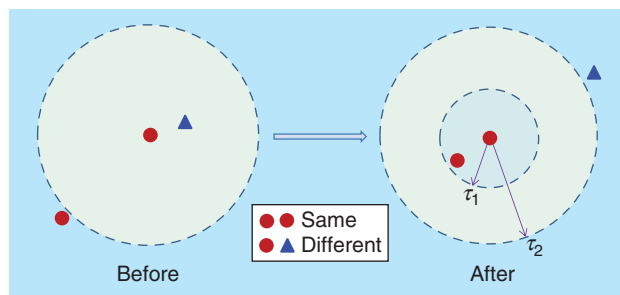


FIGURE 5. The basic idea of DML methods via Siamese network using (4) [6]. At the top layer of the network, the distance $d_f(\mathbf{x}_i, \mathbf{x}_j)$ for a positive pair is less than a smaller parameter τ_1 , and that of a negative pair is larger than a larger parameter τ_2 , respectively.

other one is the verification signal, which enforced that DeepID2 features extracted from the same class are as similar as possible. Their method showed that both the identification and verification signal contributed to the final discriminative feature learning.

Yi et al. [12] proposed a DML method with a Siamese deep neural network to learn a similarity metric from image pixels directly for person reidentification. Their method jointly learned discriminative features and similarity measures under a unified

deep framework. The network has a symmetrical structure, where two subnetworks were connected by a cosine similarity layer. There are two convolutional layers and a full connected layer for each subnetwork. Their method has two key advantages: 1) it can learn a similarity metric from image pixels directly; 2) it can learn multichannel filters to capture both the color and texture information from body images simultaneously.

Most DML methods assume that the training and testing samples are collected in similar scenarios and the same distribution assumption is usually made. This assumption does not hold in many real world applications, especially when samples are captured across different data sets. To address this, Hu et al. [23] proposed a deep transfer metric learning (DTML) method to learn hierarchical nonlinear transformations for cross-domain visual recognition, which learned transferrable discriminative knowledge from the labeled source domain to the unlabeled target domain. Specifically, DTML learned a deep metric network by maximizing the interclass variations and minimizing the intraclass variations, and minimizing the distribution divergence between the source domain and the target domain at the top layer of the network. To better exploit the discriminative information from the source domain, they also considered exploiting discriminative information from the middle layers of the deep network so that more discriminative information can be exploited.

Recently, Lu et al. [14] introduced a multimaniifold DML (MMDML) method to recognize objects from different viewpoints or under different illuminations. Specifically, MMDML jointly learns multiple nonlinear feed-forward neural networks, one for each object class, to explicitly project the instances from each image set into a common feature space at the top layer of all networks, where the maximal manifold margin constraint is enforced. In this way, class-specific discriminative information can be effectively exploited for classification. The authors' method achieved competitive performance on five widely used image set data sets.

Table 1 shows basic characteristics of several Siamese networks-based DML methods. In this table, the strongly supervised setting means that the class labels of training data are used to train neural networks, and the weakly supervised setting denotes that only the pairwise labels of similar pairs and dissimilar pairs are used to train neural networks.

Table 1. Characteristics of several DML methods using Siamese networks.

Method	Setting	End to End?	Convolutional Architecture?
DrLIM [19]	Strongly supervised	Yes	Yes
DNLM-ISA [20]	Weakly supervised	No	No
DDML [6]	Weakly supervised	No	No
DeepFace [21]	Strongly supervised	Yes	Yes
DeepID2 [7]	Strongly supervised	Yes	Yes
DTML [23]	Strongly supervised	No	No
MMDML [14]	Strongly supervised	No	No

DML via triplet networks

DML using triplet networks was trained by minimizing a triplet loss function, which exploits labels of training data to generate triplets. Given a triplet $(\mathbf{x}_i, \mathbf{x}_i^+, \mathbf{x}_i^-)$, \mathbf{x}_i^+ is a similar example to the reference \mathbf{x}_i , and \mathbf{x}_i^- is a dissimilar example to the \mathbf{x}_i . A triplet $(\mathbf{x}_i, \mathbf{x}_i^+, \mathbf{x}_i^-)$ means that \mathbf{x}_i is more similar to \mathbf{x}_i^+ in contrast to \mathbf{x}_i^- , i.e., $d_f(\mathbf{x}_i, \mathbf{x}_i^+) < d_f(\mathbf{x}_i, \mathbf{x}_i^-)$. DML via triplet networks aims to minimize the following loss function for triplets:

$$L(\{\mathbf{W}^{(m)}, \mathbf{b}^{(m)}\}_{m=1}^M) = \sum_i h(\tau + d_f(\mathbf{x}_i, \mathbf{x}_i^+) - d_f(\mathbf{x}_i, \mathbf{x}_i^-)), \quad (5)$$

where $h(x) = \max(0, x)$ is the hinge loss function, and $\tau > 0$ is a margin between $d_f(\mathbf{x}_i, \mathbf{x}_i^+)$ and $d_f(\mathbf{x}_i, \mathbf{x}_i^-)$. The triplet network pulls the similar example close to reference and pushes dissimilar example further away.

Wang et al. [9] proposed a deep ranking model with the triplet-based hinge loss functions to learn similarity metric from raw images. Specifically, they employed a multiscale neural network architecture to capture both the global visual properties and the image semantics. An efficient online triplet sampling method was presented to generate a large amount of training data to learn the parameters of the network.

Hoffer et al. [26] employed a triplet network architecture for DML, which aims to learn useful representations by distance comparisons. Their method is similar to the approach in [9] that learned a deep ranking similarity function for image retrieval. Their method made a comprehensive study of the triplet architecture, and demonstrated that the triplet approach is a strong competitor to the Siamese approach.

Schroff et al. [24] introduced a FaceNet deep model that directly learns a mapping from the original sample space to a compact Euclidean space. Once this space is produced, face recognition and clustering can be easily implemented under the network. Specifically, FaceNet used a deep convolutional network to directly optimize the embedding itself rather than using an intermediate bottleneck layer. Triplets of roughly aligned matching/nonmatching face patches were generated for training with an online triplet mining method.

Bell and Bala [27] proposed learning visual similarity for product design with the CNNs, which exploit communities of users to help each other answering questions about products in images. Their method contains two different domains of product images: products cropped from internet scenes, and products in their iconic form. With the help of a multidomain deep embedding, it can deal with several applications of visual search including identifying products in scenes and finding stylistically similar products.

Song et al. [28] introduced a DML method via lifted structured feature embedding (LiftedStruct) to learn semantic feature embeddings where similar examples are mapped close to each other and dissimilar examples are mapped farther apart. Their method took full advantage of the training batches in the network training stage by lifting the vector of pairwise distances within the batch to the matrix of pairwise distances. This step enabled the method to learn the state of the art feature embedding by optimizing a new structured prediction

objective on the lifted problem. Experiments on three large-scale data sets demonstrated significant improvements over existing deep feature embedding methods.

Cui et al. [29] presented an iterative framework for fine-grained visual categorization with humans in the loop information. Their method can handle three challenges in existing fine-grained visual categorization methods: lacking of training data, large number of fine-grained categories, and high intraclass versus low interclass variance. Using DML with humans in the loop, a low-dimensional feature embedding with anchor points on manifolds was learned for each category, where these anchor points captured intraclass variances and remained discriminative among different classes. In each round, images with high confidence scores from our model were sent to humans for labeling. By comparing these images with exemplar images, labelers marked each candidate image as either a true positive or a false positive. True positives were added into the current data set and false positives were considered as hard negatives for the DML model. Then the model was retrained with an expanded data set and hard negatives for the next round iteration. The proposed DML method was evaluated on two fine-grained data sets. Experimental evaluations showed that their method achieved significant performance gain over state-of-the-art methods.

Shi et al. [31] proposed a deep metric embedding method with triplet loss for person reidentification. Their method introduced a positive sample mining method to train robust CNN for person reidentification. In addition, a metric weight constraint was used to improve the learning, so that the learned metric has a better generalization ability. They empirically found that both of these tricks improve the reidentification performance.

Lim et al. [33] proposed a competitive approach for style similarity learning of three-dimensional (3-D) shapes using DML, which made use of recent advances in triplet based metric learning with neural networks. The key advantages of their method are four aspects:

- it explored DML techniques for perceived style similarities of 3-D shapes
- it showed that rendered images of 3-D geometry from multiple viewpoints were an appropriate representation and how salient views can be selected
- it used a triplet sampling method that does not rely on style class labels and allows for an efficient learning procedure
- it showed how heterogeneous data sources in the form of 3-D geometry and annotated photographs found online can be integrated into the DML method.

DML via other networks

There are also some DML methods via other networks. For example, Batchelor and Green [22] proposed using DML on CNNs to learn features with good locality for object recognition. In particular, they considered two metric learning methods: neighborhood components analysis and mean square error's gradient minimization (MEGM). They utilized a nonlinear form of MEGM as an alternative to neighborhood components analysis and proposed some stochastic sampling methods to apply them to larger data sets with a minibatch stochastic gradient descent algorithm.

Sohn [32] proposed a DML method using multiclass N -pair loss [32]. Their method first generated triplet loss by allowing joint comparison among more than one negative example. Then, $N - 1$ negative examples were considered to reduce the computational burden of evaluating deep embedding vectors. They demonstrated the superiority of their method over other competing loss functions for a variety of tasks such as fine-grained object recognition and verification, image clustering and retrieval, and face verification and identification.

Visual understanding applications

In this section, we show various visual understanding applications via DML, including face recognition, image classification, visual search, person reidentification, visual tracking, cross-modal matching, and image set classification.

Face recognition

Chopra et al. [18] learned a similarity metric for face verification. Their approach learned a CNN-based mapping from the input space to the target space, where the L_1 norm can directly approximate the semantic distance. Cai et al. [20] learned a nonlinear metric using the deep ISA network. Compared with kernel-based methods, deep models present strong discriminative power and better exploit the nature of the data set. Sun et al. [7] proposed a DeepID2 method to increase the interpersonal variations with the identification signals, and reduce the intrapersonal distances with the verification signals. Taigman et al. [21] presented the DeepFace network by exploiting a 3-D face model and training a nine-layer CNN network. Hu et al. [6] presented a DDML method by learning a set of hierarchical nonlinear transformations, where the distance between positive pairs is smaller than negative pairs by a threshold. They also proposed a DTML [23] for cross-data set face recognition. DTML transferred the information from the labeled source domain to the unlabeled target domain, and minimized their distribution divergence. Schroff et al. [24] proposed a FaceNet method by learning a projection to map facial images to a compact Euclidean space. With the learned embedding, feature vectors can be directly used to measure the similarity of faces. Most these DML methods achieved the state-of-the-art performance on the widely used LFW and YouTube Face data sets.

Image classification

Batchelor and Green [22] utilized CNN architecture to learn a deep nonlinear metric, where the learned features with good locality show good performance and generalization for image classification. Hoffer and Ailon [26] utilized a triplet-based network to learn deep metrics by distance comparisons. The triplet network contains three instances of networks with shared parameters, where three samples with a positive pair and a negative pair can be simultaneously fed into the network. Cui et al. [29] learned a deep metric for fine-grained categorization. Human helps to label high confidence images in each loop to expand data sets and hard negatives, where the network was further retrained in the next loop.

Visual search

Wu et al. [8] proposed an online multimodal deep similarity learning for visual search. They applied deep-learning techniques to

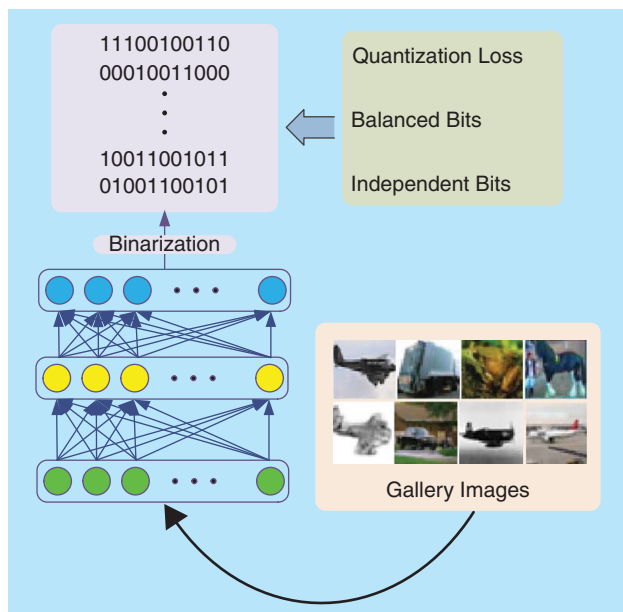


FIGURE 6. The basic idea of deep hashing for large scale visual search [35], which employed a feed-forward neural network to map each gallery image into a compact Hamming feature space with three criteria.

obtain a flexible nonlinear similarity metric from images, which have multimodal feature representations via an efficient and scalable online learning method. Their proposed technique achieved encouraging results on several multimodal images retrieval tasks. Wang et al. [9] proposed a ranking-based deep-learning method for fine-grained image search, which employed a triplet-based hinge loss ranking function and a multiscale neural network. Their method outperformed existing hand-crafted features and deep models in numerous experiments. Liong et al. [35] proposed a deep hashing approach for large scale visual search. Their method

learned compact binary codes to exploit nonlinear relationship of samples, and Figure 6 shows the key idea of their method.

Person reidentification

Yi et al. [12] proposed a DML method for person reidentification, which learned a similarity metric from image pixels directly with a Siamese networks. Hu et al. [23] proposed a DTML method for cross data set person reidentification, where the discriminative information exploited from the source domain was transferred to the target domain with limited labeled samples. Shi et al. [31] proposed a deep embedding metric method for person reidentification, which used a moderate positive sample mining method for robust CNN training, and improved the learning procedure with a metric weight constraint.

Visual tracking

Hu et al. [10] employed DML for visual object tracking. Their DML tracker adopts the marginal fisher analysis criterion to characterize the separability of the positive samples and negative samples by maximizing the variance of interclass negative sample pairs. They first learned a multilayer nonlinear feed-forward neural network to map both the sampled templates and particles into a discriminative feature space to minimize the intraclass variations of positive sample pairs and maximize the interclass variations of negative sample pairs at the top layer of the network. Then, they selected the candidate which is most similar to the template in the learned deep network and considered it as the target in the current predicted frame. Experimental results demonstrated that their DML tracker achieved very competitive performance on a challenging benchmark data set. Figure 7 shows the main procedure of their proposed DML tracker.

Cross-modal matching

Liong et al. [13] proposed a deep coupled metric learning (DCML) method for cross-modal matching. Unlike existing cross-modal

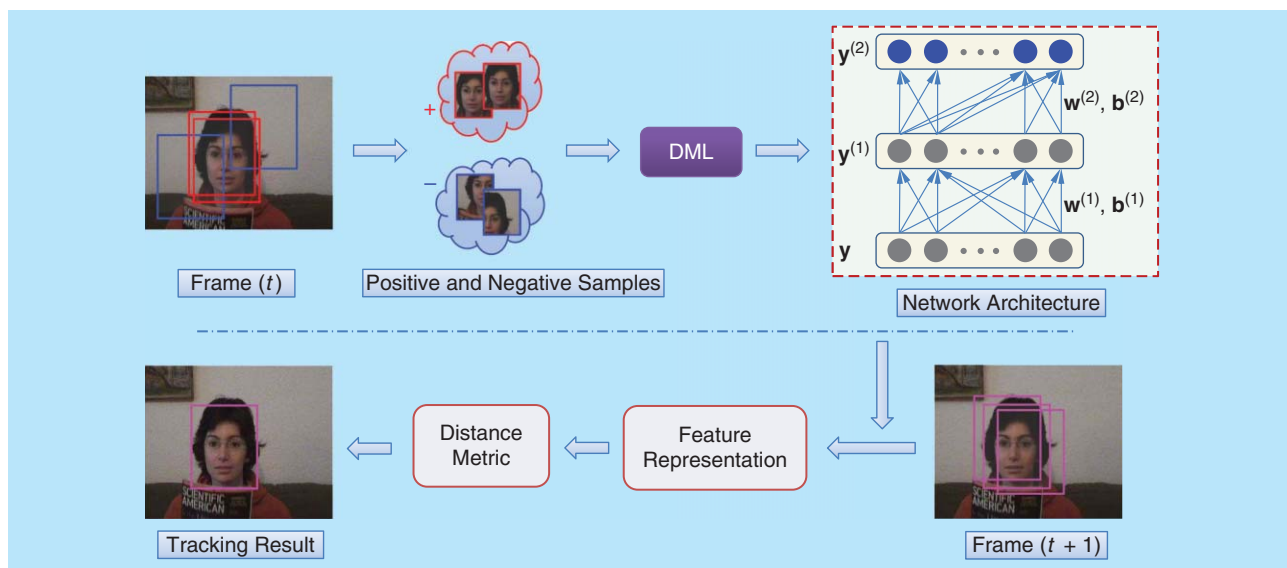


FIGURE 7. The key procedure of the DML tracker [10]. This tracker sampled some positive and negative samples to construct a training set and learn a deep metric with these samples at the t th frame. For the coming $(t + 1)$ th frame, their tracker computed the similarity between each candidate and the template and find the candidate which has the maximal similarity score as the foreground the $(t + 1)$ th frame.

learning methods such as canonical correlation analysis and partial least squares, which learn a single linear latent space to reduce the modality gap, their DCML designs two neural networks to learn two sets of hierarchical nonlinear transformations (one set for each modality) to nonlinearly map data samples into a shared feature subspace, under which the intraclass variation is minimized and the interclass variation is maximized, and the difference of each sample pair captured from two modalities of the same class is minimized, respectively. Experimental results on three different cross-modal matching applications including text-image matching, tag-image retrieval, and heterogeneous face recognition demonstrated the effectiveness of the proposed method. Lin et al. [15], Workman et al. [16], and Vo and Hays [17] employed DML techniques to address the cross-view matching problem for image-based geolocation, in which these methods were used to localize a ground-level query image by matching to a reference database of aerial/overhead images.

Image set classification

Lu et al. [14] presented an MMDML method to recognize objects from different viewpoints or under different illuminations. Specifically, MMDML jointly learns multiple nonlinear feed-forward neural networks, one for each object class, to explicitly project the instances from each image set into a common feature space at the top layer of all networks, where the maximal manifold margin constraint is enforced. In this way, class-specific discriminative information can be effectively exploited for classification. The authors' method achieved competitive performance on five widely used image set data sets. Figure 8 shows the key idea of their MMDML method.

Summary and future research directions

In this article, we have summarized the recent trends of DML and shown their wide applications of various visual understanding tasks including face recognition, image classification, visual search, person reidentification, visual tracking, cross-modal matching, and image set classification. Empirical results have clearly demonstrated that DML can significantly improve the state of the art in these visual understanding tasks.

There are five interesting directions of DML for future research:

- 1) Most existing DML methods learn one neural network from a single feature representation and cannot deal with multiple feature representations directly. In many visual understanding applications, it is easy to extract multiple features for each sample for multiple feature fusion. However, these features extracted from the same sample are usually highly correlated to each other even if they could characterize samples from different aspects. For multiple feature fusion, this highly correlated information should be preserved because it usually reflects the intrinsic information of samples. How to perform DML with multiview feature representation to preserve the correlation of different features and further improve the performance is a desirable future work.
- 2) Most existing DML methods assume that high-quality and clean samples are usually obtained so that the learned metrics are employed for visual understanding. In many real-world

applications, visual data are usually captured in wild conditions so that many noisy and low-quality samples are usually collected, so that it is desirable to develop robust DML methods that can well measure the similarity of these noisy and low-quality samples. Hence, how to develop robust DML methods is another interesting future direction for research.

- 3) Most existing DML methods are developed for a single specific task, which means that a large amount of labeled data for this task are usually required to exploit the supervision information. In some real applications, it is difficult to collect extensive labeled data for a specific task. Therefore, it is desirable to conduct multitask DML which can leverage labeled samples from multiple different yet related tasks so that it is much easier to obtain more labeled samples for DML, which is also an interesting future research direction.
- 4) Most existing DML methods are supervised. In many real applications, it is easier to collect an extensive unlabeled data rather than labeled data for practical applications. Hence, how to develop more effective unsupervised or semisupervised DML is an important future direction.
- 5) Most existing DML methods utilize the contrastive and triplet loss functions to train deep models. To complete the family of DML, employing other loss functions (e.g., quadruplet loss [37]) is also a promising path to the development of DML.

Acknowledgments

We would like to thank the anonymous reviewers for their insightful suggestions for improving the article. This work was supported in part by the National Key Research and Development

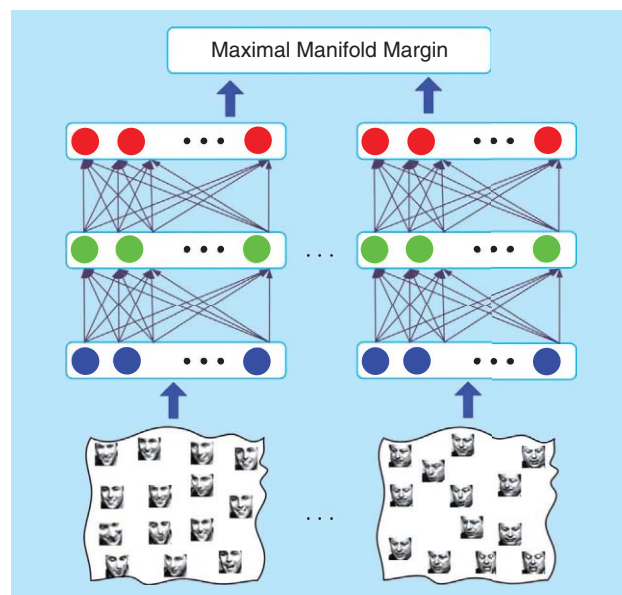


FIGURE 8. The basic idea of MMDML for image set classification [14]. MMDML models each image set as a nonlinear manifold and employs a feed-forward neural network to nonlinearly map it into a feature space. Assume there are C classes, MMDML designs C feed-forward neural networks (one for each manifold). At the top layer of the network, the manifold margin is maximized so that the parameters of these manifolds can be updated with backpropagation. Finally, the testing image set is fed to each network and the smallest distance between it and the training class is used for classification.

Program of China under grant 2016YFB1001001 and the National Natural Science Foundation of China under grant 61672306.

Authors

Jiwen Lu (lujiwen@tsinghua.edu.cn) received his B.Eng. degree in mechanical engineering and his M.Eng. degree in electrical engineering from the Xi'an University of Technology, Xi'an, China, in 2003 and 2006, respectively. He received his Ph.D. degree in electrical engineering from Nanyang Technological University, Singapore, in 2012. He is currently an associate professor with the Department of Automation, Tsinghua University, Beijing, China. His current research interests include computer vision, pattern recognition, and machine learning. He is an associate editor of four international journals including *Pattern Recognition*, and an elected member of the IEEE Technical Committees of the IEEE Circuits and Systems Society and the IEEE Signal Processing Society. He is a Senior Member of the IEEE.

Junlin Hu (jhu007@e.ntu.edu.sg) received his B.Eng. degree from the Xi'an University of Technology, Xi'an, China, in 2008, and the M.Eng. degree from Beijing Normal University, China, in 2012. He is currently pursuing his Ph.D. degree with the School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore. His research interests include computer vision, pattern recognition, and biometrics.

Jie Zhou (jzhou@tsinghua.edu.cn) received his B.S. and M.S. degrees from the Department of Mathematics, Nankai University, Tianjin, China, in 1990 and 1992, respectively, and his Ph.D. degree from the Institute of Pattern Recognition and Artificial Intelligence, Huazhong University of Science and Technology, Wuhan, China, in 1995. From 1995 to 1997, he served as a postdoctoral fellow with the Department of Automation, Tsinghua University, Beijing, China, where he has been a full professor, since 2003. His current research interests include computer vision, pattern recognition, and image processing. He received the National Outstanding Youth Foundation of China Award. He is a Senior Member of the IEEE.

References

- [1] E. P. Xing, A. Y. Ng, M. I. Jordan, and S. J. Russell, "Distance metric learning with application to clustering with side-information," in *Proc. Neural Information Processing Systems*, 2002, pp. 505–512.
- [2] J. V. Davis, B. Kulis, P. Jain, S. Sra, and I. S. Dhillon, "Information-theoretic metric learning," in *Proc. Int. Conf. Machine Learning*, 2007, pp. 209–216.
- [3] K. Q. Weinberger and L. K. Saul, "Distance metric learning for large margin nearest neighbor classification," *J. Mach. Learning Res.*, vol. 10, pp. 207–244, 2009.
- [4] B. Kulis, "Metric learning: A survey," *Found. Trends Mach. Learning*, vol. 5, no. 4, pp. 287–364, 2013.
- [5] M. Harandi, M. Salzmann, and R. Hartley, "Joint dimensionality reduction and metric learning: A geometric take," in *Proc. Int. Conf. Mach. Learning*, 2017, pp. 1404–1413.
- [6] J. Hu, J. Lu, and Y.-P. Tan, "Discriminative deep metric learning for face verification in the wild," in *Proc. IEEE Conf. Computer Vision Pattern Recognition*, 2014, pp. 1875–1882.
- [7] Y. Sun, Y. Chen, X. Wang, and X. Tang, "Deep learning face representation by joint identification-verification," in *Proc. Neural Information Processing Systems*, 2014, pp. 1988–1996.
- [8] P. Wu, S. C. H. Hoi, H. Xia, P. Zhao, D. Wang, and C. Miao, "Online multimodal deep similarity learning with application to image retrieval," in *Proc. ACM Conf. Multimedia*, 2013, pp. 153–162.

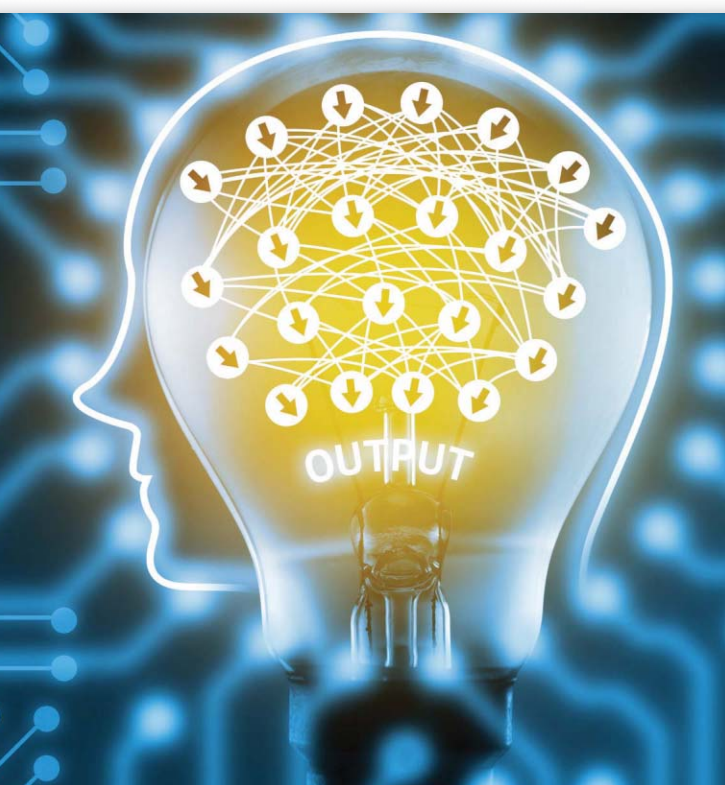
- [9] J. Wang, Y. Song, T. Leung, C. Rosenberg, J. Wang, J. Philbin, B. Chen, and Y. Wu, "Learning fine-grained image similarity with deep ranking," in *Proc. IEEE Conf. Comput. Vision and Pattern Recognition*, 2014, pp. 1386–1393.
- [10] J. Hu, J. Lu, and Y.-P. Tan, "Deep metric learning for visual tracking," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 26, no. 11, pp. 2056–2068, 2016.
- [11] H. Li, Y. Li, and F. Porikli, "Deeptrack: Learning discriminative feature representations online for robust visual tracking," *IEEE Trans. Image Proc.*, vol. 25, no. 4, pp. 1834–1848, 2016.
- [12] D. Yi, Z. Lei, S. Liao, and S. Z. Li, "Deep metric learning for person re-identification," in *Proc. Int. Conf. Pattern Recognition*, 2014, pp. 34–39.
- [13] V. E. Liong, J. Lu, Y.-P. Tan, and J. Zhou, "Deep coupled metric learning for cross-modal matching," *IEEE Trans. Multimedia*, vol. 19, no. 6, pp. 1234–1244, 2017.
- [14] J. Lu, G. Wang, W. Deng, P. Moulin, and J. Zhou, "Multi-manifold deep metric learning for image set classification," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2015, pp. 1137–1145.
- [15] T.-Y. Lin, Y. Cui, S. J. Belongie, and J. Hays, "Learning deep representations for ground-to-aerial geolocalization," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2015, pp. 5007–5015.
- [16] J. Workman, R. Souvenir, and N. Jacobs, "Wide-area image geolocalization with aerial reference imagery," in *Proc. IEEE Int. Conf. Computer Vision*, 2015, pp. 3961–3969.
- [17] N. N. Vo and J. Hays, "Localizing and orienting street views using overhead imagery," in *Proc. European Conf. Computer Vision*, 2016, pp. 494–509.
- [18] S. Chopra, R. Hadsell, and Y. LeCun, "Learning a similarity metric discriminatively, with application to face verification," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2005, pp. 539–546.
- [19] R. Hadsell, S. Chopra, and Y. LeCun, "Dimensionality reduction by learning an invariant mapping," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2006, pp. 1735–1742.
- [20] X. Cai, C. Wang, B. Xiao, X. Chen, and J. Zhou, "Deep nonlinear metric learning with independent subspace analysis for face verification," in *Proc. ACM Conf. Multimedia*, 2012, pp. 749–752.
- [21] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf, "Deepface: Closing the gap to human-level performance in face verification," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2014, pp. 1701–1708.
- [22] O. Batchelor and R. D. Green, "Object recognition by stochastic metric learning," in *Proc. Int. Conf. Simulated Evolution and Learning*, 2014, pp. 798–809.
- [23] J. Hu, J. Lu, and Y.-P. Tan, "Deep transfer metric learning," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2015, pp. 325–333.
- [24] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2015, pp. 815–823.
- [25] X. Han, T. Leung, Y. Jia, R. Sukthankar, and A. C. Berg, "Matchnet: Unifying feature and metric learning for patch-based matching," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2015, pp. 3279–3286.
- [26] E. Hoffer and N. Ailon, "Deep metric learning using triplet network," in *Proc. Similarity-Based Pattern Recognition, Third Int. Workshop*, 2015, pp. 84–92.
- [27] S. Bell and K. Bala, "Learning visual similarity for product design with convolutional neural networks," *ACM Trans. Graphics*, vol. 34, no. 4, pp. 98, 2015.
- [28] H. O. Song, Y. Xiang, S. Jegelka, and S. Savarese, "Deep metric learning via lifted structured feature embedding," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2016, pp. 4004–4012.
- [29] Y. Cui, F. Zhou, Y. Lin, and S. Belongie, "Fine-grained categorization and data set bootstrapping using deep metric learning with humans in the loop," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2016, pp. 1153–1162.
- [30] J. Hu, J. Lu, Y.-P. Tan, and J. Zhou, "Deep transfer metric learning," *IEEE Trans. Image Proc.*, vol. 25, no. 12, pp. 5576–5588, 2016.
- [31] H. Shi, Y. Yang, X. Zhu, S. Liao, Z. Lei, W.-S. Zheng, and S. Z. Li, "Embedding deep metric for person re-identification: A study against large variations," in *Proc. European Conf. Computer Vision*, 2016, pp. 732–748.
- [32] K. Sohn, "Improved deep metric learning with multi-class n-pair loss objective," in *Proc. Neural Information Processing Systems*, 2016, pp. 1849–1857.
- [33] I. Lim, A. Gehre, and L. Kobbelt, "Identifying style of 3-D shapes using deep metric learning," *Comput. Graphics Forum*, vol. 35, no. 5, pp. 207–215, 2016.
- [34] J. Lu, J. Hu, and Y.-P. Tan, "Discriminative deep metric learning for face and kinship verification," *IEEE Trans. Image Processing*, vol. 26, no. 9, pp. 4269–4282, 2017.
- [35] V. E. Liong, J. Lu, G. Wang, P. Moulin, and J. Zhou, "Deep hashing for compact binary codes learning," in *Proc. IEEE Conf. Comput. Vision and Pattern Recognition*, 2015, pp. 2475–2483.
- [36] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proc. IEEE*, vol. 86, no. 11, pp. 2278–2324, Nov 1998.
- [37] M. T. Law, N. Thome, and M. Cord, "Learning a distance metric from relative comparisons between quadruplets of images," *Int. J. Comput. Vision*, vol. 121, no. 1, pp. 65–94, 2017.



Michael T. McCann, Kyong Hwan Jin,
and Michael Unser

Convolutional Neural Networks for Inverse Problems in Imaging

A review



©ISTOCKPHOTO.COM/ZAPP2PHOTO

In this article, we review recent uses of convolutional neural networks (CNNs) to solve inverse problems in imaging. It has recently become feasible to train deep CNNs on large databases of images, and they have shown outstanding performance on object classification and segmentation tasks. Motivated by these successes, researchers have begun to apply CNNs to the resolution of inverse problems such as denoising, deconvolution, super-resolution, and medical image reconstruction, and they have started to report improvements over state-of-the-art methods, including sparsity-based techniques such as compressed sensing. Here, we review the recent experimental work in these areas, with a focus on the critical design decisions:

- From where do the training data come?
- What is the architecture of the CNN?
- How is the learning problem formulated and solved?

We also mention a few key theoretical papers that offer perspectives on why CNNs are appropriate for inverse problems, and we point to some next steps in the field.

Introduction

The basic ideas underlying the use of CNNs (also known as *ConvNets*) for inverse problems are not new. Here, we give a condensed history of CNNs to provide context to what follows. For further historical perspective, see [1]; for an accessible introduction to deep neural networks and a summary of their recent history, see [2]. The CNN architecture was proposed in 1986 [3], and neural networks were developed for solving inverse imaging problems as early as 1988 [4]. These approaches, which used networks with few parameters and did not always include learning, were largely superseded by compressed sensing (or, broadly, convex optimization with regularization) approaches in the 2000s. As computer hardware improved, it became feasible to train larger neural networks, until, in 2012, Krizhevsky et al. [5] achieved a significant improvement over the state of the art on the ImageNet classification challenge by using a graphics processing unit (GPU) to train a CNN with five convolutional layers and 60 million parameters on a set of 1.3 million images. This work spurred a resurgence of interest in neural

Digital Object Identifier 10.1109/MSP.2017.2739299
Date of publication: 13 November 2017

networks and, specifically, CNNs—not only for computer vision tasks but also for inverse problems.

The purpose of this article is to summarize the recent works using CNNs for inverse problems in imaging, i.e., in problems most naturally formulated as recovering an image from a set of noisy measurements. This criterion excludes detection, segmentation, classification, quality assessment, etc. We also focus on CNNs, avoiding other architectures such as recurrent neural networks, fully connected networks, and stacked denoising autoencoders. We organized our literature search by application, selecting topics of broad interest where we could find at least three peer-reviewed papers from the last ten years. (Much of the work on the theory and practice of CNNs is posted on the preprint server arXiv.org before eventually appearing in traditional journals. Because of the lack of peer review on arXiv.org, we have preferred not to cite these papers, except in cases where we are trying to illustrate a very recent trend or future direction for the field.) The resulting applications and references are summarized in Table 1. The aim of this constrained scope is to allow us to draw meaningful generalizations from the surveyed works.

Background

We begin by introducing inverse problems and contrasting the traditional approach to solving them with a learning-based approach. For a textbook treatment of inverse problems, see [28]. Throughout the section, we use X-ray computed tomogra-

phy (CT) as a running example, and Figure 1 shows images of the various mathematical quantities we mention.

Learning for inverse problems in imaging

Mathematically speaking, an imaging system is an operator $H : \mathcal{X} \rightarrow \mathcal{Y}$ that acts on an image $x \in \mathcal{X}$, to create a vector of measurements $y \in \mathcal{Y}$, with $H\{x\} = y$. The underlying function/vector spaces are

- the space, \mathcal{X} , of acceptable images, which can be two-dimensional (2-D), three-dimensional (3-D), or even 3-D+time, with its values representing a physical quantity of interest, such as X-ray attenuation or concentration of fluorophores
- the space, \mathcal{Y} , of measurement vectors that depends on the imaging operator and could include images (discrete arrays of pixels), Fourier samples, line integrals, etc.

We typically consider x to be a continuous object (function of space), while y is usually discrete: $\mathcal{Y} = \mathbb{R}^M$. For example, in X-ray CT, x is an image representing X-ray attenuations, H represents the physics of the X-ray source and detector, and y is the measured sinogram (see Figure 1).

In an inverse imaging problem, we aim to develop a reconstruction algorithm (which is also an operator), $R : \mathcal{Y} \rightarrow \mathcal{X}$, to recover the original image, x , from the measurements, y . The dominant approach for reconstruction, which we call the *objective function approach*, is to model H and recover an estimate of x from y by

$$R_{\text{obj}}\{y\} = \underset{x \in \mathcal{X}}{\text{argmin}} f(H\{x\}, y), \quad (1)$$

where $H : \mathcal{X} \rightarrow \mathcal{Y}$ is the system model, which is usually linear, and $f : \mathcal{Y} \times \mathcal{Y} \rightarrow \mathbb{R}^+$ is an appropriate measure of error.

Table 1. Reviewed applications and associated references.

Denoising	Deconvolution	Superresolution	MRI	CT
[6]–[11]	[10], [12]–[14]	[9], [15]–[20]	[21]–[23]	[24]–[27]

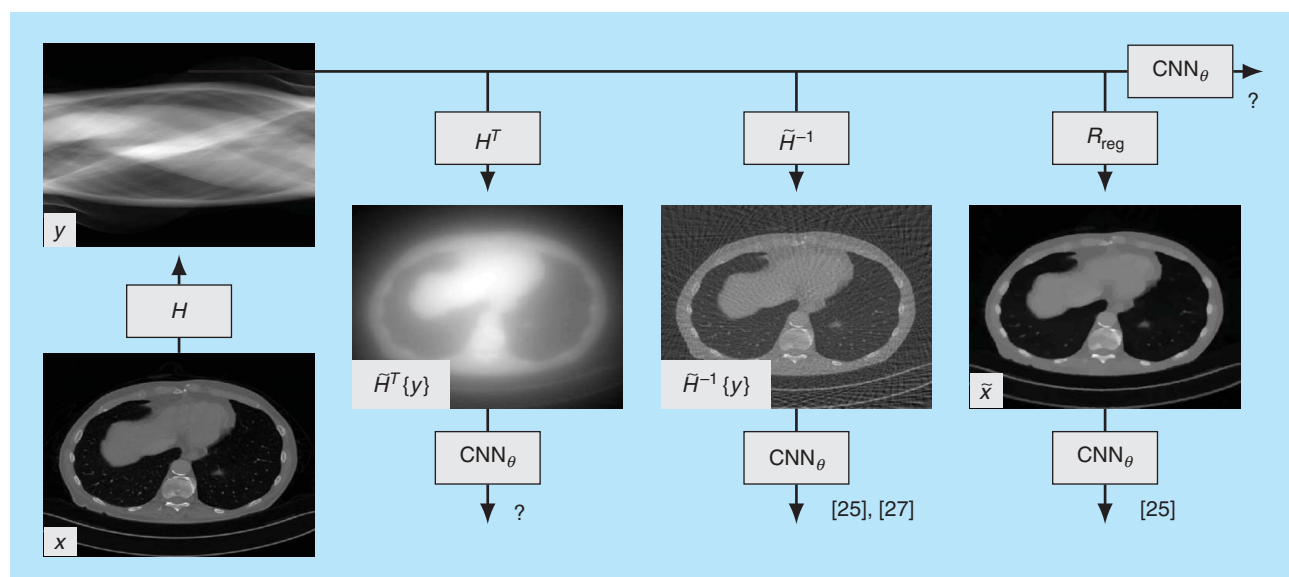


FIGURE 1. A block diagram of image reconstruction methods, using images from X-ray CT as examples. An image, x , creates measurements, y , that can be used to estimate x in a variety of ways. The traditional approach is to apply a direct inversion, \tilde{H}^{-1} , which is artifact prone in the sparse-measurement case (note the stripes in the reconstruction). The current state of the art is a regularized reconstruction, R_{reg} , written, in general, in (2). Several recent works apply CNNs to the result of the direct inversion or an iterative reconstruction, but it might also be reasonable to use as input the measurements themselves or the back projected measurements.

Continuing the CT example, H would be a discretization of the X-ray transform (such as MATLAB's `radon`), and f could be the Euclidean distance, $\|H\{x\} - y\|_2$. For many applications, decades of engineering have gone into developing a fast and reasonably accurate inverse operator, \tilde{H}^{-1} , so (1) is easily approximated by $R_{\text{obj}}\{y\} = \tilde{H}^{-1}\{y\}$; for CT, \tilde{H}^{-1} is the filtered back projection (FBP) algorithm. An important, related operator is the back projection, $H^T: \mathcal{Y} \rightarrow \mathcal{X}$, which can be interpreted as the simplest way to put measurements back into the image domain (see Figure 1).

These direct inverses begin to show significant artifacts when the number or quality of the measurements decreases, either because the underlying discretization breaks down or because the inversion of (1) becomes ill posed (lacking a solution, lacking a unique solution, or being unstable with respect to the measurements). Unfortunately, in many real-world problems, measurements are costly (in terms of time, or, e.g., X-ray damage to the patient), which motivates us to collect as few as possible. To reconstruct from sparse or noisy measurements, it is often better to use a regularized formulation,

$$R_{\text{reg}}\{y\} = \underset{x \in \mathcal{X}}{\text{argmin}} f(H\{x\}, y) + g(x), \quad (2)$$

where $g: \mathcal{X} \rightarrow \mathbb{R}^+$ is a regularization functional that promotes solutions that match our prior knowledge of x and, simultaneously, makes the problem well posed. For CT, g could be the total variation (TV) regularization, which penalizes large gradients in x .

From this perspective, the challenge of solving an inverse problem is designing and implementing (2) for a specific application. Much effort has gone into designing general-purpose regularizers and minimization algorithms. For example, compressed sensing [29] provides sparsity-promoting regularizers. Nonetheless, in the worst case, a new application necessitates developing accurate and efficient H , g , and f , along with a minimization algorithm.

An alternative to the objective function approach is called the *learning approach*, where a training set of ground-truth images and their corresponding measurements, $\{(x_n, y_n)\}_{n=1}^N$, is known. A parametric reconstruction algorithm, R_{learn} , is then learned by solving

$$R_{\text{learn}} = \underset{R, \theta \in \Theta}{\text{argmin}} \sum_{n=1}^N f(x_n, R_{\theta}\{y_n\}) + g(\theta), \quad (3)$$

where Θ is the set of all possible parameters, $f: \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}^+$ is a measure of error, and $g: \Theta \rightarrow \mathbb{R}^+$ is a regularizer on the parameters with the aim of avoiding overfitting. Once the learning step is complete, R_{learn} can then be used to reconstruct a new image from its measurements.

To summarize, in the objective function approach, the reconstruction function is itself a regularized minimization

problem, while in the learning approach, the solution of a regularized minimization problem is a parametric function that can be used to solve the inverse problem. The learning formulation is attractive because it overcomes many of the limitations of the objective function approach: there is no need to handcraft the forward model, cost function, regularizer, and optimizer from (2). On the other hand, the learning approach requires a training set, and the minimization (3) is typically more difficult than (2) and requires a problem-dependant choice of f , g , and the class of functions described by R and Θ .

Finally, we note that the learning and objective function approaches describe a spectrum rather than a dichotomy. In fact, the learning formulation is strictly more general, including the objective function formulation as a special case. As we will discuss further in the section "Network Architecture," which (if any) aspects of the objective formulation approach to retain is a critical choice in the design of learning-based approaches to inverse problems in imaging.

CNNs

Our focus here is the formulation of (3) using CNNs. Using a CNN means, roughly, fixing the set of functions, R_{θ} , to be a sequence of (linear) filtering operations alternating with simple nonlinear operations. This class of functions is parametrized by the values of the filters used (also known as *filter weights*), and these filter weights are the parameters

over which the minimization occurs. For illustration, Figure 2 shows a typical CNN architecture.

We will discuss the theoretical motivations for using CNNs as the learning architecture for inverse problems in the section "Theory," but we mention some practical advantages here. First, the forward operation of a CNN consists of (usually small) convolutions and simple, pointwise nonlinear functions. This means that, once training is complete, the execution of R_{learn} is very fast and amenable to hardware acceleration on GPUs. Second, the gradient of (3) is computable via the chain rule, and these gradients again involve small convolutions, meaning that the parameters can be learned efficiently via gradient descent.

When the first CNN-based method entered the ImageNet Large-Scale Visual Recognition Challenge in 2012 [5], its error rate on the object localization and classification task was 15.3%, as compared to an error rate 26.2% for the next closest method and 25.8% for the 2011 winner. In subsequent competitions (2013–2016), the majority of the entries (and all of the winners) were CNN based and continued to improve substantially, with the 2016 winner achieving an error rate of just 2.99%. Can we expect such large gains in inverse problems? That is, can we expect denoising results to improve by an order of magnitude (20 dB) in the next few years? Next, we answer this question by surveying the results reported by recent CNN-based approaches to image reconstruction.

In the objective function approach, the reconstruction function is itself a regularized minimization problem, while in the learning approach, the solution of a regularized minimization problem is a parametric function that can be used to solve the inverse problem.

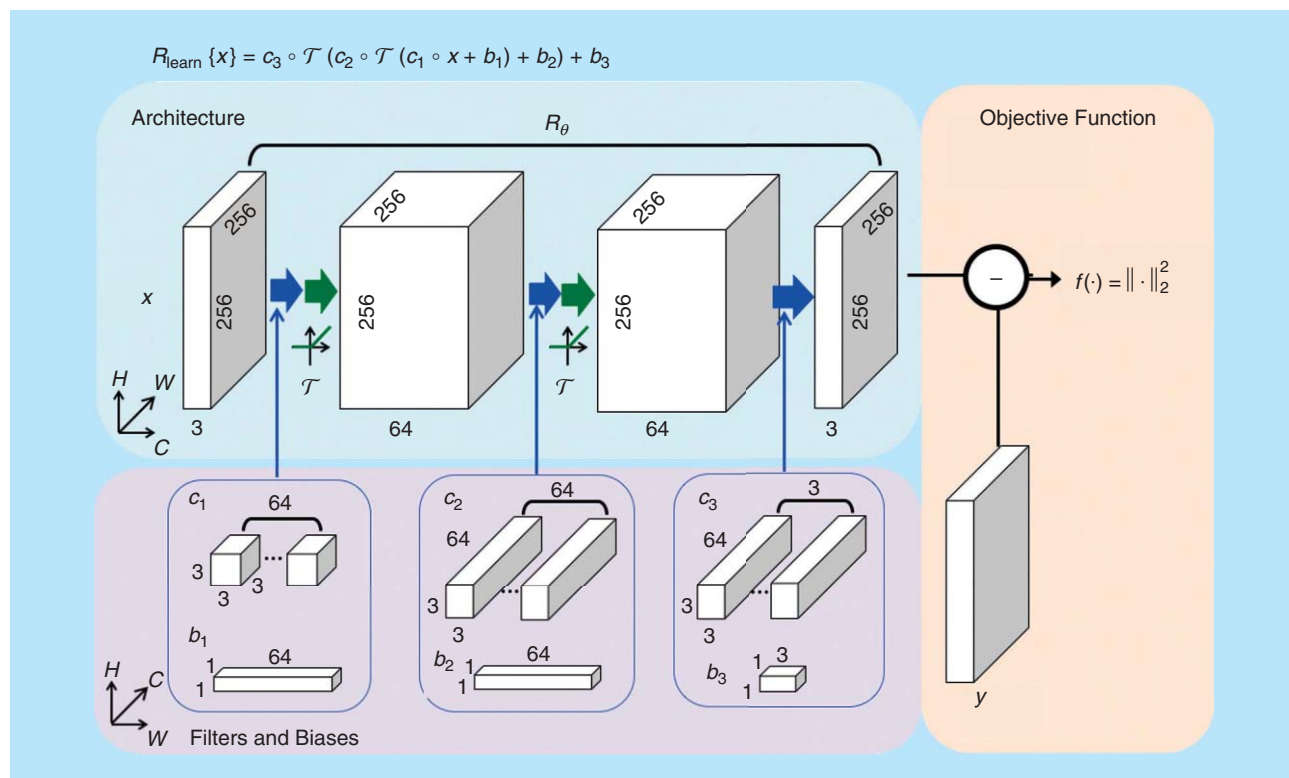


FIGURE 2. An illustration of a typical CNN architecture for 256^2 pixel RGB images, including the objective function used for training. $\mathcal{T}(\cdot)$ is the rectified linear unit function (point-wise nonlinear function). The symbol \circ denotes a 2-D convolution. The convolutions in each layer are described by a four-dimensional tensor representing a stack of 3-D filters.

Current state of performance

Of the inverse problems we review here, denoising provides the best look at recent trends in results because there are standard experiments that appear in most papers. Work on CNN-based denoising from 2009 [6] showed an average peak signal-to-noise ratio (PSNR) of 28.5 on the Berkeley segmentation data set, a less than 1-dB improvement over contemporary wavelet and Markov random field-based approaches. For comparison, one very recent denoising work [11] reported a 0.7-dB improvement on a similar experiment, which remains less than 1 dB better than contemporary non-CNN methods (including block-matching and 3-D filtering, which had remained the state of the art for years). As another point of reference, in 2012, one CNN approach [7] reported an average PSNR of 30.2 dB on a set of standard test images (Lena, peppers, etc.), less than 0.1 dB better than comparisons, and another [8] reported an average of 30.5 dB on the same experiment. Recently, [11] achieved an average of 30.4 dB under the same conditions. One important perspective on these denoising results is that the CNN is learning the distribution of natural images (or, equivalently, is learning a regularization). Such a CNN could be reused inside an iterative optimization as a proximal operator to enforce this learned regularization for any inverse problem.

The trends are similar in deblurring and superresolution, although experiments are more varied and therefore harder to compare. For deblurring, [12] showed around a 1-dB PSNR improvement over comparison methods, and [13] showed a

further improvement of approximately 1 dB. For superresolution, work from 2014 [15] reported a less than 0.5-dB improvement in PSNR over comparisons. During the next two years, [16] and [19] both reported a 0.5-dB PSNR increase over this baseline. Even more recent work, [30], improves on the 2014 work by around 1.5 dB in PSNR. For video superresolution, [18] improves on non-CNN-based methods by about 0.5 dB PSNR and [20] improves upon that result by another 0.5 dB.

For inverse problems in medical imaging, direct comparison between works is impossible due to the wide variety of experimental setups. A 2013 CNN-based work [24] shows improvement in limited-view CT reconstruction over direct methods and unregularized iterative methods but does not compare to regularized iterative methods. In 2015, [25] showed (in full-view CT) an improvement of several decibels in signal-to-noise ratio (SNR) over direct reconstruction and around 1-dB improvement over regularized iterative reconstruction. Recently, [26] showed about 0.5-dB improvement in PSNR over TV-regularized reconstruction, while [27] showed a larger (1–4 dB) improvement in SNR over a different TV-regularized method (Figure 3). In magnetic resonance imaging (MRI), [22] demonstrates performance equal to the state of the art, with advantages in running time.

Do these improvements matter? CNN-based methods have not, so far, had the profound impact on inverse problems that they have had for object classification. The difference between 30 and 30.5 dB is impossible to see by eye. On the other hand,

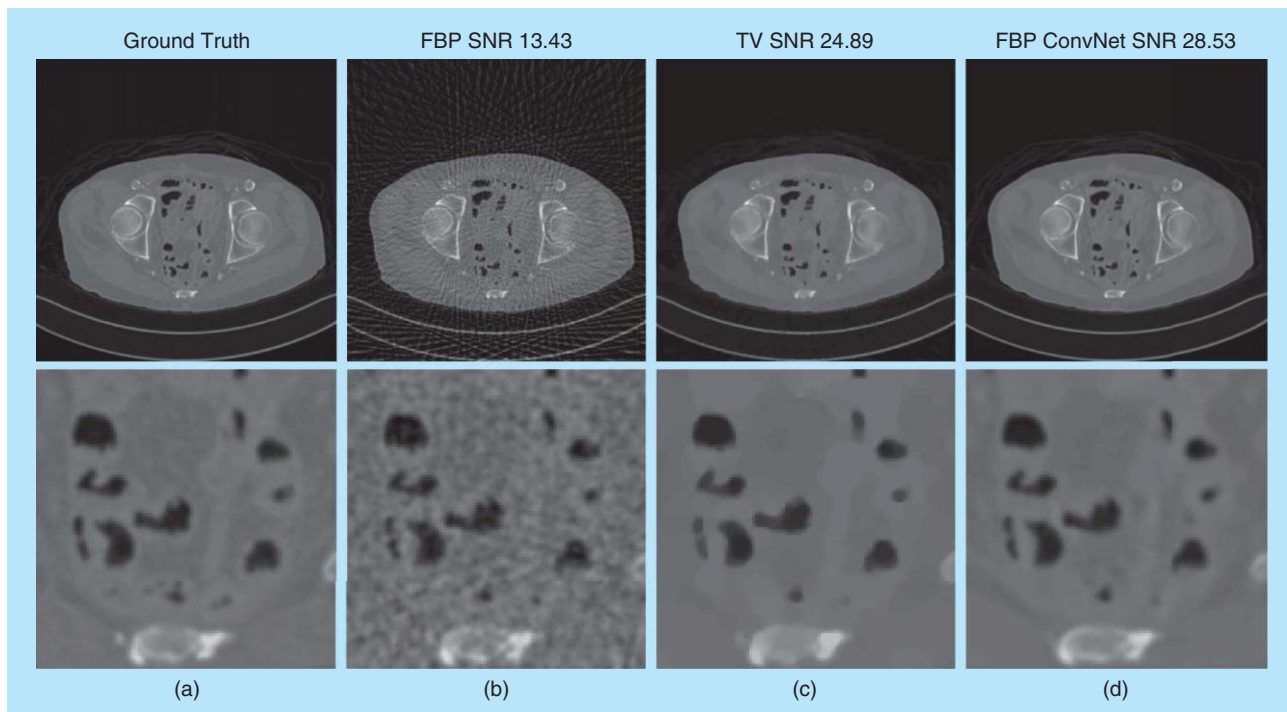


FIGURE 3. An example of X-ray CT reconstructions. (a) The ground truth comes from an FBP reconstruction using 1,000 views. (b)–(d) are reconstructions from just 50 views using FBP, a regularized reconstruction, and from a CNN-based approach. The CNN-based reconstruction preserves more of the texture present in the ground truth and results in a significant increase in SNR. (Images are reproduced with permission from [27]).

these improvements occur in heavily studied fields: we have been denoising the Lena image since the 1970s. Furthermore, CNNs offer some unique advantages over many traditional methods. The design of the CNN architecture can be more or less decoupled from the application at hand and reused from problem to problem. They can also be expanded in straightforward ways as computer memory grows, and there is some evidence that larger networks lead to better performance. Finally, once trained, running the model is fast (dozens of convolutions per image, usually less than 1 s). This means that CNN-based methods can be attractive in terms of running time even if they do not improve upon state-of-the-art performance.

Designing CNNs for inverse problems

In this section, we survey the design decisions needed to develop CNN-based approaches for inverse problems in imaging. We organize the section around the learning equation as summarized in Figure 4, first describing how the training set is created, then how the network architecture is designed, and, finally, how the learning problem is formulated and solved.

Training set

Learning requires a suitable training set, i.e., the (input, output) pairs from which the CNN will learn. In a typical learning problem, training outputs are provided by some oracle labeling a set of inputs. For example, in object classification, a set of human graders might view a large number of images and provide annotations for each. In the inverse problem setting, this is considerably more difficult because no such oracle exists.

For example, in X-ray CT, to generate a training set, we would need to image a large number of physical phantoms for which we have exact 3-D models, which is not feasible in practice. The choice of the training set also constrains the network architecture because the input and output of the network must match the dimensions of y_n and x_n , respectively.

Generating training data

In some cases, generating training data is straightforward because the forward model we aim to invert is known exactly and easily computable. In denoising, training data are generated by corrupting images with noise; the noisy image then serves as training input and the clean image as the training output, as in, e.g., [6] and [7]. Or, the noise itself can serve as the oracle output, in a scheme called *residual learning* [11], [23]. Super-resolution follows the same pattern, where training pairs are easily generated by downsampling, as in, e.g., [19]. The same is true

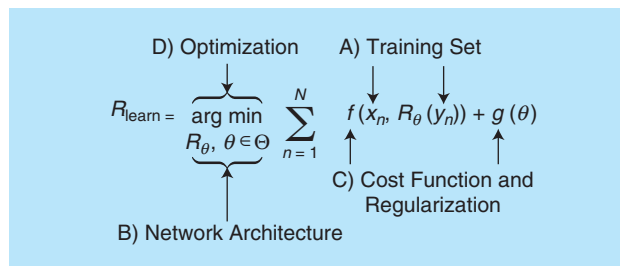


FIGURE 4. The learning equation, which we use to organize the parts of the section “Designing CNNs for Inverse Problems”.

for deblurring, where training pairs can be generated by blurring [12]–[14].

In medical imaging, the focus is on reconstructing from real measurements, and the corresponding ground truth is not usually known. The emerging paradigm is to learn to reconstruct from sparse measurements, using reconstructions from fully sampled measurements to train. For example, in MRI reconstruction, [22] trains by using undersampled k-space data as inputs and reconstructions from fully sampled k-space data as outputs. Likewise, [27] uses a low-view CT reconstruction as input and a high-view CT reconstruction as output. Or the CNN can learn from low-dose (noisy) measurements [25].

Preprocessing

Another aspect of training data preparation is whether the training inputs are the measurements themselves or whether some preprocessing occurs. In denoising, it is natural to use the raw measurements, which are of the same dimensions as the reconstruction. But, in the other applications, the trend is to use a direct inverse operator to preprocess the network input. Following the notation in the section “Learning for Inverse Problems in Imaging,” this can be viewed as a combination of the objective function and learning approach, where instead of R_{learn} being a CNN, it is the composition of a CNN with a direct inverse: $R_{\theta} \circ \tilde{H}^{-1}$. For example, in superresolution, [16], [18], and [19] first upsample and interpolate the low-resolution input images; in CT, [25] and [27] preprocess with the FBP ([25] also preprocesses with an iterative reconstruction); and, in MRI, [21] preprocesses with the inverse Fourier transform.

Without preprocessing, the CNN must learn the underlying physics of the inverse problem. It is not even clear that this is possible with CNNs (e.g., what is the meaning of filtering an X-ray CT sinogram?). Preprocessing is also a way to leverage the significant engineering effort that has gone into designing these direct inverses over the past decades. Superficially, this type of preprocessing appears to be inversion followed by denoising, which is a standard, if ad hoc, approach to inverse problems. What is unique here is that the artifacts caused by direct inversion, especially in the sparse measurement case, are usually highly structured and therefore not good candidates for generic denoising approaches. Instead, the CNN is allowed to learn the specific character of these artifacts.

A practical aspect of preprocessing is controlling the dynamic range of the input. While not typically a problem when working with natural images or standardized data sets, there may be huge fluctuations in the intensity or contrast of the measurements in certain inverse problems. To avoid a small set of images dominating the error during training, it is best to scale the dynamic range of the training set [23], [27]. Similarly, it may be advantageous to discard training patches without sufficient contrast.

Training size

CNNs typically have at least thousands of parameters to train; thus, the number of (input, output) pairs in the training set is of important practical concern. The number of training pairs varied among the papers we surveyed. The biomedical imaging papers tended to have the fewest samples (e.g., 500 brain images [21] or 2,000 CT images [24]), while papers on natural images had the most (e.g., pretraining with 395,909 natural images [20]).

A further complication is that, depending on the network architecture, images may be split into patches for training. Thus, depending on the dimensions of the training images and the chosen patch size, numerous patches can be created from a small training set. The patch size also has important ramifications for the performance of the network and is linked to its architecture, with larger filters and deeper networks requiring larger training patches [17].

With a large CNN and a small training set, overfitting must be avoided by regularization during learning and/or the use of a

validation set (e.g., [24] and discussed more in the sections “Cost Function and Regularization” and “Optimization”). These strategies are necessary to produce a CNN that generalizes at all, but they do not overcome the fact that the performance of the CNN will be limited by the size and variety of the training set. One strategy to increase the training set size is data augmentation, where new (input, output) pairs are generated by

transforming existing ones. For example, [20] augmented training pairs by scaling them in space and time, turning 20,000 pairs into 70,000 pairs. The augmentation must be application specific because the trained network will be approximately invariant to the transforms used. Another strategy to effectively increase the training set size is to use a pretrained network. For example, [18] first trains a CNN for image superresolution with a large image data set, then retrains with videos.

Network architecture

By network architecture, we mean the choice of the family of CNNs, R_{θ} parameterized by θ . In our notation, R_{θ} represents a CNN with a specific architecture, while θ are the weights to be learned during the training. There is great variety among CNN-based methods regarding their architecture: how many convolutional layers, what filter sizes, which nonlinearities, etc. For example, [19] uses 8,032 parameters, while [20] uses on the order of 100,000. In this section, we survey recent approaches to CNN architecture design for inverse problems.

The simplest approach to architecture design is simply a stack of series of convolutional layers and nonlinear functions [26], [10]; see Figure 2. This provides a baseline to check the feasibility of the network for the given application. It is straightforward to adjust the size of such a network, either by changing the number of layers, the number of channels per layer, or the size of the filters in each layer. For example, keeping the filters small (3×3 pixels) allows the network to be deeper for a given

Can we expect large gains in inverse problems, such as improving the denoising results by an order of magnitude in the next few years?

number of parameters [23]; constraining the filters to be separable [12] further reduces the number of parameters. Doing this can give the experimenter a sense of the training time required on their hardware as well as the effects of the network size on performance. From this simple starting point, the architecture can be tweaked for greater performance; for example, by adding downsampling and upsampling operations [27] or by simply adding more layers [20].

Instead of using ad hoc architecture design, one can adapt a successful CNN architecture from another application. For example, [27] adapts a network designed for biomedical image segmentation to CT reconstruction by changing the number of output layers from two (background and foreground images) to one (reconstructed image). These architectures can also be connected end to end, creating modular or hierarchical designs. For example, a four-times superresolution architecture can be created by connecting two two-times superresolution networks [16]. This is distinct from training a two-times superresolution network and applying it twice because the two modules of the CNN are trained as a unit.

A second approach is to begin with an iterative optimization algorithm and unroll it, turning each iteration into a layer of a network. In such a scheme, filters that are normally fixed in the iterative minimization are instead learned. The approach was pioneered in [31] for sparse coding; their results showed that the learned algorithms could achieve a given error in fewer iterations than the standard ones. Because many iterative optimization algorithms alternate filtering steps (linear updates) with pointwise nonlinear steps (proximal/shrinkage operations), the resulting network is often a CNN. This was the approach in [22], where the authors unrolled the alternating direction method of multipliers (ADMM) algorithm to design a CNN for MRI reconstruction, with state-of-the-art results and improvements in running time. For networks designed in this way, the original algorithm is a specific case, and, therefore, the performance of the network cannot be worse than the original algorithm if training is successful. The concept of unrolling can also be applied at a coarser scale, as in [13], where the modules of the network mimic the steps of a typical blind deconvolution pipeline: extract features, estimate kernel, estimate image, repeat.

Another promising design approach, similar to unrolling, is to learn only some part of an existing iterative method. For example, given the modular nature of popular iterative optimization schemes such as the ADMM, a CNN can be employed as a proximal (denoising) operator, while the rest of the algorithm remains unchanged [32]. This design combines many of the good aspects of both the objective function and learning-based approaches and allows a single CNN to be used for several different inverse problems without retraining.

Cost function and regularization

In this section, we survey the approaches taken to actually train the CNN, including the choice of a cost function, f ,

and regularizer, g . For a textbook coverage of the subject of learning, see [33].

Understanding the learning minimization problem as a statistical inference can provide useful insight into the selection of the cost and regularization functions. From this perspective, we can formulate the goal of learning as maximizing the conditional likelihood of each training output given the corresponding training input and CNN parameters:

$$\text{given } \{(x_n, y_n)\}_{n=1}^N,$$

$$R_{\text{learn}} = \arg \max_{R, \theta \in \Theta} \prod_{n=1}^N P(y_n | x_n, \theta),$$

where P is a conditional likelihood. When this likelihood follows a Gaussian distribution, this optimization is equivalent to the one from the “Background” section, (3), with f being

the Euclidean distance and no regularization. Put another way, learning with the standard, Euclidean cost, and no regularization implicitly assumes a Gaussian noise model; this is a well-known fact in inverse problems in general. This formulation is used in most of the works we surveyed [6], [7], [11], [12], [18], [19], [23], [25], [26], despite the fact that several raise questions about whether it is the best choice [25], [34].

An alternative is the maximum a posteriori formulation, which maximizes the joint probability of the training data and

the CNN parameters, which can be decomposed into several terms using Bayes’ rule:

$$\text{given } \{(x_n, y_n)\}_{n=1}^N,$$

$$R_{\text{learn}} = \arg \max_{R, \theta \in \Theta} \prod_{n=1}^N P(y_n | x_n, \theta) P(\theta). \quad (4)$$

This formulation explicitly allows prior information about the desired CNN parameters, θ , to be used. Under a Gaussian model for the weights of the CNN as well as the noise, this formulation results in a Euclidean cost function and a Euclidean regularization on the weights of the CNN, $g(\theta) = \sigma^{-2} \|\theta\|_2^2$. Other examples of regularizations for CNNs are the total generalized variation norm or sparsity on the coefficients. Regularized approaches are taken in [10], [15], [21], and [22].

Optimization

Once an objective function for learning has been fixed, it still must be actually minimized. This is a crucial and deep topic, but, from the practical perspective, it can be treated as a black box due to the availability of several high-quality software libraries that can perform efficient training of user-defined CNN architectures. For a comparison of these libraries, refer to [35]; here, we provide a basic overview.

The popular approaches to CNN learning are variations on gradient descent. The most common is stochastic gradient descent (SGD), used, e.g., in [16] and [25], where, at each

The notion of universal approximation tells us what the network can learn, not what it does learn, and comparison to established algorithms can help guide our understanding of CNNs in practice.

iteration, the gradient of the cost function is computed using random subsets of the available training. This reduces the overall computation compared to computing the true gradient, while still providing a good approximation. The process can be further tuned by adding momentum, i.e., combining gradients from previous iterations in clever ways or by using higher-order gradient information as in BFGS [22].

Initial weights can be set to zero or chosen from some random distribution (Gaussian or uniform). Because learning is nonconvex, the initialization does potentially change which minimum the network converges to, but not much difference is observed in practice. However, good initializations can improve the speed of convergence. This explains the popularity of taking pretrained networks, or, in the case of an unrolled architecture, initializing the network weights based on corresponding known filters. Recently, a procedure called *batch normalization*, where the inputs to each layer of the network are normalized, was proposed as a way to increase learning speed and reduce sensitivity to initialization [36].

As mentioned in the section “Training Set,” overfitting is a serious risk when training networks with potentially millions of parameters. In addition to augmenting the training set, steps can be taken during training to reduce overfitting. The simplest is to split the training data into a set used for optimization and a set used for validation. During training, the performance of the network on the validation set is monitored, and training is terminated when the performance on the validation set begins to decrease. Another method is dropout [37], where individual units of the network are randomly deleted during training. The motivation for dropout is the idea that the network should be regularized by forming a weighted average of all possible parameter settings, with weights determined by their performance. While this regularization is not feasible, removing units during training provides a reasonable approximation that performs well in practice.

Theory

The excellent performance of CNNs for various applications is undisputed, but the question of “Why?” remains mostly unanswered. Here, we bring together a few different theoretical perspectives that begin to explain why CNNs are a good fit for solving inverse problems in imaging.

Universal approximation

We know that neural networks are universal approximators. More specifically, a fully connected neural network with one hidden layer can approximate any continuous function arbitrarily well, provided that its hidden layer is large enough [38]. The result does not directly apply to CNNs because they are not fully connected, but, if we consider the network patch by patch, we see that each input patch is mapped to the corresponding output patch by a fully connected network. Thus, CNNs are universal approximators for shift-invariant functions. From this perspective, statements such as “CNNs work

well because they generalize X algorithm” are vacuously true because CNNs generalize all shift-invariant algorithms. On the other hand, the notion of universal approximation tells us what the network can learn, not what it does learn, and comparison to established algorithms can help guide our understanding of CNNs in practice.

Unrolling

The most concrete perspective on CNNs as generalizations of established algorithms comes from the idea of unrolling, which we discussed in the section “Network Architecture.” The idea originated in [31], where the authors unrolled the ISTA algorithm for sparse coding into a neural network. This network is not a typical CNN because it includes recurrent connections, but it does share the alternating linear/nonlinear motif. A more general perspective is that nearly all state-of-the-art iterative reconstruction algorithms alternate between linear steps and pointwise nonlinear steps, so it follows that CNNs should be able to perform similarly well given appropriate training. One

refinement of this idea comes from [27], which establishes conditions on the forward model, H , that ensure that the linear step of the iterative method is a convolution. All of the inverse problems surveyed here meet these conditions, but the theory predicts that certain inverse problems, e.g., structured illumination microscopy, should not be amenable to reconstruction via CNNs. Another

refinement concerns the popular rectified linear unit (ReLU) employed as the nonlinearity by most CNNs: results from spline theory can be adapted to show that combinations of ReLUs can approximate any continuous function. This suggests that the combinations of ReLUs usually employed in CNNs are able to closely approximate the proximal operators used in traditional iterative methods.

Invariance

Another perspective comes from work on scattering transforms, which are cascades of linear operations (convolutions with wavelets) and nonlinearities (absolute value) [39] with no combinations formed between the different channels. This simplified model shows invariance to translation and, more importantly, to small deformations of the input (diffeomorphisms). CNNs generalize the scattering transform, giving the potential for additional invariances, e.g., to rigid transformations, frequency shifts, etc. Such invariances are attractive for image classification, but more work is needed to connect these results to inverse problems.

Critiques

While the papers we have surveyed present many reasons to be optimistic about CNNs for inverse problems, we also want to mention a few general critiques of the approach. We hope these can be useful points to think about when writing or reviewing manuscripts in the area, as well as jumping-off points for future research.

The most concrete perspective on CNNs as generalizations of established algorithms comes from the idea of unrolling.

Algorithm descriptions and reproducibility

When planning this survey, we aimed to measure quantitative trends in the literature, e.g., to plot the number of training samples versus the number of parameters for each network. We quickly discovered this is nearly impossible. Very few manuscripts clearly noted the number of parameters they were training, and only some provided a clear-enough description of the network for us to calculate the value. Many more included a figure of network architecture along the lines of Figure 2, but without a clear statement of the dimensions of each layer. Similar problems exist in the description of the training and evaluation procedures, where it is not always clear whether the evaluation data come from simulation or from a real data set. As the field matures, we hope papers converge on a standard way to describe network architecture, training, and evaluation.

The lack of clarity presents a barrier to the reproducibility of the work. Another barrier is the fact that training often requires specialized or expensive hardware. While GPUs have become more ubiquitous, the largest (and best-performing) CNNs remain difficult for small research groups to train. For example, the CNN that won the ImageNet Large-Scale Visual Recognition Challenge in 2012 took “between five and six days to train on two GTX 580 3GB GPU” [5].

Robustness of learning

The success of any CNN-based algorithm hinges on finding a reasonable solution to the learning problem (3). As stated previously, this is a nonconvex problem, where the best solution we can hope for is to find one of many local minima of the cost. This raises questions about the robustness of the learning to changes in the initialization of parameters and the specifics of the optimization method employed. This is in contrast to the typical convex formulations of inverse problems, where the specifics of the initialization and optimization scheme probably do not affect the quality of the result.

The uncertainty about learning complicates the comparison of any two CNN-based methods. Does A outperform B because of its superior architecture, or simply because the optimization of A fell into a superior local minimum? As an example of the confusion this can cause, [34] shows, in the context of denoising, superresolution, and JPEG deblocking that a network trained with the l_1 cost function can outperform a network trained with the l_2 cost function even with regard to the l_2 cost. In the authors' analysis of this highly disturbing result, they attribute it to the l_2 learning being stuck in a local optimum. Regardless, the vast majority of work relies on the l_2 cost, which is computationally convenient and provides excellent results.

There is some indication that large networks trained with lots of data can overcome this problem. In [40], the authors show that larger networks have more local minima, but that most local minima are equivalent in terms of testing performance. They also identify that the global minima likely correspond to parameter settings that overfit the training set. More work on the stability of the learning process will be an important step toward wider acceptance of CNNs in the inverse problem community.

More generally, how sensitive are the results of a given experiment to small changes in the training set, network architecture, or optimization procedure? Is it possible for the experimenter to overfit the testing set by iteratively tweaking the network architecture (or the experimental parameters) until state-of-the-art results are achieved? To combat this, CNN-based approaches should provide carefully constructed experiments with results reported on a large number of testing images. Even better are competition data sets, where the testing data is hidden until algorithm development is complete.

Can we trust the results?

Once trained, CNNs remain nonlinear and highly complex. Can we trust reconstructions generated by such systems? One way to look at this is to evaluate the sensitivity of the network to noise: ideally, small changes to the input should cause only small changes to the output; data augmentation during training can help achieve this. Similarly, demonstrating generalization between data sets (where the network learns on one data set, but is evaluated on another) can help improve confidence in the results by showing that the performance of the network is not dependent on some systematic bias of the data set.

A related question is how to measure the quality of the results. Even if a robust SNR improvement can be demonstrated, practitioners will inevitably want to know, e.g., whether the resulting images can be reliably used for diagnosis. To this end, as much as possible, methods should be assessed with respect to the ultimate application of the reconstruction (diagnosis, quantification of biological phenomenon, etc.) rather than an intermediate measure such as SNR or structural similarity (SSIM). While this critique can be made of any approach to inverse problems, it is especially relevant for CNNs because they are often treated as black boxes and because the reconstructions they generate are plausible-looking by design, hiding areas where the algorithm is less sure of the result.

Next steps

So far, we have given a small look into the wide variety of ways that researchers have applied CNNs to solve inverse problems in imaging. Because CNNs are so powerful and flexible, we believe there is plenty of room to create even better systems. Next, we suggest a few directions that this future research might take.

Biomedical imaging

CNNs have so far been applied mostly to inverse problems where the measurements take the form of an image and the measurement model is simple, and less so for CT and MRI, which have relatively more complicated models. A search on arXiv.org reveals dozens more CT and MRI papers submitted within the last few months, suggesting many incoming contributions in these areas. We expect diffusion into other modalities such as positron-emission tomography, single-photon emission CT, transmission electron microscopy, structured illumination microscopy, ultrasound, superresolution microscopy, etc. to follow.

Central to this work will be questions of how best to combine CNNs with knowledge of the underlying physics as well as direct and iterative inversion techniques. Most of the surveyed works involve using a CNN to correct the artifacts created by direct or iterative methods, where it remains an open question what is the best such prereconstruction method. One creative approach is to build the inverse operator into the network architecture as in [22], where the network can compute inverse Fourier transforms. Another would be to use the back-projected measurements, $H^T y$, which at least take the form of an image and could reduce the burden on the CNN to learn the physics of the forward model. CNNs could be deployed in a variety of other ways here, too, e.g. using a CNN to approximate a high quality, but extremely slow reconstruction method. With enough computing power, a training set could be generated by running the slow method on real data, and, once trained, the resulting network could provide very fast and accurate reconstructions.

Cross-task learning

In cross-task learning (also called *transfer learning*, although this can have other meanings as well), an algorithm is trained with one data set and deployed on a different, but related, task. This is attractive in the inverse problem setting because it avoids the costly retraining of the network when imaging parameters change (different noise levels, image dimensions, etc.), which may occur often. Or we could imagine a network that transfers between completely different imaging modalities, especially when training data for the target modality are scarce; e.g., a network could train on denoising natural images and then be used to reconstruct MRI images. Recent work has made progress in this direction by learning a CNN-based proximal operator, which can be used inside an iterative optimization method for any inverse problem [32].

Multidimensional signals

Modern inverse problems in imaging increasingly involve reconstruction of 3-D or 3-D+time images. However, most CNN-based approaches to these problems involve 2-D inputs and outputs. This is likely because much of the work on deep neural networks in general has been in two dimensions and because of practical considerations. Specifically, learning strongly relies on GPU computation, but current GPUs have maximally 24 GB of physical memory. This limitation makes training a large network with 3-D inputs and outputs infeasible.

One way to overcome this issue is model parallelism, in which a large model is partitioned onto separable computers. Another is data parallelism, where it is the data that are split. When used together, large computational gains are achieved [41]. Such approaches will be key in tackling multidimensional imaging problems.

Generative adversarial networks and perceptual loss

CNN-based approaches to inverse problems also stand to benefit from new developments in neural network research. One such development is the generative adversarial network (GAN) [42], which may offer a way to break current limits in supervised

learning. Basically, two networks are trained in competition: the generator tries to learn a mapping between training samples, while the discriminator attempts to distinguish between the output of the generator and real data. Such a setup can, e.g., produce a generator capable of creating plausible natural images from noise. The GAN essentially revises the learning formulation (3) by replacing the cost function f with another neural network. In contrast to a designed cost function, which will be suboptimal if the assumed noise model is incorrect, the discriminator network learns a cost function that models the probability density of the real data. GANs have already begun to be used for inverse problems, e.g., for superresolution in [30] and deblurring in [14].

A related approach is perceptual loss, where a network is trained to compute a loss function that matches human perception. The method has already been used for style transfer and superresolution [43]. Compared to the standard Euclidean loss, networks trained with perceptual loss give better looking results, but do not typically improve the SNR. It remains to be seen whether these ideas can gain acceptance for applications such as medical imaging, where the results must be quantitatively accurate.

Acknowledgment

As stated in the introduction, [30], [32], and [35] from arXiv.org are not peer reviewed. They have been included only to illustrate recent trends.

Authors

Michael T. McCann (michael.mccann@epfl.ch) received his B.S.E. degree in biomedical engineering in 2010 from the University of Michigan and the Ph.D. degree in biomedical engineering from Carnegie Mellon University, Pittsburgh, Pennsylvania, in 2015. He is currently a scientist with the Laboratoire d'imagerie biomédicale and Centre d'imagerie biomédicale, École Polytechnique Fédérale de Lausanne, where he works on X-ray computed tomography reconstruction. His research interest centers on developing signal processing tools to answer biomedical questions.

Kyong Hwan Jin (kyong.jin@epfl.ch) received his B.S. degree and integrated M.S. and Ph.D. degrees from the Department of Bio and Brain Engineering at Korea Advanced Institute of Science and Technology (KAIST), Daejeon, in 2008 and 2015, respectively. He was a postdoctoral scholar at KAIST from 2015 to 2016. He is currently a postdoctoral scholar in the Biomedical Imaging group, École Polytechnique Fédérale de Lausanne, Switzerland. His research interests include low-rank matrix completion, sparsity-promoted signal recovery, sampling theory, biomedical imaging, and image processing in various applications.

Michael Unser (michael.unser@epfl.ch) received the M.S. and Ph.D. degrees in electrical engineering in 1981 and 1984, respectively, from the École Polytechnique Fédérale de Lausanne (EPFL). He is a professor and the director of the EPFL Biomedical Imaging Group, Lausanne, Switzerland. His primary area of investigation is biomedical image processing.

He was the associate editor-in-chief of *IEEE Transactions on Medical Imaging* from 2003 to 2005 and the founding chair of the technical committee on Bioimaging and Signal Processing of the IEEE Signal Processing Society (SPS). He is a member of the editorial boards of *SIAM Journal on Imaging Sciences* and *Foundations and Trends in Signal Processing*, a fellow of EURASIP, a member of the Swiss Academy of Engineering Sciences, and a Fellow of the IEEE. He received several international prizes, including three IEEE SPS Best Paper Awards and two Technical Achievement Awards from the IEEE (2008 SPS and Engineering in Medicine and Biology 2010).

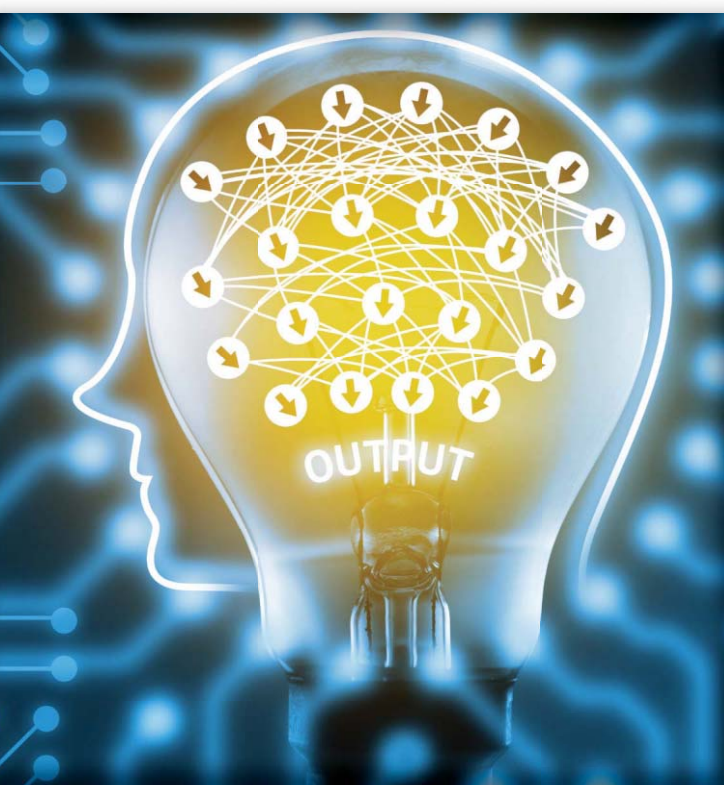
References

- [1] J. Schmidhuber, "Deep learning in neural networks: An overview," *Neural Netw.*, vol. 61, pp. 85–117, Jan. 2015.
- [2] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, May 2015.
- [3] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning internal representations by error propagation," in *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, vol. 1. Cambridge, MA: MIT Press, 1986, pp. 318–362.
- [4] Y. T. Zhou, R. Chellappa, A. Vaid, and B. K. Jenkins, "Image restoration using a neural network," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 36, no. 7, pp. 1141–1151, July 1988.
- [5] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proc. 25th Int. Conf. Neural Information Processing Systems*, Lake Tahoe, NV, 2012, pp. 1097–1105.
- [6] V. Jain and S. Seung, "Natural image denoising with convolutional networks," in *Proc. 21st Int. Conf. Neural Information Processing Systems*, Vancouver, British Columbia, Canada, 2008, pp. 769–776.
- [7] H. C. Burger, C. J. Schuler, and S. Harmeling, "Image denoising: Can plain neural networks compete with BM3D?" in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, Providence, RI, June 2012, pp. 2392–2399.
- [8] J. Xie, L. Xu, and E. Chen, "Image denoising and inpainting with deep neural networks," in *Proc. 25th Int. Conf. Neural Information Processing Systems*, Lake Tahoe, NV, 2012, pp. 341–349.
- [9] R. Wang and D. Tao, "Non-local auto-encoder with collaborative stabilization for image restoration," *IEEE Trans. Image Process.*, vol. 25, no. 5, pp. 2117–2129, May 2016.
- [10] Y. Chen and T. Pock, "Trainable nonlinear reaction diffusion: a flexible framework for fast and effective image restoration," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1256–1272, 2016.
- [11] K. Zhang, W. Zuo, Y. Chen, D. Meng, and L. Zhang, "Beyond a Gaussian denoiser: Residual learning of deep CNN for image denoising," *IEEE Trans. Image Process.*, vol. 26, no. 7, pp. 3142–3155, July 2017.
- [12] L. Xu, J. S. Ren, C. Liu, and J. Jia, "Deep convolutional neural network for image deconvolution," in *Proc. 27th Int. Conf. Neural Information Processing Systems*, Cambridge, MA, 2014, pp. 1790–1798.
- [13] C. J. Schuler, M. Hirsch, S. Harmeling, and B. Schlopp, "Learning to deblur," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 7, pp. 1439–1451, July 2016.
- [14] K. Schawinski, C. Zhang, H. Zhang, L. Fowler, and G. K. Santhanam, "Generative adversarial networks recover features in astrophysical images of galaxies beyond the deconvolution limit," *Monthly Notices Royal Astronomical Soc.: Lett.*, vol. 467, no. 1, pp. L110–L114, May 2017.
- [15] Z. Cui, H. Chang, S. Shan, B. Zhong, and X. Chen, "Deep network cascade for image super-resolution," in *Proc. Computer Vision Conf.*, Sept. 2014, pp. 49–64.
- [16] Z. Wang, D. Liu, J. Yang, W. Han, and T. Huang, "Deep networks for image super-resolution with sparse prior," in *Proc. IEEE Int. Conf. Computer Vision*, Santiago, Chile, 2015, pp. 370–378.
- [17] J. Kim, J. Kwon Lee, and K. Mu Lee, "Accurate image super-resolution using very deep convolutional networks," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, Las Vegas, NV, 2016, pp. 1646–1654.
- [18] A. Kappeler, S. Yoo, Q. Dai, and A. K. Katsaggelos, "Video super-resolution with convolutional neural networks," *IEEE Trans. Comput. Imaging*, vol. 2, no. 2, pp. 109–122, June 2016.
- [19] C. Dong, C. C. Loy, K. He, and X. Tang, "Image super-resolution using deep convolutional networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 2, pp. 295–307, Feb. 2016.
- [20] D. Li and Z. Wang, "Video super-resolution via motion compensation and deep residual learning," *IEEE Trans. Comput. Imaging*, vol. PP, no. 99, pp. 1–1, 2017.
- [21] S. Wang, Z. Su, L. Ying, X. Peng, S. Zhu, F. Liang, D. Feng, and D. Liang, "Accelerating magnetic resonance imaging via deep learning," in *Proc. IEEE 13th Int. Symp. Biomedical Imaging*, Apr. 2016, pp. 514–517.
- [22] Y. Yang, J. Sun, H. Li, and Z. Xu, "Deep ADMM-net for compressive sensing MRI," in *Proc. 29th Int. Conf. Neural Information Processing Systems*, Barcelona, Spain, 2016, pp. 10–18.
- [23] O. Oktay, W. Bai, M. Lee, R. Guerrero, K. Kamnitsas, J. Caballero, A. D. Marva, S. Cook, D. O'Regan, and D. Rueckert, "Multi-input cardiac image super-resolution using convolutional neural networks," in *Proc. Medical Image Computing and Computer-Assisted Intervention*. Cham, Switzerland: Springer, 2016, pp. 246–254.
- [24] D. M. Pelt and K. J. Batenburg, "Fast tomographic reconstruction from limited data using artificial neural networks," *IEEE Trans. Image Process.*, vol. 22, no. 12, pp. 5238–5251, Dec. 2013.
- [25] D. Boubil, M. Elad, J. Shtok, and M. Zibulevsky, "Spatially-adaptive reconstruction in computed tomography using neural networks," *IEEE Trans. Med. Imag.*, vol. 34, no. 7, pp. 1474–1485, July 2015.
- [26] H. Chen, Y. Zhang, W. Zhang, P. Liao, K. Li, J. Zhou, and G. Wang, "Low-dose CT via convolutional neural network," *Biomed. Opt. Express*, vol. 8, no. 2, pp. 679–694, Feb. 2017.
- [27] K. H. Jin, M. T. McCann, E. Froustey, and M. Unser, "Deep convolutional neural network for inverse problems in imaging," to be published.
- [28] A. Kirsch, *An Introduction to the Mathematical Theory of Inverse Problems*. New York: Springer, 2011, vol. 120.
- [29] E. J. Candes, J. Romberg, and T. Tao, "Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information," *IEEE Trans. Inf. Theory*, vol. 52, no. 2, pp. 489–509, Feb. 2006.
- [30] C. Ledig, L. Theis, F. Huszar, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, and W. Shi, "Photo-realistic single image super-resolution using a generative adversarial network," *arXiv Preprint*, arXiv:1609.04802 [cs, stat], Sept. 2016.
- [31] K. Gregor and Y. LeCun, "Learning fast approximations of sparse coding," in *Proc. 27th Int. Conf. Machine Learning*, Haifa, Israel, 2010, pp. 399–406.
- [32] K. Zhang, W. Zuo, S. Gu, and L. Zhang, "Learning deep CNN denoiser prior for image restoration," *arXiv Preprint*, arXiv:1704.03264 [cs], Apr. 2017.
- [33] T. M. Mitchell, *Machine Learning*. New York: McGraw-Hill Education, Mar. 1997.
- [34] H. Zhao, O. Gallo, I. Frosio, and J. Kautz, "Loss functions for image restoration with neural networks," *IEEE Trans. Comput. Imaging*, vol. 3, no. 1, pp. 47–57, Mar. 2017.
- [35] S. Bahrapour, N. Ramakrishnan, L. Schott, and M. Shah, "Comparative study of deep learning software frameworks," *arXiv Preprint*, arXiv:1511.06435 [cs], Nov. 2015.
- [36] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proc. Int. Conf. Machine Learning*, 2015, Lille, France, pp. 448–456.
- [37] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 1929–1958, Jan. 2014.
- [38] K. Hornik, "Approximation capabilities of multilayer feedforward networks," *Neural Netw.*, vol. 4, no. 2, pp. 251–257, Mar. 1991.
- [39] S. Mallat, "Understanding deep convolutional networks," *Phil. Trans. R. Soc. A*, vol. 374, no. 2065, pp. 20150203, Apr. 2016.
- [40] A. Choromanska, M. Henaff, M. Mathieu, G. Ben Arous, and Y. LeCun, "The loss surfaces of multilayer networks," in *Proc. 18th Int. Conf. Artificial Intelligence and Statistics*, 2015, pp. 192–204.
- [41] J. Dean, G. Corrado, R. Monga, K. Chen, M. Devin, M. Mao, A. Senior, P. Tucker, K. Yang, Q. V. Le, et al., "Large scale distributed deep networks," in *Proc. 25th Int. Conf. Neural Information Processing Systems*, Lake Tahoe, Nevada, 2012, pp. 1223–1231.
- [42] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in Neural Information Processing Systems*, vol. 27, Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, Eds. Cambridge, MA: Curran Associates, 2014, pp. 2672–2680.
- [43] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *Computer Vision ECCV 2016 (Lecture Notes Series in Computer Science)*. Cham, Switzerland: Springer, 2016, pp. 694–711.

Dhanesh Ramachandram and
Graham W. Taylor

Deep Multimodal Learning

A survey on recent advances and trends



©ISTOCKPHOTO.COM/ZAPP2PHOTO

The success of deep learning has been a catalyst to solving increasingly complex machine-learning problems, which often involve multiple data modalities. We review recent advances in deep multimodal learning and highlight the state-of-the-art, as well as gaps and challenges in this active research field. We first classify deep multimodal learning architectures and then discuss methods to fuse learned multimodal representations in deep-learning architectures. We highlight two areas of research—regularization strategies and methods that learn or optimize multimodal fusion structures—as exciting areas for future work.

Introduction

Neural networks have made an impressive resurgence in recent years, after long-standing concerns about the ability to train deep models were successfully abated by a pioneering group of researchers who leveraged advances in algorithms, data, and computation [1]. This active research area now interests researchers in academia, but also industry, and it has resulted in state-of-the-art performance for many practical problems, especially in areas involving high-dimensional unstructured data such as in computer vision, speech, and natural language processing.

With the undeniable success of deep learning in the visual domain, the natural progression of deep-learning research points to problems involving larger and more complex multimodal data. Such multimodal data sets consist of data from different sensors observing a common phenomena, and the goal is to use the data in a complementary manner toward learning a complex task. One of the main advantages of deep learning is that a hierarchical representation can be automatically learned for each modality, instead of manually designing or handcrafting modality-specific features that are then fed to a machine-learning algorithm.

The goal of this article is to provide a comprehensive survey of the state of the art in deep multimodal learning and suggest future research directions by highlighting advances, gaps, and challenges in this active field. We believe this review is timely given the increasing number of deep-learning techniques

Digital Object Identifier 10.1109/MSP.2017.2738401
Date of publication: 13 November 2017

applied to multimodal data published in leading conferences and journals, as shown in Figure 1.

The crux of this article centers around two important areas of focus in deep multimodal learning research: 1) methods that use regularization techniques to improve cross-modality learning (see the section “Multimodal Regularization”) and 2) methods that attempt to find optimal deep multimodal architectures through search, optimization, or some learning procedure (see the section “Fusion Structure Learning and Optimization”).

Background

For the purposes of our review, we adopt the definition provided by Lahat et al. [2], where we consider phenomena or systems that are observed using multiple sensors and each sensor output can be termed a *modality* associated with a single data set. The underlying motivation to use multimodal data is that complementary information could be extracted from each of the modalities considered for a given learning task, yielding a richer representation that could be used to produce much improved performance compared to using only a single modality. There are many practical tasks that benefit from the use of multimodal data. In medical image analysis, for example, the use of multiple imaging modalities, such as computed tomography (CT), magnetic resonance imaging (MRI), and ultrasound imaging provides complementary information that is routinely used by medical experts in diagnosis and treatment. Applications involving human–computer interaction use depth and vision cues extensively for applications like immersive gaming and autonomous driving. Similar improvements in performance can be seen in biometric applications. In remote sensing applications, data from different sensors [intensity images, synthetic aperture radar, and light detection and ranging (LIDAR)] are often fused.

Techniques for multimodal data fusion, which cover different application domains, have long been investigated by the research community [3], [4]. Traditionally, combining the signals of multiple sensors has been investigated from a data fusion perspective. This is called *early fusion* or *data-level fusion* and focuses on how best to combine data from multiple sources, either by removing correlations between modalities or representing the fused data in a lower-dimensional common subspace. Techniques that accomplish one or both of these objectives include principal component analysis (PCA), independent components analysis, and canonical correlation analysis. The fused data are then presented to a machine-learning algorithm. When ensemble classifiers became popular in the early 2000s [5], researchers began applying multimodal fusion techniques that fell into the category known as *late fusion* or *decision-level fusion*. In general, these late-fusion strategies were much simpler to implement than early fusion, particularly when the different modalities varied significantly in terms of data dimensionality and sampling rates, and often resulted in improved performance.

As shown in the section “Intermediate Fusion,” popular deep neural network (DNN) architectures allow yet a third form of multimodal fusion, i.e., intermediate fusion of learned representations, offering a truly flexible approach to multimodal fusion for numerous practical problems. As deep-learning architectures

learn a hierarchical representation of the underlying data across its hidden layers, learned representations between different modalities can be fused at various levels of abstraction.

Deep-learning-based multimodal learning offers several advantages over conventional machine-learning methods, which are highlighted in Table 1. For many practical problems, deep-learning models often offer much improved performance for problems involving multimodal data. However, this entails several architectural design choices that we discuss next.

The first of these design choices relates to when to fuse different modalities. From a traditional data fusion standpoint, the practitioner could fuse the various input modalities at the data level and proceed to train a single machine-learning model, but, as we discuss in the section “Early Fusion,” this option can be rather challenging. Alternatively, a late-fusion option can also be considered, and we review several works in this category in the section “Late Fusion.” An important feature of deep learning is its ability to learn hierarchical representations from raw data. This feature can be exploited in multimodal learning to have a fine-grained control over how learned representations are fused. Therefore, a common practice in multimodal deep learning is to construct a shared representation or fusion layer that can merge incoming representations of modalities, thereby forcing the network to learn a joint representation of its inputs. The simplest fusion layer is a layer of hidden units, each of which receives input from all modalities. The flexibility of learning cross-modality shared representations at different levels of abstraction could be exploited to achieve better multimodal fusion results; however,

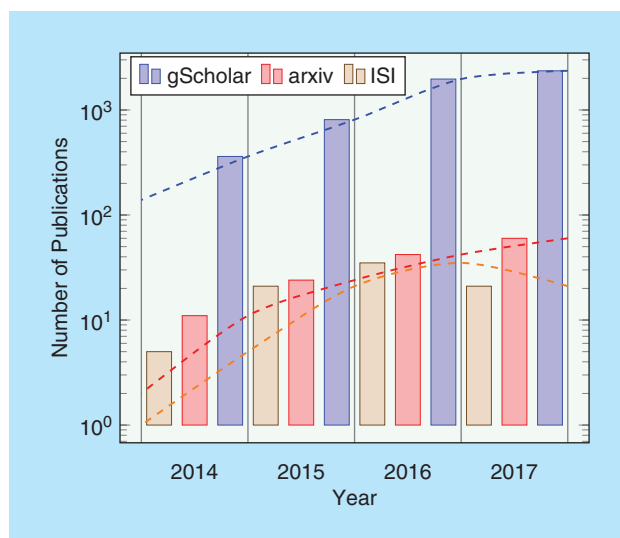


FIGURE 1. The increasing research interest in deep multimodal learning. Data were generated by analyzing the search results from leading search engines for technical publications: Google Scholar, the arXiv, and Thomson-Reuters’ ISI. We used the search terms *multimodal*, *fusion*, and *deep learning* and applied search filters to include references from engineering, computer science, and mathematics fields, excluding results related to social sciences, neurobiology, and business. Note that, due to differences in scale in terms of returned results, we use a semilog scale.

Table 1. A comparison between deep multimodal learning and conventional approaches.
Deep Multimodal Learning

Both modality-wise representations (features) and shared (fused) representations are learned from data.

Requires little or no preprocessing of input data (end-to-end training).

Implicit dimensionality reduction within architecture.

Supports early, late, or intermediate fusion.

Easily scalable in terms of data size and number of modalities for all fusion methods.

Fusion architecture can be learned during training.

Deeper, complex networks typically require large amounts of training data (if trained from scratch).

Numerous hyperparameter tunings vital for state-of-art performance.

Compute intensive, requires powerful graphics processing units (GPUs) for reasonable training time.

Conventional Multimodal Learning

Features are manually designed and require prior knowledge about the underlying problem and data.

Some techniques, like early fusion, may be sensitive to data preprocessing.

Feature selection and dimensionality reduction are often explicitly performed.

Typically performs early or late fusion.

Early fusion (data-level fusion) can be challenging and not scalable: late-fusion rules may need to be defined.

Rigid fusion architecture usually handcrafted.

May not require as much training data.

May not have as many hyperparameters as deep-learning architectures.

GPUs may offer speed-up, but not critical.

the question remains: at which depth of representation would the fusion be optimal?

The second architectural design choice for deep multimodal learning concerns which modalities to fuse. The underlying assumption in multimodal fusion is that different modalities provide complementary information toward solving the task at hand. However, it could be the case that the inclusion of all available modalities ends up being detrimental to the performance of the machine-learning algorithm—and, as such, some form of feature selection may be required. In the section “Fusion Structure Learning and Optimization,” we discuss a number of techniques that, during training, can automatically learn the optimal ordering and depth of fusion.

The third design choice involves dealing with missing data or modalities. Deep multimodal learning models should be robust enough to compensate for missing data or modalities during inference. Generative models are typically used in such instances.

Most deep multimodal learning approaches also involve representation learning from raw data. It is often the case that a deep multimodal architecture utilizes several standard modules or “building blocks” that are optimized for a specific kind of data. The choice of which deep-learning module is best for extracting pertinent information for a given modality is also an important architectural design choice. For example, when two-dimensional (2-D) pixel-based data are considered, convolutional architectures are often preferred. Three-dimensional (3-D) convolutional networks can be used for volumetric data, like CT, MRI, or even video. When temporal data are used, variants of recurrent neural networks (RNNs) such as long short-term memory (LSTM) or gated recurrent units can be incorporated.

The choice of modality-wise deep-learning architecture is mainly dependent upon the dimensionality of the input or whether temporal trends need to be learned. Beyond these common architectural choices, it is up to the reader to decide, given application-specific requirements that may involve properties of the data set or even the hardware used for training or deployment.

Applications

This section aims to provide an overview of the various application domains where deep multimodal learning has garnered much interest. Although multimodal learning and fusion is a widely researched topic, deep multimodal learning only started to gain attention following the works of Ngiam et al. [6] and Srivastava and Salakhutdinov [7]. These early works on deep multimodal fusion involved only two modalities: images and text. Ngiam et al. [6] investigated several approaches for multimodal fusion that include simple concatenation of inputs and shared representation learning, as well as cross-modality learning (where data from all modalities are present during training, but only a single modality is available during test). At around the same time, Srivastava and Salakhutdinov [7] also demonstrated the utility of fusing higher-level representations of disparate modalities involving images and text in a deep-learning framework. A notable finding was that constructing a multimodal fusion layer by way of simple concatenation of incoming connections resulted in relatively poorer results—revealing that hidden units have strong connections to variables from individual modalities but few units that connect across modalities. They also found that capturing cross-modality correlations required at least one nonlinear stage to be successful since higher-level representations of individual modalities will be relatively “modality free” and therefore more amenable to fusion. These early explorations became the basis of a number of proceeding works in deep multimodal learning that investigated various regularization strategies (see the section “Multimodal Regularization”) to enforce constraints for learning intermodality relationships.

Human activity recognition

An important area of research that heavily utilizes multimodal data is human activity recognition. Under this large umbrella of research, there are numerous subfields of research that relate to some aspect of human understanding. Given that humans

Table 2. Multimodal learning data sets and public multimodal machine-learning challenges.

Data Set	Modalities	Problem	Reference	Year
UTD-MHAD	Depth and inertial sensor data	Human action recognition	Chen et al. [93]	2015
Chalearn looking at people	RGB-D, audio, skeletal pose	Human activity recognition	Escalera et al. [94]	2014
Berkeley MHAD	Multiviewpoint RGB-D and skeletal pose data	Human activity recognition	Oflin et al. [95]	2013
MHRI data set	Chest, top RGB-D, face, video, and audio	Human-robot interaction	Pablo et al. [96]	2016
H-MOG	Nine smartphone sensors and interaction data	Continuous authentication in smartphones	Sitová et al. [12]	2016
RECOLA	Audio, visual, and physiological	Emotion recognition	Ringeval et al. [97]	2013
MHEALTH	Accelerometer, electrocardiogram, magnetometer, and gyroscopes	Health monitoring	Banos et al. [98]	2015
Pinterest Multimodal	Images and text (40M)	Multimodal word embeddings	Mao et al. [99]	2016
MM-IMDb	Video, images, and text metadata	Movie genre prediction	Arevalo et al. [100]	2017
FCVID	Video and audio	Action recognition	Jiang et al. [101]	2017
KITTI	Stereo gray- and color video, 3-D-LIDAR, inertial and GPS navigation data	Autonomous driving	Geiger et al. [91]	2017
KinectFaceDB	RGB-D and facial landmarks	Face recognition	Min et al. [102]	2014
Oxford RobotCar	Six cameras, LIDAR, GPS, and inertial navigation data	Autonomous driving	Maddern et al. [92]	2016
Multimodal BRATS	T2-, FLAIR-, post-Gadolinium T1-MRI, perfusion, and diffusion MRI and MRSI	Brain tumor segmentation	Menze et al. [103]	2015

RGB-D: RGB-depth

exhibit highly complex behavior in social settings, it is only natural that multimodal data are required for machine-learning algorithms to classify, or “understand,” their human behavior. Not surprisingly, we found that many works in deep multimodal fusion reported in recent years have focused on multimedia data that typically involve modalities such as audio, video, depth, and skeletal motion information. Multimodal deep-learning methods have been applied to various problems involving human activity such as action recognition (an activity can be composed of two or more sequences of shorter actions), gesture recognition [8], gaze-direction estimation [9], face recognition [10], and emotion recognition [11]. The ubiquity of mobile smartphones, which have no fewer than ten sensors, has given rise to novel applications that involve multimodal data such as continuous biometric authentication [12] and activity recognition [13]. Related subfields of research include human pose estimation [14] and semantic scene understanding [15].

We expect that the deep-learning research community will continue to focus on these problems in the foreseeable future. This is evidenced not only by the number of multimodal deep-learning papers being published but also the increasing number of data sets and public challenges made available online (see Table 2).

Medical applications

Deep learning in medical applications has become an important application domain that has attracted substantial interest. Medical imaging, for example, consists of a multitude

of multimodal data in the form of different medical imaging modalities such as MRI, CT, positron emission tomography (PET), functional MRI (fMRI), X-ray, and ultrasound. Although there have been notable improvements in new medical imaging technologies, the interpretation of these modalities for diagnosis still requires highly trained human experts. Conventional computer vision approaches required manually designed morphological feature representations. However, transforming the tacit knowledge of human experts into a computational form is not trivial. In the medical imaging field, manually designing suitable image features is extremely challenging, as it not only involves the interpretation of subtle visual markers and abnormalities, often needing years of medical training, but also the need to fuse complementary as well as possibly conflicting information from multiple imaging modalities. Therefore, the ability to learn these multimodal features through examples, as seen in the success of deep learning applied to computer vision, has prompted researchers to investigate their applicability in the medical domain. It is therefore not surprising that an increasing amount of medical image analysis research in recent years [16], both unimodal as well as multimodal, involves deep-learning-based methods.

Multimodal deep learning has been used in problems involving tissue and organ segmentation [17], multimodal medical image retrieval [18], multimodal medical image registration [19], and computer-aided diagnosis [20]. A recent review article by Mamoshina et al. [21] demonstrates the

rising popularity of DNNs, including models that implement multimodal fusion in biomedical applications involving genomic, proteomic, and drug data.

Two major challenges in applying deep-learning-based approaches for medical applications are 1) the difficulty in obtaining sufficiently labeled data and 2) the problem of class imbalance (where the number of negative examples far outnumber the case of positive examples). To overcome the first challenge, early approaches resorted to patch-based training [22]. Recently, techniques that utilize transfer learning have been surprisingly successful [23]. This involves reusing a portion of the data-agnostic representations learned by very deep networks on a separate, large data set, for example, ImageNet, and then fine-tuning or retraining only the upper layers of the network using a much smaller medical data set. Another common technique is to perform training data augmentation, for example, applying different affine transformations or randomly perturbing the brightness and contrast of images to increase the amount of training data available. To address the data imbalance problem, it is common to apply some form of weighting to the loss function such that mistakes made on the majority classes are less penalized than mistakes the network makes on the minority class. These challenges, although very common to problems in the medical domain, may also occur in other domains—and the solutions, as such, are equally applicable. However, despite the success of deep learning in medical applications, the medical community is still rather apprehensive about deploying them in the real world, as deep learning is often seen as opaque. This view is likely to gradually change given the increasing efforts to design interpretable DNNs [24], [25].

Autonomous systems

Following the success of deep learning, there has been a surge of interest in autonomous navigation (also known as *autonomous driving*) applications, which typically involve multimodal data acquired from sensors mounted on the vehicle. A self-driving car, for example, could include a number of external and internal sensors including radar, stereoscopic visible-light cameras, LIDAR, infrared (IR) cameras, global positioning system (GPS), and audio. To perform autonomous navigation, the heterogeneous data collected from sensors are used for learning a number of interrelated but complex tasks such as localization and mapping, scene understanding, movement planning, and driver-state recognition.

One of the biggest challenges for autonomous navigation is the dynamic nature of the operating environment—the system has to adapt and be reactive to weather variations, lighting variability, pedestrians and other traffic, road conditions, and traffic signs, as well as the driver's state. Nevertheless, deep-learning and reinforcement-learning techniques [26] have been instrumental in advancing this field of application with industry players like Uber, Nvidia, Baidu, and Tesla actively involved in the development of commercial self-driving cars.

An important task in autonomous driving is real-time scene understanding. It requires the learning system to recognize objects in the scene, like lanes, traffic signs, pedestrians, and other traffic. It follows that, for each frame of the multimodal video feed, semantic segmentation has to first be

performed. Each semantic concept identified in the scene then has to be localized. For such tasks, deep fully convolutional architectures that perform pixel-wise labeling of each frame are often used [27]. For multimodal inputs, a common strategy is to concatenate synchronized frames across the channel dimensions before being input to a fully convolutional neural network (CNN) (this is, in a sense, early fusion) or, alternatively, to train separate modality-wise networks and then fuse at a deeper stage in the multimodal network. We further discuss such fusion strategies in the section “Fusion Structure.” Semantic segmentation can be extended to video by using a 3-D variant of fully CNNs. The basic techniques used in self-driving vehicle technology can be extended to other robotic applications such as mobile robots or drone navigation, grasp configuration learning [28], and robotic manipulation [29].

Summary

We have highlighted three major areas where deep multimodal learning approaches have gained a foothold and continue to experience rapid advancements. In addition to the work already highlighted with respect to each of these key areas, we list additional representative work involving deep multimodal learning in Table 3. Several other application areas that involve text, images, and video, e.g., visual question answering (VQA) and image and video annotations, are highlighted in subsequent sections where we discuss specific deep-learning models.

Models

Applying multimodal deep learning to a new problem involves the selection of both an architecture and a learning algorithm. Together, we will call these choices a *model*. A plethora of different deep-learning models have been proposed for multimodal data. While there are several ways that they could be partitioned and organized for review, we have opted to group notable examples according to their learning paradigm, specifically whether they are generative, discriminative, or hybrid models. Our reason for choosing this categorization is that this choice impacts the available architectures and learning algorithms from which to select.

Generative models implicitly or explicitly represent a data distribution, often allowing for new data to be sampled or “generated” through a process, hence their name. Discriminative models, on the other hand, are less ambitious. Rather than modeling distributions, they attempt to model class boundaries. In the supervised learning setup, where we have data X and labels Y , generative models learn the joint probability $P(X, Y)$. In contrast, discriminative models are used for primarily prediction tasks, and these models learn the conditional $P(Y|X)$. Yet generative models can still have discriminative properties. An advantage of generative models is that they are much more flexible. For example, $P(X, Y)$ can be sampled in the case of missing modalities during inference.

Discriminative models

Discriminative deep architectures directly model the mapping from inputs to outputs, and the model parameters are learned

Table 3. Diverse applications of multimodal deep learning.

Reference	Year	Modalities	Problem	Fusion Method	Model	Architecture
Ngiam et al. [6]	2011	Audio, video	Speech classification	Intermediate	Generative	Sparse RBM
Srivastava and Salakhutdinov [30]	2012	Image, text	Image annotation	Intermediate	Generative	DBN
Cao et al. [31]	2014	Medical images, textual descriptions	Content-based medical image retrieval	Intermediate	Generative	DBM
Liang et al. [32]	2015	Gene expression, DNA methylation, and drug response	Cancer subtype clustering	Intermediate	Generative	DBM
Valada et al. [15]	2016	Multispectral imagery	Semantic segmentation	Early	Discriminative	FCNN
Simonyan and Zisserman [33]	2014	Image and optical flow	Action recognition	Late	Discriminative	CNN
Kahou et al. [11]	2015	Video, audio	Emotion recognition	Late	Discriminative	CNN, RNN, SVM, and AE
Liu et al. [20]	2015	MRI, PET	Medical diagnosis	Intermediate	Discriminative	Stacked AE, SVM
Poria et al. [34]	2015	Video, audio, text	Sentiment analysis	Intermediate, late	Discriminative	CNN, SVM
Lenz et al. [28]	2015	Intensity, depth video	Robotic grasping	Intermediate	Discriminative	Stacked AE and MLP
Jain et al. [35]	2016	Video features, GPS coordinates, vehicle dynamics	Driver activity anticipation	Intermediate	Discriminative	LSTM

by minimizing some regularized loss function. Such models compose the bulk of the proposed models for multimodal learning, while tasks include classification or recognition for a variety of problem domains.

In addition to the aforementioned active research problems, image captioning and VQA [36], both of which combine natural language processing and high-level scene interpretation by machine-learning algorithms, have garnered active research interest. In deep image captioning, the model is required to generate a textual description of image content, and this could be achieved by using both discriminative techniques [37], [38] and generative approaches [39]. On the other hand, VQA typically requires the model to answer complex questions based on image content, which is a generative task. This problem can also be cast into a discriminative setting (e.g., multiple choice questions) [40]. Recently, Kim et al. [41] extended the highly successful deep residual network model for a multimodal VQA problem. As multiple modalities may have correlations, the authors carefully designed joint residual mappings across modalities and achieved state-of-the-art results for VQA.

Discriminative deep multimodal learning models have also been proposed for human activity recognition. With the cheap availability of RGB-depth (RGB-D) cameras and ubiquity of smartphones with numerous sensors, deep multimodal learning architectures that involve from four to five modalities have been reported. These problems involve temporal data (video, joint motion, audio), and it is essential that spatiotemporal dependencies be learned effectively. To capture temporal structures and relationships, deep multimodal learning approaches typically use temporal components such as LSTMs or hidden Markov models combined with visual rep-

resentation learning layers like CNNs or 3-D-CNNs [42], [43]. These models have benefited from the combination of CNNs and recurrent layers that can collectively capture spatiotemporal relationships.

There are also instances of work where generative models have been adapted to perform discriminative tasks. For example, a discriminative variant of the RBM [building block of deep belief networks (DBNs) and deep Boltzmann machines] was proposed by Larochelle and Bengio [44]. Other discriminative models have previously been mentioned while discussing application areas in the sections “Human Activity Recognition,” “Medical Applications,” and “Autonomous Systems.” In addition, Table 3 also lists examples of discriminative models for other multimodal problems. While discriminative models excel at the task of classification or regression, they cannot cope when there are missing data or modalities. Discriminative models also require a large set of labeled data, which could be expensive to obtain in certain applications. Next, we review deep generative multimodal models, which offer some advantages, considering the drawbacks of discriminative models, in the context of learning multimodal representations.

Generative models

Deep generative models typically characterize the high-order correlation properties of the observed or visible data for pattern analysis or synthesis purposes. They can also be used to characterize the joint statistical distributions of the visible data and their associated classes. Generative models like DBNs can also be used for classification and regression tasks by exploiting their capability to learn (unsupervised) from unlabeled data and fine-tuned in a discriminative setting using the

backpropagation algorithm or by using the learned representation in conjunction with other classifiers such as support vector machines (SVMs).

For multimodal learning problems, generative models are useful in situations where there could be missing modalities during test time or when there is a lack of labeled data. The early works of Ngiam et al. [6] and Srivastava and Salakhutdinov [30] proved that generative models are indeed capable of handling such learning problems. Since then, a number of works have been reported in the literature that specifically deal with using generative deep multimodal networks in cases where there are missing data [31], [45].

While energy-based models based on stacking RBMs have received most of the attention in deep generative multimodal learning, the landscape of generative models is changing. Recently, generative adversarial networks [46], deep directed models trained with variational inference [47], are gaining traction in multi- and unimodal settings [48]–[50].

Hybrid models

While discriminative models are trained to maximize the separation between classes, generative models excel at modeling data distributions. Hybrid models combine both discriminative and generative components in a unified framework. Deng [51] defines hybrid deep architectures as architectures where the goal is discrimination but is assisted (often in a significant way) with the outcomes of generative architectures. For example, the generative component in a hybrid model may learn a deep representation of input modalities and use the discriminative component for classification or regression tasks.

Hybrid models can be divided into three groups as per [52]:

- 1) *joint methods* that optimize a single objective function to learn a joint representation using the generative and discriminative components
- 2) *iterative methods* that learn a shared representation using an iterative method such as expectation maximization using representations updated from both generative and discriminative components
- 3) *staged methods*, where the generative and discriminative components are trained separately in stages.

Representations learned by the generative model in an unsupervised manner can then be used as features for the discriminative component using supervised training.

An example of a joint model is reported in [53], where short-term temporal characteristics and long-range temporal dependencies for audio-video modalities are modeled by combining a conditional RBM temporal generative model for the former and a discriminative component consisting of a conditional random field for the latter. This model also is able to handle missing modalities due to the generative component. Other related hybrid architectures include those of Sachan et al. [54] and Liu et al. [55].

Summary

In this section, we have highlighted multimodal architectures according to their primary learning paradigm. In some sense,

deep-learning models can be thought of as building blocks that allow us to “mix and match” different models to create elaborate deep multimodal architectures. While this can be seen as an advantage of deep learning, a common issue is that architecture design has been more an art than a science. Notwithstanding, there are numerous hyperparameters associated with each model that have to be carefully fine-tuned, and this process may be possibly even more complicated when dealing with hybrid architectures. Another aspect to be concerned with is the choice of the fusion structure between modalities and their representations. Next, we discuss several choices for multimodal fusion structure and direct our discussion to the attractive notion of optimizing and learning this fusion architecture for improved performance.

Fusion structure

Deep architectures offer the flexibility of implementing multimodal fusion either as early, intermediate, or late fusion. Multimodal fusion approaches predating the advent of deep learning often referred to early fusion as *feature-level fusion* and late fusion as *decision-level fusion*. With deep-learning approaches, however, the idea of feature-level fusion can be extended further to the concept of intermediate fusion.

Early fusion

Early fusion involves the integration of multiple sources of data, at times very disparate, into a single feature vector, before being used as input to a machine-learning algorithm as illustrated in Figure 2(a). The data to be fused are the raw or pre-processed data from the sensor; hence, the terms *data fusion* or *multisensor fusion* are often used.

If data fusion is performed without feature extraction, this could be quite challenging. For instance, the sampling rate between different sensors could vary, or synchronized data from multiple data sources might not be available if one source produces discrete data, while another source provides a continuous data stream.

To alleviate some of the issues related to fusing raw data, higher-level representations can first be extracted from each modality, which could be either handcrafted features or learned representations, as is common in deep learning, before fusing at the feature level. When nonhierarchical features are used, as often the case in handcrafted features, features extracted from multiple modalities can be fused at only one level, prior to being input to the machine-learning algorithm. Since deep learning essentially involves learning hierarchical representations from raw data, this gives rise to intermediate-level fusion.

Most early-fusion models make the simplifying assumption that there is conditional independence between the states of various sources of information. This may not be true in practice, as multiple modalities tend to be highly correlated (for example, video and depth cues). Sebe [56] argues that different streams contain information that is correlated to another stream only at a high level. An excellent example of this can be seen in [57]. This assumption allows the output of each modality to be processed independently of the others.

In its simplest form, early fusion involves concatenation of multimodal features as was implemented by Poria et al. [34]. Early fusion of multimodal data may not fully exploit the complementary nature of the modalities involved and may lead to very large input vectors that may contain redundancies. Typically, dimensionality reduction techniques like PCA are applied to remove these redundancies in the input space. Autoencoders, which are nonlinear generalizations of PCA [58], are popularly used in deep learning to extract a distributed representation from raw data. This idea has been extended to learn a multimodal embedded space with the aim to represent multimodal data within a common feature space [59], [60].

One of the issues faced in early fusion of multimodal data is to determine the time-synchronicity between different data sources. Commonly, these signals are resampled at a common sampling rate. To mitigate this issue, Martínez and Yanakakis [61] proposed several methods (convolution, training, and pooling fusion) to integrate sequences of discrete events with continuous signals.

Late fusion

Late- or decision-level fusion refers to the aggregation of decisions from multiple classifiers, each trained on separate modalities [see Figure 2(b)]. This fusion architecture is often favored because errors from multiple classifiers tend to be uncorrelated and the method is feature independent. Various rules exist to determine how decisions from different classifiers are combined.

These fusion rules could be max-fusion, averaged-fusion, Bayes' rule based, or even learned using a metaclassifier. Decision-level fusion was popular in the early- to mid-2000s, when ensemble classifiers received widespread interest within the machine-learning community.

There have been several works that employ late- or decision-level fusion for deep multimodal learning [33], [43], [62] in addition to some works listed in Table 3. Based on the papers that we have reviewed, we do not find conclusive evidence that late fusion is better than early fusion—the performance is very much problem dependent. Undoubtedly, when

input modalities are significantly uncorrelated, of very different dimensionality and sampling rates, it is much simpler to implement a late-fusion approach for multimodal learning problems. An alternative approach, intermediate fusion, offers much more flexibility as to how and when representations learned from multimodal data can be fused.

Intermediate fusion

Neural networks transform raw inputs to higher-level representations by mapping the input through a pipeline of layers. Each layer typically alternates linear and nonlinear operations that scale, shift, and skew its input, producing a new representation of the original data. In the multimodal context, when all of the modalities are transformed into representations, then it becomes amenable to fuse different representations into a single hidden layer and then learn a joint multimodal representation. The majority of work in deep multimodal fusion adopts this intermediate-fusion approach, where a shared representation layer is constructed by merging units with connections coming into this layer from multiple modality-specific paths. Figure 2(c) illustrates a simple intermediate fusion model with three modalities. Representations (features) are learned using different kinds of layers (e.g., 2-D-convolution, 3-D-convolution, or fully connected), and representations are fused using a fusion layer, also known as a *shared representation layer*.

This shared representation layer can be a single shared layer that fuses multiple channels at some depth or could be gradually fused, one or more modalities at a time. A naïve concatenation of features or weights in the shared representation layer may lead to overfitting or the network failing to learn associations between modalities due to distinct underlying distributions. A simple method of improving performance of multimodal fusion is to apply some form of dimensionality reduction like PCA [63] or stacked autoencoders [10] after constructing a shared representation layer (or fusion layer) via simple concatenation of weights from different modalities. This choice of fusing various representations at different depths is perhaps the most powerful and flexible aspect of deep multimodal fusion as opposed to other fusion techniques. The advantage of a flexible fusion scheme can be seen in the

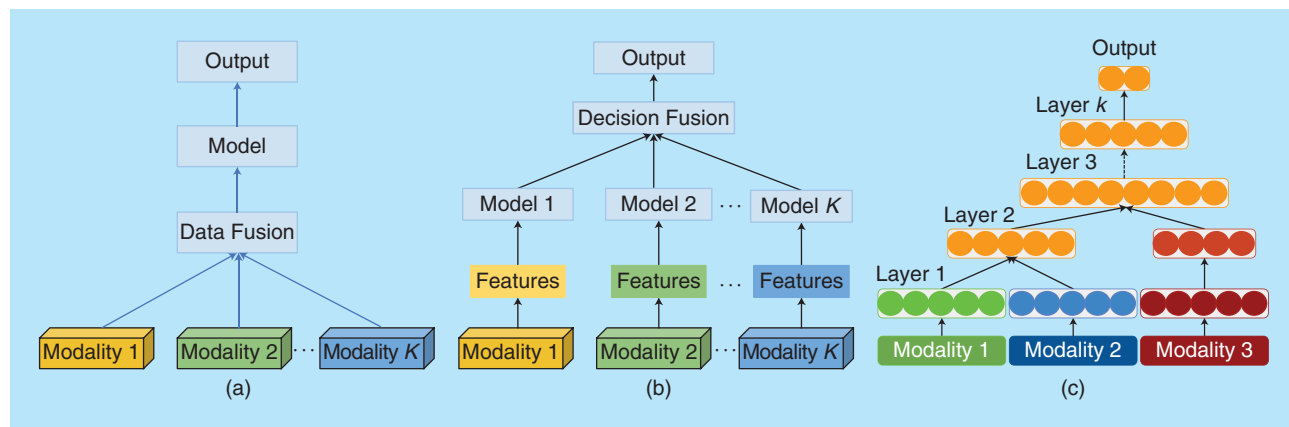


FIGURE 2. An illustration of various fusion models for multimodal learning. (a) Early or data-level fusion, (b) late or decision-level fusion, and (c) intermediate fusion.

work of Karpathy et al. [64], who showed that using a “slow-fusion” model, where learned representations of video streams are gradually fused across multiple fusion layers during training, consistently produced better results for a large-scale video classification problem, as opposed to early-fusion and late-fusion models. Similarly, Neverova et al. [8] empirically showed that implementing a gradual fusion strategy, by first fusing highly correlated modalities, to less correlated ones in a progressive manner (e.g., visual modalities first, then motion capture, then audio), produced state-of-art results for communicative gesture recognition.

Although learning a multimodal representation using the shared representation layer is indeed flexible, many current architectures require careful design in terms of how, when, and which modalities can be fused. In the “Fusion Structure Learning and Optimization” section, we discuss further attempts at optimizing the tedious architecture design process required by multimodal learning.

Multimodal regularization

Deep-learning techniques iteratively optimize a set of model parameters (typically, the weights and biases between each layer) by minimizing a loss function. To improve generalization, one or more regularization strategies are employed, often as an additional term added to the loss function. From a computational perspective, regularization provides stability to the optimization problem leading to algorithmic speed-ups, and, from a statistical point of view, regularization reduces overfitting [65].

In the deep multimodal learning context, an important design consideration is the formulation of cost functions and regularizers that enforce intermodality and intramodality relationships such as information-theoretic regularizers and structured regularization, which we now briefly review.

Information-theoretic regularizers are formulated using measures such as mutual information and variation of information. For example, Sohn et al. [66] proposed a cost function that minimized the variation of information between modalities to learn relationships between modalities. The intuition behind this formulation is that learning to maximize the amount of information that one data modality has about the others would allow generative models to reason about the missing data modality given partial observations. Alternatively, a mutual information term could also be maximized during training [67]. Another information theoretic loss formulation based on the Kulback–Leibler (KL) divergence was proposed by Zhu et al. [68] for a multilabel image annotation problem. In a pretraining stage, they first trained CNNs to learn intermediate representations from each modality using unlabeled data and then, in the fine-tuning stage, used backpropagation to minimize the KL-divergence between the predictive and ground-truth distributions. Finally, to learn the optimal combination of multimodal weights, they adopted the exponentiated online learning algorithm to sequentially find an optimal set of combinational weights.

Taking inspiration from structured feature selection in multitask learning [69], Wu et al. [70] designed a model that

uses a trace norm regularization term, which encourages similar modalities to share similar representations for a video classification problem using video and audio modalities.

Cost functions that enforce inter- and intramodality correlations have also been explored by Wang et al. [71]. Their formulation includes a discriminative term, and a correlative term based on canonical correlation analysis. In a follow-up work [72], they proposed a multimodal fusion layer that uses matrix transformations to explicitly enforce a common part to be shared by features of different modalities while retaining modality-specific learning.

Lenz et al. [28] formulated a structured regularization term in the cost function, which allows a model to learn correlated features between multiple input modalities but regularizes the number of modalities used per feature, thereby discouraging the model from learning weak correlations between modalities. Structured regularization essentially applies some form of regularization separately for each set of modality-specific weights. They considered several variants of structure regularization for a multimodal robotic grasping task. One that worked well in their case incorporated the L_0 norm on top of the max-norm penalty

$$f(W) = \sum_{j=1}^K \sum_{r=1}^R \mathbb{I}\left\{\left(\max_i S_{r,i} | W_{i,j}\right) > 0\right\},$$

where $S_{r,i}$ is one if feature i belongs to group r and is otherwise zero. S is a binary modality matrix of size $R \times N$, where each element $S_{r,i}$ indicates the membership of a visible unit, x_i , in a particular modality, r . \mathbb{I} is an indicator function, which takes a value of one if its argument is true and is otherwise zero.

In some problems, temporal context can play an important role, for example, driver activity anticipation. Unlike human activity recognition, where complete temporal context is available, in driver activity anticipation, the machine-learning system must predict using only partial context within a short span of time before the event occurs. To solve this problem, Jain et al. [35] incorporated a temporal term that grows exponentially in time into their cost function for a multimodal RNN with LSTM units. This encourages the model to fix mistakes as early as it can.

Multimodal-aware regularizers have resulted in marginal to notable improvements in model performance. Despite including these multimodal regularization strategies, the deep-learning architectures discussed in this section have input modalities merging into a single fusion layer. A possible extension could be to investigate a gradual fusion model that takes advantage of these regularization strategies.

Fusion structure learning and optimization

Most multimodal deep-learning architectures proposed to date are meticulously handcrafted. While many models adopt a single fusion layer (shared representation layer), several stand-out works [8], [64] implemented a gradual fusion strategy. The choice of which modality is fused, and at which

depth of representation, is usually based on intuition (for example, fusing similar modalities early, and then fusing disparate modalities at a deeper layer). When more than two modalities are involved, also depending upon the nature of the modalities being used in the problem, choosing an optimal fusion architecture may be more challenging. A natural progression would be to search for an optimal multimodal fusion architecture by casting this as a model search or structure learning problem.

Neural network structure optimization for unimodal problems has long been investigated by machine-learning researchers. These mainly involved determining the optimal number of neurons and number of layers in a network. There is a tradeoff between good generalization ability of the network, and the number of parameters and availability of training data. Too large a network might perform well or overfit, depending if it is trained with sufficiently large training data, while too small a network, might underfit and may result in poor generalization.

A common approach is to adopt a bottom-up constructive approach. The basic idea proposed by Elman [73] is to start with a relatively small network and add hidden units or layers incrementally until the best performing architecture is found. More recently, and in the large-scale setting, Chen et al. [74] gradually added depth and width to an inception-style [75] network by knowledge transfer between one neural network to another.

Pruning algorithms [76] address the same problem from a top-down approach. Recent approaches for DNNs include the works of Feng and Darrel [77], who proposed an evolving grow-and-prune algorithm that optimizes the structure of an Indian buffet process-CNN model, and Yang et al. [78], who introduced network pruning for large, diverse data sets based on sparse representations.

Genetic algorithm (GA)-based structure optimization of neural networks was one of the earliest metaheuristic search algorithms used for neural network structure search and optimization [79]. In the early 2000s, an algorithm called *Neuro Evolution of Augmenting Topologies (NEAT)* [80] that also used GAs to evolve increasingly complex neural network architectures received much attention. More recently, Shinzaki and Watanabe [81] applied GAs and a covariance matrix evolution strategy to optimize the structure of a DNN, parameterizing the structure of the DNN as a simple binary vector based on a directed acyclic graph representation. As the GA search space can be very large, and each model evaluation in the search space is expensive, a parallel search using a large GPU cluster was used to speed up the process.

These neural network structural search and optimization techniques can readily be extended to the multimodal setting if a suitable representation of the network architecture is devised and provided that the cost of training and testing multiple architectures during the search process is not prohibitively expensive. With data set sizes approaching gigabytes, and even terabyte levels, and deep network architectures involving millions of parameters and multiple modalities, search and optimization of multimodal fusion structure can

be prohibitively expensive unless some parallel search procedure is implemented or an efficient optimization algorithm is used. While Bayesian optimization (BO) [82] has been a popular choice for hyperparameter optimization, it has been recently used for multimodal fusion architecture optimization [83]. Architecture optimization was cast as a discrete optimization problem by searching a space of all possible multimodal fusion architectures using a Gaussian process-based BO. A novel graph-induced kernel was proposed to quantify the distance between different architectures in the search space.

Reinforcement learning [84] has also been used for deep neural architecture search [85]. This work proposed a novel method of using an RNN to generate variable-length model descriptions of neural networks. The RNN was trained with reinforcement learning to maximize the expected accuracy of the generated architectures on a validation set.

A number of recent works have approached structure learning as a means of regularization, or capacity control, in a network. By pruning the network in a stochastic manner, stochastic regularization methods can be considered as a kind of ensemble that improves generalization via model averaging. Kulkarni et al. [86] implemented a method of learning the structure of DNNs via deterministic regularization. They insert, between each pair of fully connected layers, a sparse diagonal matrix whose entries are l_1 penalized. This implicitly defines the size of the effective weight matrices at each layer. The approach has a similar effect to Dropout [87]. Blockout [88] can perform simultaneous regularization and model selection through a clever technique that stochastically assigns hidden units to “clusters,” forming block-structured weight matrices. In addition, by averaging the outputs of multiple stochastic inference passes (which can be viewed as a case of ensemble classifiers), results better than ResNets were achieved. This architecture effectively implements a late fusion of multiple architectures to achieve better results.

Stochastic regularization has been extended to the multimodal setting by Neverova et al. [8] and, more recently, by Li et al. [89]. In the latter work, the authors show that, when the intermodality correlation is high, an early-fusion approach (whose fusion structure was learned by the network) produced better results, while a late-fusion approach worked better when the input modalities are less correlated. This concurred with the empirical choice made by the former.

In this section, we have covered a number of recent works that use either stochastic regularization or optimization resulting in deep multimodal fusion architectures that perform at par with or better than meticulously designed ones. While feature engineering has been largely solved by deep representation learning, the next logical step would be to do away with meticulous engineering of deep architectures and pursue techniques that achieve this automatically.

Data sets

To facilitate research in multimodal learning, a number of data sets have been released to the public. We note that the majority of

these data sets typically involve person-centric visual understanding, with variants including emotion recognition, group behavior analysis, etc. Table 2 lists a number of such data sets, the modalities involved, and the problem domain. While this list is not exhaustive, we cover more recent data sets (many of which were released in the past three years) that are available for multimodal research. While most data sets include at least two modalities (images and text, for example) or up to four (RGB-D, audio, and skeletal pose), some data sets, for example, H-MOG [12], include up to nine different modalities. For the interested reader, Firman [90] presents an extensive survey of 102 RGB-D data sets. Autonomous driving and driver assistance systems (using driver behavior prediction) are being pursued as a popular research topic in deep learning. Such data sets are not only highly multimodal [91], with data from up to six individual sensors, but also very large—hours of data available. The Oxford RobotCar [92] data set, for example, contains more than 23 TB of year-long driving data in various weather conditions.

We note that there are relatively fewer multimodal medical data sets available, possibly due to the cost and ethical and privacy concerns. Most medical data sets also tend to be much smaller, involving between ten and 50 subjects and also suffer from high class imbalances (for example, it is much more common to have normal cases in comparison to abnormal cases). Medical informatics and imaging studies rely heavily on multimodal information, and this can be leveraged to improve computer-aided diagnosis. Efforts to gather and make such data sets publicly available are encouraged.

Conclusions and future directions

In this article, we have reviewed recent advancements in deep multimodal learning. It is undeniable that the incorporation of multiple modalities into the learning problem almost always results in much better performance for a wide range of problems. From a fusion perspective, we see that techniques in deep multimodal learning can be classified into early- and late-fusion approaches and that deep-learning methods facilitate a flexible intermediate-fusion approach, which not only makes it simpler to fuse modality-wise representations and learn a joint representation but also allows multimodal fusion at various depths in the architecture. Although deep learning has, in many cases, reduced the need for feature engineering, deep-learning architectures still involve a great deal of manual design, and experimenters may not have explored the full space of possible fusion architectures. It is only natural that researchers should extend the notion of learning to architectures in an effort to have a truly generic learning method, which can be adapted, with minimal or no human intervention, to a specific task.

We reviewed several options for learning an optimal architecture. This includes stochastic regularization, casting architecture optimization as a hyperparameter optimization problem using, for example, BO, and incremental online reinforcement learning. This is, in our opinion, the most exciting area of research for deep multimodal learning. Architecture learning can be extremely compute-intensive, so researchers should take advantage of advances in hardware acceleration and distributed deep learning.

We have also identified several application domains that are gaining the most attention in deep multimodal learning. This includes RGB-D and data from the multitude of sensors on mobile phones that have been used for a range of problems involving multimodal data such as human activity recognition and their variants. We foresee that this area will gain more attention in the coming years for novel applications, which will profoundly impact our daily lives. Another important area highlighted is medical research, which involves numerous modalities of data, some of which are very difficult to interpret without human experts in the loop. With the medical community opening up to the rise of artificial intelligence-assisted diagnosis, we will see more significant progress being made in this domain. Finally, two more application areas that are gaining the attention of deep-learning researchers involve autonomous vehicles or robotics and multimedia applications, for example, video transcription, image captioning, etc. Novel applications like online chatbots that use multimodal inputs, like images, and text or recommender systems that utilize multimodal data may become widespread in the near future.

We conclude by acknowledging that this is very much a fast-evolving field, and, at the rate of the amount of new research being published, many new innovations in deep multimodal learning architectures are bound to be presented. We have tried not to provide specific suggestions to architecture design, as we found many problems require application-specific considerations. Regardless, we feel this is a timely publication as the directions of future research that we have highlighted, hopefully, can act as a guide toward a more organized effort into advancing the research field.

Authors

Dhanesh Ramachandram (dramacha@uoguelph.ca) received his B.Tech degree in industrial technology and his Ph.D. degree in computer vision and robotics from the Universiti Sains Malaysia in 1997 and 2003, respectively, where he was formerly an associate professor. He is a researcher at the University of Guelph, Ontario, Canada, and a Senior Member of the IEEE. He is interested in deep learning for computer vision, medical imaging, and multimodal problems.

Graham W. Taylor (gwtaylor@uoguelph.ca) received his bachelor's and master's degrees in applied science from the University of Waterloo, Canada, in 2003 and 2004, respectively. He received his Ph.D. degree in computer science from the University of Toronto, Canada, in 2009, where his thesis coadvisors were Geoffrey Hinton and Sam Roweis. He is an associate professor at the University of Guelph, Ontario, Canada, a member of the Vector Institute for Artificial Intelligence, and a Canadian Institute for Advanced Research Azrieli Global Scholar. He is interested in statistical machine learning and biologically inspired computer vision, with an emphasis on unsupervised learning and time-series analysis.

References

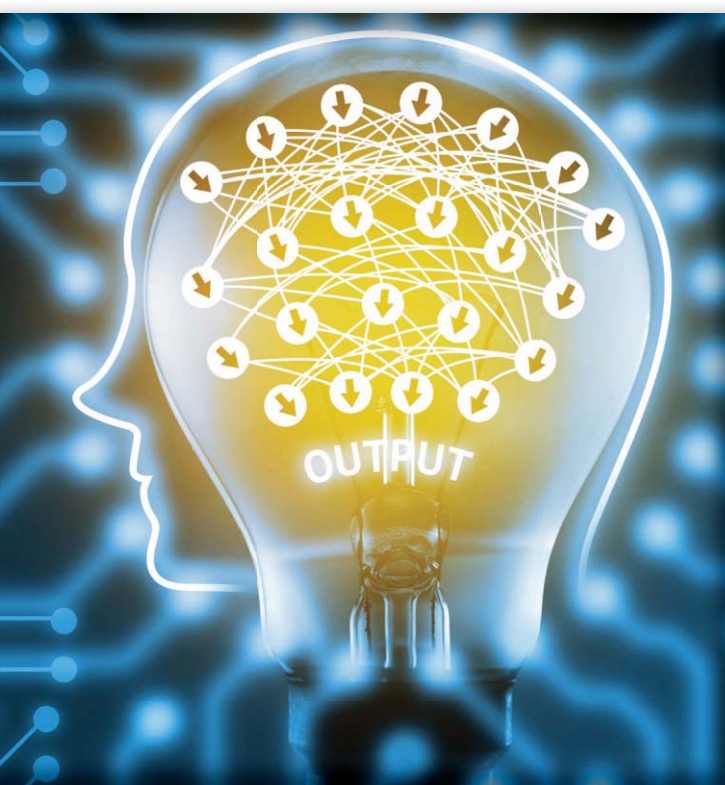
- [1] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

- [2] D. Lahat, T. Adali, and C. Jutten, "Multimodal data fusion: An overview of methods, challenges, and prospects," *Proc. IEEE*, vol. 103, no. 9, pp. 1449–1477, 2015.
- [3] P. K. Atrey, M. A. Hossain, A. El Saddik, and M. S. Kankanhalli, "Multimodal fusion for multimedia analysis: A survey," *Multimedia Systems*, vol. 16, no. 6, pp. 345–379, 2010.
- [4] B. Khaleghi, A. Khamis, F. O. Karray, and S. N. Razavi, "Multisensor data fusion: A review of the state-of-the-art," *Inform. Fusion*, vol. 14, no. 1, pp. 28–44, 2013.
- [5] L. I. Kuncheva, *Combining Pattern Classifiers: Methods and Algorithms*. Hoboken, NJ: Wiley, 2004.
- [6] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Y. Ng, "Multimodal deep learning," in *Proc. 28th Int. Conf. Machine Learning (ICML-11)*, 2011, pp. 689–696.
- [7] N. Srivastava and R. R. Salakhutdinov, "Multimodal learning with deep Boltzmann machines," in *Proc. Advances in Neural Inform. Processing Syst.*, 2012, pp. 2222–2230.
- [8] N. Neverova, C. Wolf, G. Taylor, and F. Nebout, "ModDrop: Adaptive multi-modal gesture recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 8, pp. 1692–1706, 2016.
- [9] S. S. Mukherjee and N. M. Robertson, "Deep head pose: Gaze-direction estimation in multimodal video," *IEEE Trans. Multimedia*, vol. 17, no. 11, pp. 2094–2107, 2015.
- [10] C. Ding and D. Tao, "Robust face recognition via multimodal deep face representation," *IEEE Trans. Multimedia*, vol. 17, no. 11, pp. 2049–2058, 2015.
- [11] S. E. Kahou, X. Bouthillier, P. Lamblin, C. Gulcehre, V. Michalski, K. Konda, S. Jean, P. Froumenty, et al., "EmoNets: Multimodal deep learning approaches for emotion recognition in video," *J. Multimedia User Interfaces*, vol. 10, no. 2, pp. 99–111, 2015.
- [12] Z. Sitová, J. Šeděnka, Q. Yang, G. Peng, G. Zhou, P. Gasti, and K. S. Balagani, "HMOG: New behavioral biometric features for continuous authentication of smartphone users," *IEEE Trans. Inf. Forensics Security*, vol. 11, no. 5, pp. 877–892, 2016.
- [13] V. Radu, N. D. Lane, S. Bhattacharya, C. Mascolo, M. K. Marina, and F. Kawars, "Toward multimodal deep learning for activity recognition on mobile devices," in *Proc. ACM Int. Joint Conf. Pervasive and Ubiquitous Computing: Adjunct*, 2016, pp. 185–188.
- [14] A. Jain, J. Tompson, Y. LeCun, and C. Bregler, "Modeep: A deep learning framework using motion features for human pose estimation," in *Proc. Asian Conf. Computer Vision*, 2014, pp. 302–315.
- [15] A. Valada, G. L. Oliveira, T. Brox, and W. Burgard, "Deep multispectral semantic scene understanding of forested environments using multimodal fusion," in *Proc. Int. Symp. Experimental Robotics (ISER 2016)*, 2016, pp. 465–477.
- [16] D. Shen, G. Wu, and H.-I. Suk, "Deep learning in medical image analysis," *Annu. Review Biomedical Eng.*, vol. 19, pp. 221–248, 2017.
- [17] R. Kiros, K. Popuri, D. Cobzas, and M. Jagersand, "Stacked multiscale feature learning for domain independent medical image segmentation," in *Proc. Int. Workshop on Mach. Learning in Medical Imaging*, 2014, pp. 25–32.
- [18] P. Wu, S. C. Hoi, H. Xia, P. Zhao, D. Wang, and C. Miao, "Online multimodal deep similarity learning with application to image retrieval," in *Proc. 21st ACM Int. Conf. Multimedia*, 2013, pp. 153–162.
- [19] M. Simonovsky, B. Gutiérrez-Becker, D. Mateus, N. Navab, and N. Komodakis, "A deep metric for multimodal registration," in *Proc. Int. Conf. Medical Image Computer and Computer-Assisted Intervention*, 2016, pp. 10–18.
- [20] S. Liu, S. Liu, W. Cai, H. Che, S. Pujol, R. Kikinis, D. Feng, M. J. Fulham, et al., "Multimodal neuroimaging feature learning for multiclass diagnosis of Alzheimer's disease," *IEEE Trans. Biomed. Eng.*, vol. 62, no. 4, pp. 1132–1140, 2015.
- [21] P. Mamoshina, A. Vieira, E. Putin, and A. Zhavoronkov, "Applications of deep learning in biomedicine," *Molecular Pharmaceutics*, vol. 13, no. 5, pp. 1445–1454, 2016.
- [22] Y. Guo, G. Wu, L. A. Commander, S. Szary, V. Jewells, W. Lin, and D. Shen, "Segmenting hippocampus from infant brains by sparse patch matching with deep-learned features," in *Proc. Int. Conf. Medical Image Computer and Computer-Assisted Intervention*, 2014, p. 308.
- [23] N. Tajbakhsh, J. Y. Shin, S. R. Gurudu, R. T. Hurst, C. B. Kendall, M. B. Gotway, and J. Liang, "Convolutional neural networks for medical image analysis: Full training or fine tuning?" *IEEE Trans. Med. Imag.*, vol. 35, no. 5, pp. 1299–1312, 2016.
- [24] I. Sturm, S. Lapuschkin, W. Samek, and K.-R. Müller, "Interpretable deep neural networks for single-trial eeg classification," *J. Neuroscience Methods*, vol. 274, pp. 141–145, Dec. 2016.
- [25] S. G. Kim, N. Theera-Ampornpunt, C.-H. Fang, M. Harwani, A. Grama, and S. Chaterji, "Opening up the blackbox: an interpretable deep neural network-based classifier for cell-type specific enhancer predictions," *BMC Syst. Biology*, vol. 10, no. 2, p. 54, 2016.
- [26] C. Chen, A. Seff, A. Kornhauser, and J. Xiao, "Deepdriving: Learning affordance for direct perception in autonomous driving," in *Proc. IEEE Int. Conf. Computer Vision*, 2015, pp. 2722–2730.
- [27] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2015, pp. 3431–3440.
- [28] I. Lenz, H. Lee, and A. Saxena, "Deep learning for detecting robotic grasps," *Int. J. Robotics Res.*, vol. 34, no. 4-5, pp. 705–724, 2015.
- [29] S. Gu, E. Holly, T. Lillicrap, and S. Levine, (2016). Deep reinforcement learning for robotic manipulation with asynchronous off-policy updates. *arXiv*. [Online]. Available: <https://arxiv.org/abs/1610.00633>
- [30] N. Srivastava and R. Salakhutdinov, "Learning representations for multimodal data with deep belief nets," presented at *Proc. 29th Int. Conf. Machine Learning (Workshop)*, 2012.
- [31] Y. Cao, S. Steffey, J. He, D. Xiao, C. Tao, P. Chen, and H. Müller, "Medical image retrieval: A multimodal approach," *Cancer Informatics*, vol. 13, no. Suppl 3, p. 125, 2014.
- [32] M. Liang, Z. Li, T. Chen, and J. Zeng, "Integrative data analysis of multi-platform cancer data with a multimodal deep learning approach," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 12, no. 4, pp. 928–937, 2015.
- [33] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," in *Proc. Advances in Neural Information Processing Systems*, 2014, pp. 568–576.
- [34] S. Poria, E. Cambria, and A. Gelbukh, "Deep convolutional neural network textual features and multiple kernel learning for utterance-level multimodal sentiment analysis," in *Proc. Conf. Empirical Methods on Natural Language Processing*, 2015, pp. 2539–2544.
- [35] A. Jain, A. Singh, H. S. Koppula, S. Soh, and A. Saxena, "Recurrent neural networks for driver activity anticipation via sensory-fusion architecture," in *Proc. 2016 IEEE Int. Conf. Robotics and Automation (ICRA)*, 2016, pp. 3118–3125.
- [36] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. L. Zitnick, and D. Parikh, "VQA: Visual question answering," in *Proc. Int. Conf. Computer Vision (ICCV)*, 2015, pp. 2425–2433.
- [37] J. Donahue, L. Anne Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell, "Long-term recurrent convolutional networks for visual recognition and description," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2015, pp. 2625–2634.
- [38] A. Karpathy, A. Joulin, and F. F. Li, "Deep fragment embeddings for bidirectional image sentence mapping," in *Proc. Advances in Neural Information Processing Systems*, 2014, pp. 1889–1897.
- [39] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: A neural image caption generator," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2015, pp. 3156–3164.
- [40] M. Ren, R. Kiros, and R. Zemel, "Exploring models and data for image question answering," in *Proc. Advances in Neural Information Processing Systems*, 2015, pp. 2953–2961.
- [41] J.-H. Kim, S.-W. Lee, D.-H. Kwak, M.-O. Heo, J. Kim, J.-W. Ha, and B.-T. Zhang, (2016). Multimodal residual learning for visual QA. *arXiv*. [Online]. Available: <https://arxiv.org/abs/1606.01455>
- [42] F. J. Ordóñez and D. Roggen, "Deep convolutional and LSTM recurrent neural networks for multimodal wearable activity recognition," *Sensors*, vol. 16, no. 1, p. 115, 2016.
- [43] D. Wu, L. Pigou, P.-J. Kindermans, N. D.-H. Le, L. Shao, J. Dambre, and J.-M. Odobez, "Deep dynamic neural networks for multimodal gesture segmentation and recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 8, pp. 1583–1597, 2016.
- [44] H. Larochelle and Y. Bengio, "Classification using discriminative restricted Boltzmann machines," in *Proc. 25th Int. Conf. Machine Learning*, 2008, pp. 536–543.
- [45] Y. Huang, W. Wang, and L. Wang, "Unconstrained multimodal multi-label learning," *IEEE Trans. Multimedia*, vol. 17, no. 11, pp. 1923–1935, 2015.
- [46] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Proc. Advances in Neural Information Processing Systems*, 2014, pp. 2672–2680.
- [47] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," in *Proc. Int. Conf. Learning Representations (ICLR)*, 2014.
- [48] S. Reed, Z. Akata, X. Yan, L. Logeswaran, B. Schiele, and H. Lee, "Generative adversarial text to image synthesis," in *Proc. 33rd Int. Conf. Machine Learning (ICML)*, 2016, pp. 1060–1069.
- [49] M. Suzuki, K. Nakayama, and Y. Matsuo, (2016). Joint multimodal learning with deep generative models. *arXiv*. [Online]. Available: <https://arxiv.org/abs/1611.01891>
- [50] G. Pandey and A. Dukkupati, (2016). Variational methods for conditional multimodal deep learning. *arXiv*. [Online]. Available: <https://arxiv.org/abs/1603.01801>
- [51] L. Deng, "A tutorial survey of architectures, algorithms, and applications for deep learning," *APSIPA Trans. Signal and Inform. Processing*, vol. 3, pp. 1–29, 2014.
- [52] M. R. Amer, T. Shields, B. Siddiquie, A. Tamrakar, A. Divakaran, and S. Chai, "Deep multimodal fusion: A hybrid approach," *Int. J. Comput. Vision*, pp. 1–17, 2017. DOI: 10.1007/s11263-017-0997-7.

- [53] M. R. Amer, B. Siddique, S. Khan, A. Divakaran, and H. Sawhney, "Multimodal fusion using dynamic hybrid models," in *Proc. IEEE 2014 Applications of Computer Vision Winter Conf.*, 2014, pp. 556–563.
- [54] D. S. Sachan, U. Tekwani, and A. Sethi, "Sports video classification from multi-modal information using deep neural networks," in *Proc. 2013 Association for the Advancement of Artificial Intelligence Fall Symp.*, 2013, pp. 102–107.
- [55] Y. Liu, X. Feng, and Z. Zhou, "Multimodal video classification with stacked contractive autoencoders," *Signal Processing*, vol. 120, pp. 761–766, Mar. 2016.
- [56] N. Sebe, *Machine Learning in Computer Vision*, vol. 29. Dordrecht, The Netherlands: Springer, 2005.
- [57] A. Owens, J. Wu, J. H. McDermott, W. T. Freeman, and A. Torralba, *Ambient Sound Provides Supervision for Visual Learning*. Cham, Switzerland: Springer International Publishing, 2016, pp. 801–816.
- [58] G. E. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *Science*, vol. 313, no. 5786, pp. 504–507, 2006.
- [59] D. Wang, P. Cui, M. Ou, and W. Zhu, "Learning compact hash codes for multi-modal representations using orthogonal deep structure," *IEEE Trans. Multimedia*, vol. 17, no. 9, pp. 1404–1416, 2015.
- [60] J. Masci, M. M. Bronstein, A. M. Bronstein, and J. Schmidhuber, "Multimodal similarity-preserving hashing," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 4, pp. 824–830, 2014.
- [61] H. P. Martínez and G. N. Yannakakis, "Deep multimodal fusion," in *Proc. 16th Int. Conf. Multimodal Interaction*, 2014, pp. 34–41.
- [62] S. E. Kahou, C. Pal, X. Bouthillier, P. Froumenty, Ç. Gülçehre, R. Memisevic, P. Vincent, A. Courville, et al., "Combining modality specific deep neural networks for emotion recognition in video," in *Proc. 15th ACM Int. Conf. Multimodal Interaction*, 2013, pp. 543–550.
- [63] D. Yi, Z. Lei, and S. Z. Li, "Shared representation learning for heterogeneous face recognition," in *Proc. Automatic Face and Gesture Recognition 11th IEEE Int. Conf. Workshops*, 2015, pp. 1–7.
- [64] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei, "Large-scale video classification with convolutional neural networks," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2014, pp. 1725–1732.
- [65] M. J. Wainwright, "Structured regularizers for high-dimensional problems: Statistical and computational issues," *Annu. Rev. Statistics Application*, vol. 1, pp. 233–253, Apr. 2014.
- [66] K. Sohn, W. Shang, and H. Lee, "Improved multimodal deep learning with variation of information," in *Proc. Advances in Neural Information Processing Systems*, 2014, pp. 2141–2149.
- [67] J. J.-Y. Wang, Y. Wang, S. Zhao, and X. Gao, "Maximum mutual information regularized classification," *Eng. Applicat. Artificial Intell.*, vol. 37, pp. 1–8, Jan. 2015.
- [68] S. Zhu, X. Li, and S. Shen, "Multimodal deep network learning-based image annotation," *IET Electron. Lett.*, vol. 51, no. 12, pp. 905–906, 2015.
- [69] H. Fei and J. Huan, "Structured feature selection and task relationship inference for multi-task learning," *Knowledge and Inform. Syst.*, vol. 35, no. 2, pp. 345–364, 2013.
- [70] Z. Wu, Y.-G. Jiang, J. Wang, J. Pu, and X. Xue, "Exploring inter-feature and inter-class relationships with deep neural networks for video classification," in *Proc. ACM Int. Conf. Multimedia*, 2014, pp. 167–176.
- [71] A. Wang, J. Lu, J. Cai, T. J. Cham, and G. Wang, "Large-margin multi-modal deep learning for RGB-D object recognition," *IEEE Trans. Multimedia*, vol. 17, no. 11, pp. 1887–1898, Nov. 2015.
- [72] A. Wang, J. Cai, J. Lu, and T.-J. Cham, "MMSS: Multi-modal sharable and specific feature learning for RGB-D object recognition," in *Proc. IEEE Int. Conf. Computer Vision*, 2015, pp. 1125–1133.
- [73] J. Elman, "Learning and development in neural networks: The importance of starting small," *Cognition*, vol. 48, no. 1, pp. 71–99, 1993.
- [74] T. Chen, I. Goodfellow, and J. Shlens. (2015). Net2Net: Accelerating learning via knowledge transfer. *arXiv*. [Online]. Available: <https://arxiv.org/abs/1511.05641>
- [75] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2015, pp. 1–9.
- [76] R. Reed, "Pruning algorithms—A survey," *IEEE Trans. Neural Netw.*, vol. 4, no. 5, pp. 740–747, 1993.
- [77] J. Feng and T. Darrel, "Learning the structure of deep convolutional networks," in *Proc. Int. Conf. Computer Vision*, 2015, pp. 2749–2757.
- [78] J. Yang, J. Ma, M. Berryman, and P. Perez, "A structure optimization algorithm of neural networks for large-scale data sets," in *Proc. 2014 IEEE Int. Conf. Fuzzy Syst. (FUZZ-IEEE)*, 2014, pp. 956–961.
- [79] D. Whitley, T. Starkweather, and C. Bogart, "Genetic algorithms and neural networks: Optimizing connections and connectivity," *Parallel Comput.*, vol. 14, no. 3, pp. 347–361, 1990.
- [80] K. O. Stanley and R. Miikkulainen, "Efficient evolution of neural network topologies," in *Proc. Congr. Evolutionary Computation (CEC02)*, 2002, pp. 1757–1762.
- [81] T. Shinozaki and S. Watanabe, "Structure discovery of deep neural network based on evolutionary algorithms," in *Proc. 2015 IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 4979–4983.
- [82] B. Shahriari, K. Swersky, Z. Wang, R. P. Adams, and N. de Freitas, "Taking the human out of the loop: A review of Bayesian optimization," *Proc. IEEE*, vol. 104, no. 1, pp. 148–175, 2016.
- [83] D. Ramachandram, M. Lisicki, T. Shields, M. Amer, and G. Taylor, "Structure optimization for deep multimodal fusion networks using graph-induced kernels," in *Proc. 25th European Symp. Artificial Neural Networks, Computational Intelligence, and Machine Learning (ESANN)*, Bruges, Belgium, 2017, pp. 11–16.
- [84] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*, vol. 1. Cambridge, MA: MIT Press, 1998.
- [85] B. Zoph and Q. V. Le. (2016). Neural architecture search with reinforcement learning. *arXiv*. [Online]. Available: <https://arxiv.org/abs/1611.01578>
- [86] P. Kulkarni, J. Zepeda, F. Jurie, P. Pérez, and L. Chevallier, "Learning the structure of deep architectures using L1 regularization," in *Proc. British Machine Vision Conf.*, 2015, pp. 23.1–23.11.
- [87] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *J. Mach. Learning Res.*, vol. 15, no. 1, pp. 1929–1958, 1 Jan. 2014.
- [88] C. Murdock, Z. Li, H. Zhou, and T. Duerig, "Blockout: Dynamic model selection for hierarchical deep networks," in *Proc. Int. Conf. Computer Vision and Pattern Recognition*, 2016, pp. 2583–2591.
- [89] F. Li, N. Neverova, C. Wolf, and G. Taylor, "Modout: Learning multi-modal architectures by stochastic regularization," in *Proc. 2017 IEEE Conf. Automatic Face and Gesture Recognition*, 2017, pp. 422–429.
- [90] M. Firman, "RGBD data sets: Past, present and future," in *Proc. CVPR Workshop on Large Scale 3D Data: Acquisition, Modelling, and Analysis*, 2016.
- [91] A. Geiger, P. Lenz, C. Stillner, and R. Urtasun, "Vision meets robotics: The KITTI data set," *Int. J. Robotics Res.*, vol. 32, no. 11, pp. 1231–1237, 2013.
- [92] W. Maddern, G. Pascoe, C. Linegar, and P. Newman, "1 Year, 1000km: The Oxford RobotCar data set," *Int. J. Robotics Res.*, vol. 36, no. 1, pp. 3–15, 2017.
- [93] C. Chen, R. Jafari, and N. Kehtarnavaz, "UTD-MHAD: A multimodal data set for human action recognition utilizing a depth camera and a wearable inertial sensor," in *Proc. 2015 IEEE Int. Conf. Image Processing (ICIP)*, 2015, pp. 168–172.
- [94] S. Escalera, X. Baró, J. Gonzalez, M. A. Bautista, M. Madadi, M. Reyes, V. Ponce-López, H. J. Escalante, et al., "Chaleam looking at people challenge 2014: Data set and results," in *Proc. Workshop at the European Conf. Computer Vision*, 2014, pp. 459–473.
- [95] F. Ofli, R. Chaudhry, G. Kurillo, R. Vidal, and R. Bajcsy, "Berkeley MHAD: A comprehensive multimodal human action database," in *Proc. 2013 IEEE Workshop on Applications of Computer Vision*, 2013, pp. 53–60.
- [96] A. Pablo, Y. Mollard, F. Golemo, A. C. Murillo, M. Lopes, and J. Civera, "A multimodal human-robot interaction data set," in *Proc. Neural Information Processing Systems*, 2016, pp. 1–5.
- [97] F. Ringeval, A. Sonderegger, J. Sauer, and D. Lalanne, "Introducing the Recola multimodal corpus of remote collaborative and affective interactions," in *Proc. Automatic Face and Gesture Recognition 10th IEEE Int. Conf. Workshops*, 2013, pp. 1–8.
- [98] O. Banos, C. Villalonga, R. Garcia, A. Saez, M. Damas, J. A. Holgado-Terriza, S. Lee, H. Pomares, and I. Rojas, "Design, implementation and validation of a novel open framework for agile development of mobile health applications," *Biomedical Eng. Online*, vol. 14, no. 2, p. S6, 2015. [Online]. Available: <https://doi.org/10.1186/1475-925X-14-S2-S6>
- [99] J. Mao, J. Xu, K. Jing, and A. L. Yuille, "Training and evaluating multimodal word embeddings with large-scale web annotated images," in *Proc. Advances in Neural Information Processing Systems*, 2016, pp. 442–450.
- [100] J. Arevalo, T. Solorio, M. Montes-y Gómez, and F. A. González. (2017). Gated multimodal units for information fusion. *arXiv*. [Online]. Available: <https://arxiv.org/abs/1702.01992>
- [101] Y.-G. Jiang, Z. Wu, J. Wang, X. Xue, and S.-F. Chang, "Exploiting feature and class relationships in video categorization with regularized deep neural networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, 2017. doi: <https://doi.org/10.1109/TPAMI.2017.2670560>
- [102] R. Min, N. Kose, and J.-L. Dugelay, "KinectFaceDB: A Kinect database for face recognition," *IEEE Trans. Syst., Man, Cybern., Syst.*, vol. 44, no. 11, pp. 1534–1548, Nov. 2014.
- [103] B. H. Menze, A. Jakob, S. Bauer, J. Kalpathy-Cramer, K. Farahani, J. Kirby, Y. Burren, N. Porz, et al., "The multimodal brain tumor image segmentation benchmark (BRATS)," *IEEE Trans. Med. Imag.*, vol. 34, no. 10, pp. 1993–2024, 2015.

Xiaodong He and Li Deng

Deep Learning for Image-to-Text Generation

A technical overview

©ISTOCKPHOTO.COM/ZAPP2PHOTO

Generating a natural language description from an image is an emerging interdisciplinary problem at the intersection of computer vision, natural language processing, and artificial intelligence (AI). This task, often referred to as *image* or *visual captioning*, forms the technical foundation of many important applications, such as semantic visual search, visual intelligence in chatting robots, photo and video sharing in social media, and aid for visually impaired people to perceive surrounding visual content. Thanks to the recent advances in deep learning, the AI research community has witnessed tremendous progress in visual captioning in recent years. In this article, we will first summarize this exciting emerging visual captioning area. We will then analyze the key development and the major progress the community has made, their impact in both research and industry deployment, and what lies ahead in future breakthroughs.

Introduction

It has been long envisioned that machines one day will understand the visual world at a human level of intelligence. Thanks to the progress in deep learning [15], [36], [59], [60], [69], researchers can now build very deep convolutional neural networks (CNNs) and achieve an impressively low error rate for tasks like large-scale image classification [9], [15], [23]. In these tasks, one way for researchers to train a model to predict the category of a given image is to first annotate each image in a training set with a label from a predefined set of categories. Through such fully supervised training, the computer learns how to classify an image.

However, in tasks like image classification, the content of an image is usually simple, containing a predominant object to be classified. The situation could be much more challenging when we want computers to understand complex scenes. Image captioning is one such task. The challenges come from two perspectives. First, to generate a semantically meaningful and syntactically fluent caption, the system needs to detect salient semantic concepts in the image, understand the relationships among them, and compose a coherent description about the overall content of the image, which involves language and common-sense knowledge

Digital Object Identifier 10.1109/MSP.2017.2741510
Date of publication: 13 November 2017

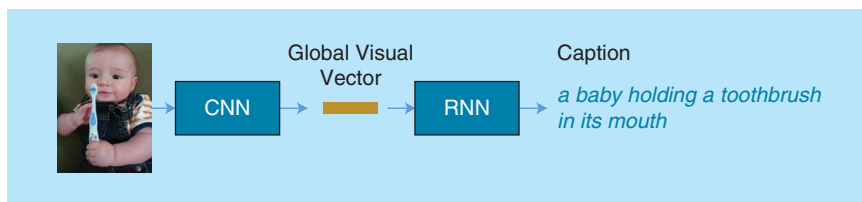


FIGURE 1. An illustration of the CNN-RNN-based image captioning framework.

modeling beyond object recognition. In addition, due to the complexity of scenes in the image, it is difficult to represent all fine-grained, subtle differences among them with the simple attribute of category. The supervision for training image captioning models is a full description of the content of the image in natural language, which is sometimes ambiguous and lacks fine-grained alignments between the subregions in the image and the words in the description.

Moreover, unlike image classification tasks, where we can easily tell if the classification output is correct or wrong after comparing it to the ground truth, there are multiple valid ways to describe the content of an image. It is not easy to tell if the generated caption is correct or not, at what degree. In practice, human studies are often employed to judge the quality of the caption given an image. However, since human evaluation is costly and time-consuming, many automatic metrics are proposed, which could serve as a proxy mainly for speeding up the development cycle of the system.

Early approaches to image captioning can be roughly divided into two families. The first one is based on template matching [6], [16], [17]. These approaches start from detecting objects, actions, scenes, and attributes in images and then fill them into a hand-designed and rigid sentence template. The captions generated by these approaches are not always fluent and expressive. The second family is grounded on retrieval-

based approaches, which first select a set of the visually similar images from a large database and then transfer the captions of retrieved images to fit the query image [10], [20]. There is little flexibility to modify words based on the content of the query image, since they directly rely on captions of training images and cannot generate new captions.

Deep neural networks can potentially address both of these issues by generating fluent and expressive captions, which can also generalize beyond those in the train set. In particular, recent successes of using neural networks in image classification [9], [15], [23] and object detection [8] have motivated strong interest in using neural networks for visual captioning.

Major deep-learning paradigms for image captioning

The end-to-end framework

Vector-to-sequence learning

Motivated by the recent success of sequence-to-sequence learning in machine translation [37]–[39], researchers studied an end-to-end encoder-decoder framework for image captioning [2]–[4], [12], [26]. Figure 1 illustrates a typical encoder-decoder-based captioning system [26].

In such a framework, first the raw image is encoded by a global visual feature vector which represents the overall semantic information of the image, via deep CNNs. As illustrated in Figure 2, a CNN consists of several convolutional, max-pooling, response-normalization, and fully connected layers. This architecture has been very successful for large-scale image classification [21], and the learned features have shown to transfer to a broad variety of vision tasks [40]. Usually, given a raw image, the activation

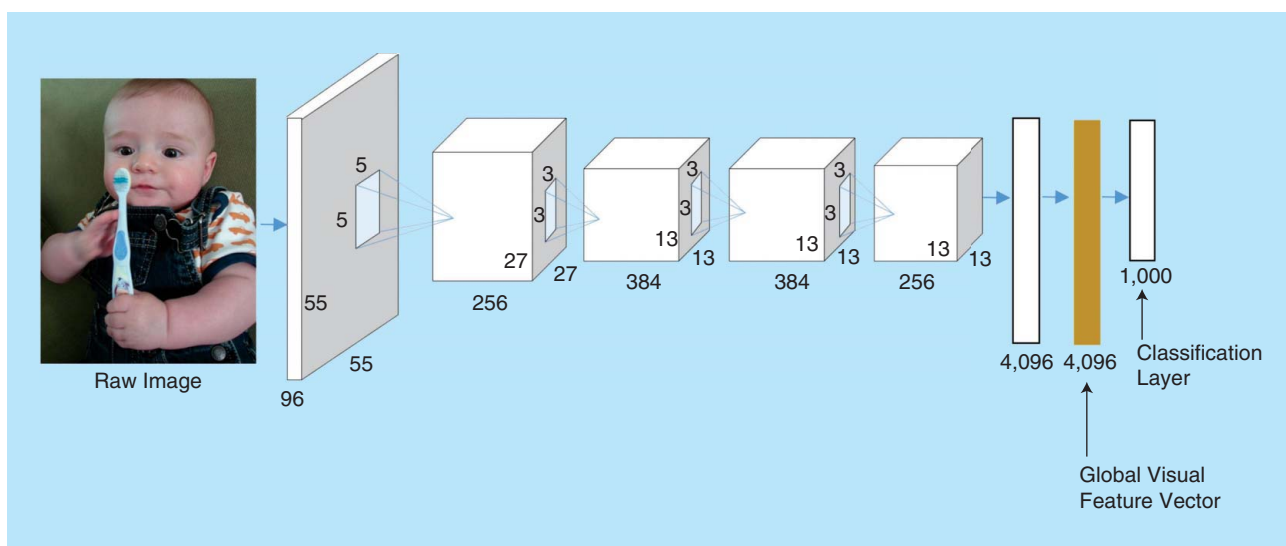


FIGURE 2. An illustration of a deep CNN such as the AlexNet [15]. The CNN is trained for a 1,000-class image classification task on the large-scale ImageNet data set [41]. The last layer of the AlexNet contains 1,000 nodes, each corresponding to a category. The second last fully connected dense layer is usually extracted as the global visual feature vector, representing the semantic content of the overall images.

values at the second last fully connected layer are extracted as the global visual feature vector.

Once the global visual vector is extracted, it is then fed into a recurrent neural network (RNN)-based decoder for caption generation, as illustrated in Figure 3. In practice, a long-short memory network (LSTM) [40] or gated recurrent unit (GRU) [39] variation of the RNN is often used; both have been shown to be more efficient and effective in training and capturing long-span language dependencies than vanilla RNNs [38], [39], and both have found successful applications in action recognition tasks [62], [63].

The representative set of studies using the aforementioned end-to-end framework include [2]–[4], [7], [11]–[13], [19], and [26] for image captioning and [1], [21] [24], [25], and [32] for video captioning. The differences of the various methods mainly lie in the types of CNN architectures and the RNN-based language models. For example, the vanilla RNN was used in [12] and [19], while the LSTM was used in [26]. The visual feature vector was only fed into the RNN once at the first time step in [26], while it was used at each time step of the RNN in [19].

The attention mechanism

Most recently, [29] utilized an attention-based mechanism to learn where to focus in the image during caption generation. The attention architecture is illustrated in Figure 4. Different from the simple encoder-decoder approach, the attention-based approach first uses the CNN to not only generate a global visual vector but also generate a set of visual vectors for subregions in the image. These subregion vectors can be extracted from a lower convolutional layer in the CNN. Then, in language generation, at each step of generating a new word, the RNN will refer to these subregion vectors and determine the likelihood that each of the subregions is relevant to the current state to generate the word. Eventually, the attention mechanism will form a contextual vector, which is a sum of subregional visual vectors weighted by the likelihood of relevance, for the RNN to decode the next new word.

This work was followed by [30], which introduced a “review” module to improve the attention mechanism and further by [18], which proposed a method to improve the correctness of visual attention. More recently, based on object detection, a bottom-up attention model was proposed in [64],

which demonstrated a state-of-the-art performance on image captioning. In the end-to-end framework, all of the model parameters, including the CNN, the RNN, and the attention model, are trained jointly in an end-to-end fashion; hence, the term *end to end*.

A compositional framework

Different from the end-to-end encoder-decoder framework previously described, a separate class of image-to-text approaches uses an explicit semantic-concept-detection process for caption generation. The detection model and other modules are often trained separately. Figure 5 illustrates a semantic-concept-detection-based compositional approach proposed by Fang et al. [5].

In this framework, the first step in the caption generation pipeline detects a set of semantic concepts, known as *tags* or *attributes*, that are likely to be part of the image’s description. These tags may belong to any part of speech, including nouns, verbs, and adjectives. Unlike image classification, standard supervised learning techniques are not directly applicable for learning detectors since the supervision only contains the whole image and the human-annotated whole sentence of caption, while the image bounding boxes corresponding to the words are unknown. To address this issue, [5] proposed learning the detectors using the weakly supervised approach of multiple instance learning (MIL) [42], [43], while in [33], this problem is treated as a multilabel classification task.

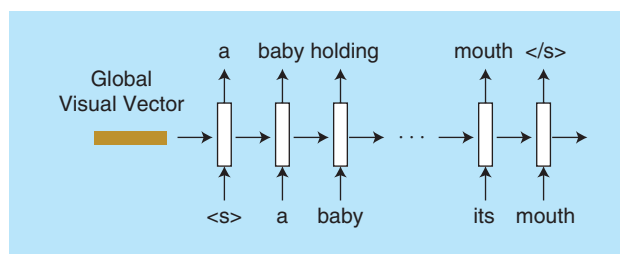


FIGURE 3. An illustration of an RNN-based caption decoder. At the initial step, the global visual vector, which represents the overall semantic meaning of the image, is fed into the RNN to compute the hidden layer at the first step while the sentence-start symbol $\langle s \rangle$ is used as the input to the hidden layer at the first step. Then the first word is generated from the hidden layer. Continuing this process, the word generated in the previous step becomes the input to the hidden layer at the next step to generate the next word. This generation process keeps going until the sentence-end symbol, $\langle /s \rangle$, is generated.

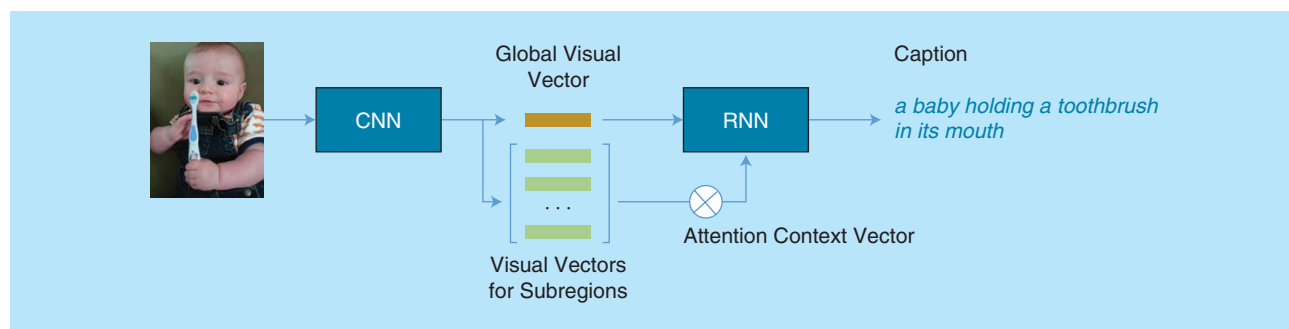


FIGURE 4. An illustration of the attention mechanism in the image caption generation process.

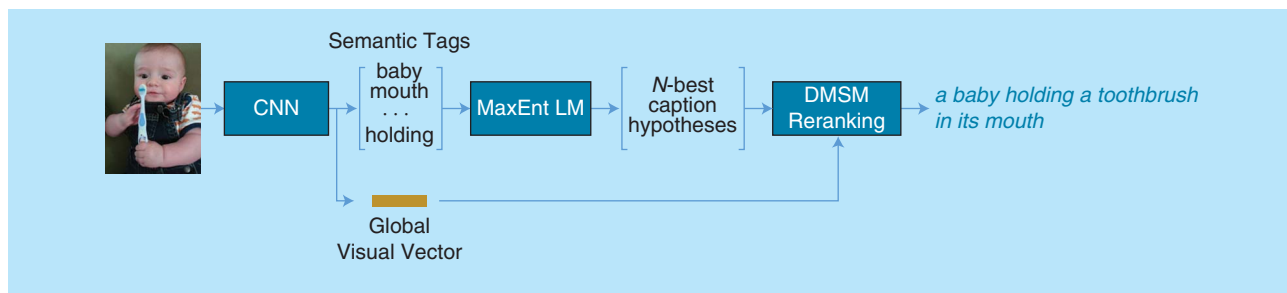


FIGURE 5. An illustration of a semantic-concept-detection-based compositional approach [5].

In [5], the detected tags are then fed into an n -gram-based max-entropy language model to generate a list of caption hypotheses. Each hypothesis is a full sentence that covers certain tags and is regularized by the syntax modeled by the language model, which defines the probability distribution over word sequences.

All of these hypotheses are then reranked by a linear combination of features computed over an entire sentence and the whole image, including sentence length, language model scores, and semantic similarity between the overall image and an entire caption hypothesis. Among them, the image-caption semantic similarity is computed by a deep multimodal similarity model (DMSM), which consists of a pair of neural networks, one for mapping each input modality, image, and language, to be vectors in a common semantic space. Image-caption semantic similarity is then defined as the cosine similarity between their vectors.

Compared to the end-to-end framework, the compositional approach provides better flexibility and scalability in system development and deployment and facilitates exploiting various data sources to optimizing the performance of different modules more effectively, rather than learn all of the models on limited image-caption paired data. On the other hand, the end-to-end model usually has a simpler architecture and can optimize the overall system jointly for a better performance.

More recently, a class of models has been proposed to integrate explicit semantic-concept detection in an encoder-decoder framework. A general diagram of this class of models is illustrated in Figure 6. For example, [1] applied retrieved sentences as additional semantic information to guide the LSTM when generating captions, while [31] and [33] applied a semantic-concept-detection process before generating sentences. In [7], a semantic compositional network is constructed based on the probability of detected semantic concepts for composing captions.

Other related work

Other related work also learns a joint embedding of visual features and associated captions, including [5] for image captioning and [21] for video captioning. Most recently, [27] has looked into generating dense image captions for individual regions in images. In addition, a variational autoencoder was developed in [22] for image captioning. Also motivated by its recent success, researchers proposed a set of reinforcement learning-based

algorithms to directly optimize the model for specific rewards. For example, [67] proposed a self-critical sequence training algorithm. It uses the REINFORCE algorithm to optimize a particular evaluation metric that is usually not differentiable and therefore not easy to optimize by conventional gradient-based methods. In [65], within the actor-critic framework, a policy network and a value network are learned to generate the caption by optimizing a visual semantic reward, which measures the similarity between the image and generated caption. Relevant to image-caption generation, models based on generative adversarial networks (GANs) recently have been proposed for text generation. Among them, SeqGAN [68] models the generator as a stochastic policy in reinforcement learning for discrete outputs like texts, while RankGAN [66] proposed a ranking-based loss for the discriminator, which gives better assessment of the quality of the generated text and therefore leads to a better generator.

Metrics

The quality of the automatically generated captions is evaluated and reported in the literature in both automatic metrics and human studies. Commonly used automatic metrics include BLEU [45], METEOR [44], CIDEr [46], and SPICE [47]. BLEU [45] is widely used in machine translation and measures the fraction of n -grams (up to four grams) that are in common between a hypothesis and a reference or set of references. METEOR [44] instead measures unigram precision and recall, but extending exact word matches to include similar words based on WordNet synonyms and stemmed tokens. CIDEr [46] also measures the n -gram match between the caption hypothesis and the references, while the n -grams are weighted by term frequency-inverse document frequency (TF-IDF). On the other hand, SPICE [47] measures the F1 score of semantic propositional content contained in image captions given the references, and therefore it has the best correlation to human judgment [47]. These automatic metrics can be computed efficiently. They can greatly speed up the development of image captioning algorithms. However, all of these automatic metrics are known to only roughly correlate with human judgment [50].

Benchmarks

Researchers have created many data sets to facilitate the research of image captioning. The Flickr data set [49] and the PASCAL sentence data set [48] were created for facilitating

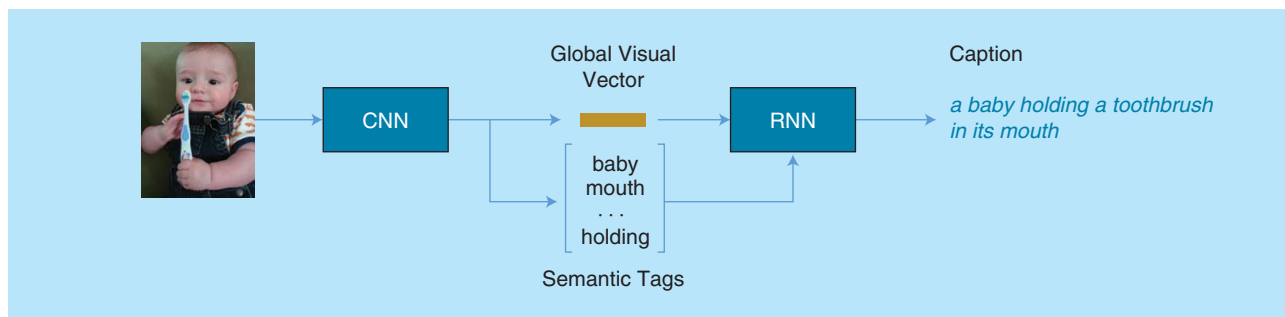


FIGURE 6. An illustration of integrate explicit semantic-concept-detection in an encoder-decoder framework.

the research of image captioning. More recently, Microsoft sponsored the creation of the Common Objects in Context (COCO) data set [51], the largest image captioning data set available to the public today. The availability of the large-scale data sets significantly prompted research in image captioning in the last several years.

In 2015, approximately 15 groups participated in the COCO Captioning Challenge [52]. The entries in the challenge are evaluated by human judgment. Five human judge metrics are listed in Table 1. In the competition, all entries are assessed based on the results from metric 1 (M1) and metric 2 (M2). The other metrics have been used as diagnostic and interpretation of the results. Specifically, in evaluation, each task presents a human judge with an image and two captions: one is automatically generated, and the other is a human caption. For M1, the judge is asked to select which caption better describes the image, or to choose the “same” option when they are of equal quality. For M2, the judge is asked to tell which of the two captions is generated by a human. If the judge chooses the automatically generated caption, or chooses the “cannot tell” option, it is considered to have passed the Turing test. Table 2 tabulates the results of the 15 entries in the 2015 COCO Captioning Challenge. Among them, the Microsoft Research entry (MSR) achieves the best performance on the Turing test metric, while the Google team outperforms others in the percentage

Microsoft Research and Google jointly received first prize in the 2015 COCO Image Captioning Challenge.

of captions that were as good or better than human captions. Overall, Microsoft Research and Google jointly received first prize in the 2015 COCO Image Captioning Challenge. The results of two special systems, human and random, are also included for reference.

There are more systems that have been developed since the 2015 COCO competition. However, due to the high cost, human judging was no longer performed. Instead, the organizers of the COCO benchmark set up an automatic evaluation server. The server can receive the captions generated by a new system and then evaluate and report the results on the blind test set in automatic metrics. Table 3 summarizes the top 24 entries plus the human system as of August 2017, ranked by SPICE, using 40 references per image [52].

Note that these 24 systems outperform the human system in all automatic metrics except SPICE. However, in human judgment, it is likely that the human system still has a lead, given that in Table 2 there is a huge gap between the best systems and a human.

Industrial deployment

Given the fast progress in the research community, the industry started deploying image captioning services. In March 2016, Microsoft released the first public image captioning application programming interface as a cloud service [53]. To showcase the usage of the functionality, it deployed a web application called CaptionBot (<http://CaptionBot.ai>) which captions arbitrary pictures users uploaded [33]. The service also supports applications like Seeing AI, designed for the low-vision community, that narrate the world around people who are blind or visually impaired [71]. More recently, Microsoft further deployed the caption service in its widely used product Office, specifically, Word and PowerPoint, for automatically generating alt-text, i.e., text descriptions of pictures, for accessibility [61]. Facebook released an automatic image captioning tool that provides a list of objects and scenes identified in a photo [34]. Meanwhile, although the service has not yet been deployed publicly, Google open sourced its image captioning system for the community [35]. With all of these industrial-scale deployment and open-source projects, a massive number of images and user feedback in real-world scenarios are collected and serve as training data to continuously

Table 1. Human evaluation metrics in the 2015 COCO Captioning Challenge.

Metric	Comment
M1	Percentage of captions that are evaluated as better or equal to human caption.
M2	Percentage of captions that pass the Turing test.
M3	Average correctness of the captions on a scale from one to five (incorrect–correct).
M4	Average amount of detail of the captions on a scale from one to five (lack of details–very detailed).
M5	Percentage of captions that are similar to human description.

Table 2. Human evaluation results of entries in the 2015 COCO Captioning Challenge.

Entry	M1	M2	M3	M4	M5	Date
Human	0.638	0.675	4.836	3.428	0.352	23 March 2015
Google	0.273	0.317	4.107	2.742	0.233	29 May 2015
MSR	0.268	0.322	4.137	2.662	0.234	8 April 2015
Montreal/Toronto	0.262	0.272	3.932	2.832	0.197	14 May 2015
MSR Captivator	0.25	0.301	4.149	2.565	0.233	28 May 2015
Berkeley LRCN	0.246	0.268	3.924	2.786	0.204	25 April 2015
m-RNN	0.223	0.252	3.897	2.595	0.202	30 May 2015
Nearest Neighbor	0.216	0.255	3.801	2.716	0.196	15 May 2015
PicSOM	0.202	0.25	3.965	2.552	0.182	26 May 2015
Brno University	0.194	0.213	3.079	3.482	0.154	29 May 2015
m-RNN (Baidu/UCLA)	0.19	0.241	3.831	2.548	0.195	26 May 2015
MIL	0.168	0.197	3.349	2.915	0.159	29 May 2015
MLBL	0.167	0.196	3.659	2.42	0.156	10 April 2015
NeuralTalk	0.166	0.192	3.436	2.742	0.147	15 April 2015
ACVT	0.154	0.19	3.516	2.599	0.155	26 May 2015
Tsinghua Bigeye	0.1	0.146	3.51	2.163	0.116	23 April 2015
Random	0.007	0.02	1.084	3.247	0.013	29 May 2015

Table 3. The state-of-the-art image captioning systems in automatic metrics (as of 8 December 2016).

Entry	CIDEr-D	METEOR	BLEU-4	SPICE (x10)	Date
Watson Multimodal	1.123	0.268	0.344	0.204	16 November 2016
DONOT_FAIL_AGAIN	1.01	0.262	0.32	0.199	22 November 2016
Human	0.854	0.252	0.217	0.198	23 March 2015
MSM@MSRA	1.049	0.266	0.343	0.197	25 October 2016
MetaMind/VT_GT	1.042	0.264	0.336	0.197	1 December 2016
ATT-IMG (MSM@MSRA)	1.023	0.262	0.34	0.193	13 June 2016
G-RMI(PG-SPIDEr-TAG)	1.042	0.255	0.331	0.192	11 November 2016
DLTC@MSR	1.003	0.257	0.331	0.19	4 September 2016
Postech_CV	0.987	0.255	0.321	0.19	13 June 2016
G-RMI (PG-BCMR)	1.013	0.257	0.332	0.187	30 October 2016
feng	0.986	0.255	0.323	0.187	6 November 2016
THU_MIG	0.969	0.251	0.323	0.186	3 June 2016
MSR	0.912	0.247	0.291	0.186	8 April 2015
reviewnet	0.965	0.256	0.313	0.185	24 October 2016
Dalab_Master_Thesis	0.96	0.253	0.316	0.183	28 November 2016
ChallS	0.955	0.252	0.309	0.183	21 May 2016
ATT_VC_REG	0.964	0.254	0.317	0.182	3 December 2016
AugmentCNNwithDe	0.956	0.251	0.315	0.182	29 March 2016
AT	0.943	0.25	0.316	0.182	29 October 2015
Google	0.943	0.254	0.309	0.182	29 May 2015
TsinghuaBigeye	0.939	0.248	0.314	0.181	9 May 2016

improve the performance of the system and stimulate new researches in deep visual understanding.

Outlook

Image-to-text generation is an important interdisciplinary area across computer vision and natural language processing. It also forms the technical foundation of many important applications. Thanks to deep-learning technologies, we have seen significant progress in this area in recent years. In this article, we have reviewed the key developments that the community has made and their impact in both research and industry deployment. Looking forward, image captioning will be a key subarea in the image–natural language multimodal intelligence field. A number of new problems in this field have been proposed lately, including visual question answering [54], [55], [70], visual storytelling [58], visually grounded dialog [56], and image synthesis from text description [57], [72]. The progress in multimodal intelligence is critical for building more general AI abilities in the future, and we hope the overview provided in this article can encourage students and researchers to enter and contribute to this exciting AI area.

Authors

Xiaodong He (xiaohex@microsoft.com) received his bachelor's degree from Tsinghua University, Beijing, China, in 1996, his M.S. degree from the Chinese Academy of Sciences, Beijing, in 1999, and his Ph.D. degree from the University of Missouri–Columbia in 2003. He is a principal researcher in the Deep Learning Group of Microsoft Research, Redmond, Washington. He is also an affiliate professor in the Department of Electrical Engineering and Computer Engineering at the University of Washington, Seattle. His research interests are mainly in artificial intelligence areas including deep learning, natural language processing, computer vision, speech, information retrieval, and knowledge representation. He received several awards including the Outstanding Paper Award at the 2015 Conference of the Association for Computational Linguistics (ACL). He has held editorial positions on several IEEE journals, was the area chair for the North American Chapter of the 2015 Conference of the ACL, and served on the organizing committee/program committee of major speech and language processing conferences. He is a Senior Member of the IEEE.

Li Deng (l.deng@ieee.org) received the Ph.D. degree from the University of Wisconsin–Madison in 1987. He was an assistant professor (1989–1992), tenured associate professor (1992–1996), and full professor (1996–1999) at the University of Waterloo, Ontario, Canada. In 1999, he joined Microsoft Research, Redmond, Washington, where he currently leads the research and development of deep learning as a partner research manager of its Deep Learning Technology Center, and where he is a chief scientist of artificial intelligence. Since 2000, he has also been an affiliate full professor and graduate committee member at the University of Washington, Seattle. He is a Fellow of the IEEE, the Acoustical Society of America, and the International Speech Communication Association. He served on the Board of Governors of the IEEE Signal Processing Society (SPS) (2008–2010), and as editor-in-chief of *IEEE Signal*

Processing Magazine (2009–2011), which earned the highest impact factor in 2010 and 2011 among all IEEE publications and for which he received the 2012 IEEE SPS Meritorious Service Award. He recently joined Citadel as its chief artificial intelligence officer.

References

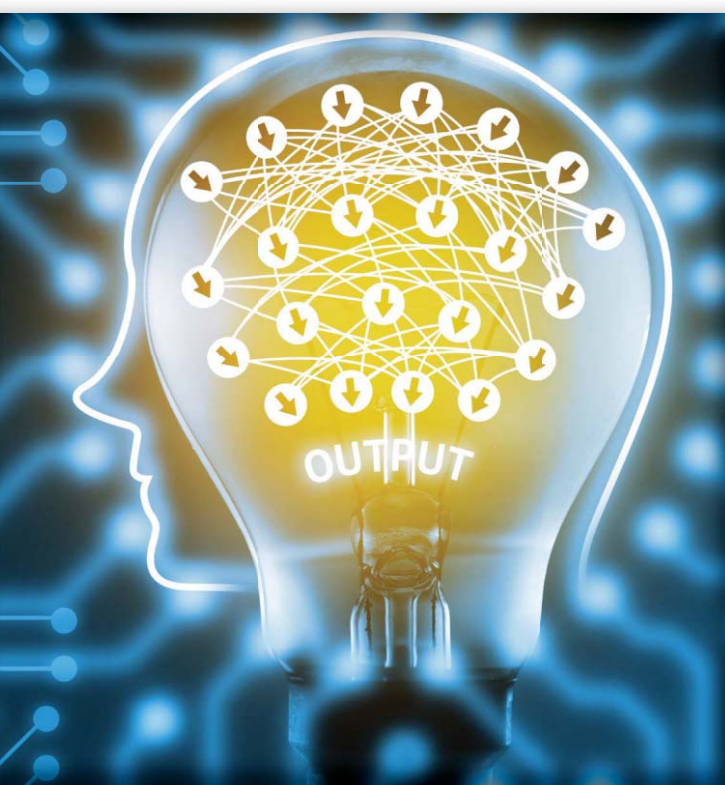
- [1] N. Ballas, L. Yao, C. Pal, and A. Courville, "Delving deeper into convolutional networks for learning video representations," in *Proc. Int. Conf. Learning Representations*, 2016.
- [2] X. Chen and C. Lawrence Zitnick, "Mind's eye: A recurrent visual representation for image caption generation," in *Proc. Conf. Computer Vision and Pattern Recognition*, 2015, pp. 2422–2431.
- [3] J. Devlin, H. Cheng, H. Fang, S. Gupta, L. Deng, X. He, G. Zweig, and M. Mitchell, "Language models for image captioning: The quirks and what works," in *roc. 53rd Annu. Meeting Association Computational Linguistics and the 7th Int. Joint Conf. Natural Language Processing*, 2015, Beijing, China, pp. 100–105.
- [4] J. Donahue, L. Anne Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell, "Long-term recurrent convolutional networks for visual recognition and description," in *Proc. Conf. Computer Vision and Pattern Recognition*, 2015, pp. 2625–2634.
- [5] H. Fang, S. Gupta, F. Iandola, R. K. Srivastava, L. Deng, P. Dollár, J. Gao, X. He, M. Mitchell, J. C. Platt, et al. "From captions to visual concepts and back," in *Proc. Conf. Computer Vision and Pattern Recognition*, 2015, pp. 1473–1482.
- [6] A. Farhadi, M. Hejrati, M. A. Sadeghi, P. Young, C. Rashtchian, J. Hockenmaier, and D. Forsyth, "Every picture tells a story: Generating sentences from images," in *Proc. European Conf. Computer Vision*, 2010.
- [7] Z. Gan, C. Gan, X. He, Y. Pu, K. Tran, J. Gao, L. Carin, and L. Deng, "Semantic compositional networks for visual captioning," *Proc. Conf. Computer Vision and Pattern Recognition*, 2017.
- [8] R. Girshick. "Fast r-CNN," in *Proc. Int. Conf. Computer Vision*, 2015, pp. 1440–1448.
- [9] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. Conf. Computer Vision and Pattern Recognition*, 2016, pp. 770–778.
- [10] M. Hodosh, P. Young, and J. Hockenmaier, "Framing image description as a ranking task: Data, models and evaluation metrics," *J. Artif. Intell. Res.*, vol. 47, pp. 853–899, 2013.
- [11] X. Jia, E. Gavves, B. Fernando, and T. Tuytelaars, "Guiding the long-short term memory model for image caption generation," in *Proc. Int. Conf. Computer Vision*, 2015, pp. 2407–2415.
- [12] Karpathy and L. Fei-Fei, "Deep visual-semantic alignments for generating image descriptions," in *Proc. Conf. Computer Vision and Pattern Recognition*, 2015, pp. 3128–3137.
- [13] R. Kiros, R. Salakhutdinov, and R. S. Zemel, "Multimodal neural language models," in *Proc. Int. Conf. Machine Learning*, 2014.
- [14] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L.-J. Li, D. A. Shamma, et al. "Visual genome: Connecting language and vision using crowdsourced dense image annotations," *arXiv Preprint*, arXiv:1602.07332, 2016.
- [15] Krizhevsky, I. Sutskever and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proc. Conf. Neural Information Processing Systems*, 2012.
- [16] G. Kulkarni, V. Premraj, V. Ordonez, S. Dhar, S. Li, Y. Choi, A. C. Berg, and T. L. Berg. "Babytalk: Understanding and generating simple image descriptions," in *Proc. Conf. Computer Vision and Pattern Recognition*, 2011, pp. 1601–1608.
- [17] S. Li, G. Kulkarni, T. L. Berg, A. C. Berg, and Y. Choi, "Composing simple image descriptions using web-scale n-grams," in *Proc. 15th Conf. Computational Natural Language Learning*, 2011, pp. 220–228.
- [18] C. Liu, J. Mao, F. Sha, and A. Yuille, "Attention correctness in neural image captioning," *arXiv Preprint*, arXiv:1605.09553, 2016.
- [19] J. Mao, W. Xu, Y. Yang, J. Wang, Z. Huang, and A. Yuille, "Deep captioning with multimodal recurrent neural networks (m-RNN)," in *Proc. Int. Conf. Learning Representations*, 2015.
- [20] V. Ordonez, G. Kulkarni, T. L. Berg, V. Ordonez, G. Kulkarni, and T. L. Berg, "Im2text: Describing images using 1 million captioned photographs," in *Proc. Conf. Neural Information Processing Systems*, 2011.
- [21] Y. Pan, T. Mei, T. Yao, H. Li, and Y. Rui, "Jointly modeling embedding and translation to bridge video and language," in *Proc. Conf. Computer Vision and Pattern Recognition*, 2016, pp. 4594–4602.
- [22] Y. Pu, Z. Gan, R. Henao, X. Yuan, C. Li, A. Stevens, and L. Carin, "Variational autoencoder for deep learning of images, labels and captions," in *Proc. Conf. Neural Information Processing Systems*, 2016.

- [23] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. Comput. Sci. Conf.*, 2014.
- [24] S. Venugopalan, H. Xu, J. Donahue, M. Rohrbach, R. Mooney, and K. Saenko, "Translating videos to natural language using deep recurrent neural networks," in *Proc. Conf. North American Chapter Association Computational Linguistics: Human Language Technologies*, 2015, pp. 1494–1505.
- [25] S. Venugopalan, M. Rohrbach, J. Donahue, R. Mooney, T. Darrell, and K. Saenko, "Sequence to sequence-video to text," in *Proc. Int. Conf. Computer Vision*, 2015, pp. 4534–4542.
- [26] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: A neural image caption generator," in *Proc. Conf. Computer Vision and Pattern Recognition*, 2015, pp. 3156–3164.
- [27] J. Johnson, A. Karpathy, and L. Fei-Fei, "Densecap: Fully convolutional localization networks for dense captioning," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 4565–4574.
- [28] Q. Wu, C. Shen, L. Liu, A. Dick, and A. v d. Hengel, "What value do explicit high level concepts have in vision to language problems?" in *Proc. Conf. Computer Vision and Pattern Recognition*, 2016, pp. 203–212.
- [29] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhutdinov, R. Zemel, and Y. Bengio, "Show, attend and tell: Neural image caption generation with visual attention," in *Proc. Int. Conf. Machine Learning*, 2015.
- [30] Z. Yang, Y. Yuan, Y. Wu, R. Salakhutdinov, and W. W. Cohen, "Review networks for caption generation," in *Proc. Conf. Neural Information Processing Systems*, 2016.
- [31] Q. You, H. Jin, Z. Wang, C. Fang, and J. Luo, "Image captioning with semantic attention," in *Proc. Conf. Computer Vision and Pattern Recognition*, 2016, pp. 4651–4659.
- [32] H. Yu, J. Wang, Z. Huang, Y. Yang, and W. Xu, "Video paragraph captioning using hierarchical recurrent neural networks," in *Proc. Conf. Computer Vision and Pattern Recognition*, 2016, pp. 4584–4593.
- [33] K. Tran, X. He, L. Zhang, J. Sun, C. Carapcea, C. Thrasher, C. Buehler, and C. Sienkiewicz, "Rich image captioning in the wild. Deep Vision Workshop," in *Proc. Conf. Computer Vision and Pattern Recognition*, 2016, pp. 434–441.
- [34] S. Wu, J. Wieland, O. Farivar, and J. Schiller, "Automatic Alt-text: Computer-generated image descriptions for blind users on a social network service," in *Proc. 20th ACM Conf. Computer Supported Cooperative Work and Social Computing*, 2017.
- [35] C. Shalloe. (2016). Open-source code on show and tell: A neural image caption generator. [Online]. Available: <https://github.com/tensorflow/models/tree/master/im2txt>
- [36] L. Deng and D. Yu, *Deep Learning: Methods and Applications*, NOW Publishers, 2014.
- [37] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," in *Proc. Conf. Neural Information Processing Systems*, 2014.
- [38] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," in *Proc. Int. Conf. Learning Representations*, 2015.
- [39] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, "Gated feedback recurrent neural networks," in *Proc. Int. Conf. Machine Learning*, 2015.
- [40] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 98, pp. 1735–1780 1997.
- [41] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. Conf. Computer Vision and Pattern Recognition*, 2009, pp. 248–255.
- [42] H. O. Song, R. Girshick, S. Jegelka, J. Mairal, Z. Harchaoui, and T. Darrell, "On learning to localize objects with minimal supervision," in *Proc. Int. Conf. Machine Learning*, 2014.
- [43] C. Zhang, J. C. Platt, and P. A. Viola, "Multiple instance boosting for object detection," in *Proc. Conf. Neural Information Processing Systems*, 2005.
- [44] S. Banerjee and A. Lavie, "METEOR: An automatic metric for MT evaluation with improved correlation with human judgments," in *Proc. ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, 2005.
- [45] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: a method for automatic evaluation of machine translation," in *Proc. 40th Annu. Meeting Association Computational Linguistics*, 2002, pp. 311–318.
- [46] R. Vedantam, L. Zitnick, and D. Parikh, "CIDEr: Consensus-based image description evaluation," in *Proc. European Conf. Computer Vision*, 2015, pp. 4566–4575.
- [47] P. Anderson, B. Fernando, M. Johnson, and S. Gould, "SPICE: Semantic propositional image caption evaluation," in *Proc. European Conf. Computer Vision*, 2016.
- [48] C. Rashtchian, P. Young, M. Hodosh, and J. Hockenmaier, "Collecting image annotations using Amazon's mechanical turk," in *Proc. NAACL HLT Workshop Creating Speech and Language Data with Amazon's Mechanical Turk*, 2010.
- [49] P. Young, A. Lai, M. Hodosh, and J. Hockenmaier, "From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions," in *Proc. Association Computational Linguistics*, vol. 2, 2014, pp. 67–78.
- [50] D. Elliott and F. Keller, "Comparing automatic evaluation measures for image description," in *Proc. 52nd Annu. Meeting Association Computational Linguistics*, 2014, pp. 452–457.
- [51] T.-Y. Lin, M. Maire, S. Belongie, L. Bourdev, R. Girshick, J. Hays, P. Perona, D. Ramanan, C. Lawrence Zitnick, and P. Dollár, "Microsoft COCO: Common objects in context," in *Proc. European Conf. Computer Vision*, 2015.
- [52] Y. Cui, M. R. Ronchi, T.-Y. Lin, P. Dollár, L. Zitnick. (2015) COCO captioning challenge. [Online]. Available: <http://mscoco.org/dataset/#captions-challenge>
- [53] Microsoft Cognitive Services Computer Vision API. [Online]. Available: <https://www.microsoft.com/cognitive-services/en-us/computer-vision-api>
- [54] Z. Yang, X. He, J. Gao, L. Deng, and A. Smola, "Stacked attention networks for image question answering," in *Proc. Conf. Computer Vision and Pattern Recognition*, 2016, pp. 21–29.
- [55] A. Agrawal, J. Lu, S. Antol, M. Mitchell, L. Zitnick, D. Batra, and D. Parikh, "VQA: Visual question answering," in *Proc. Int. Conf. Computer Vision*, 2015, pp. 2425–2433.
- [56] A. Das, S. Kottur, K. Gupta, A. Singh, D. Yadav, J. M. F. Moura, D. Parikh, and D. Batra, "Visual dialog," in *Proc. Conf. Computer Vision and Pattern Recognition*, 2017.
- [57] H. Zhang, T. Xu, H. Li, S. Zhang, X. Huang, X. Wang, and D. Metaxas, "StackGAN: Text to photo-realistic image synthesis with stacked generative adversarial networks," in *Proc. Int. Conf. Computer Vision*, 2017.
- [58] T.-H. (K.). Huang, F. Ferraro, N. Mostafazadeh, I. Misra, A. Agrawal, J. Devlin, R. Girshick, X. He, P. Kohli, D. Batra, C. Lawrence Zitnick, D. Parikh, L. Vanderwende, M. Galley, and M. Mitchell, "Visual storytelling," in *Proc. 2016 Conf. North American Chapter Association Computational Linguistics: Human Language Technologies*, 2016, pp. 1233–1239.
- [59] G. Dahl, D. Yu, L. Deng, and A. Acero, "Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, pp. 30–42, Jan. 2012.
- [60] G. Hinton, L. Deng, D. Yu, G. E. Dahl, A. Mohamed, and N. Jaitly, A, "Deep neural networks for acoustic modeling in speech recognition," *IEEE Signal Process. Mag.*, vol. 29, pp. 82–97, Dec. 2012.
- [61] K. Koenigsbauer, Microsoft Office Blogs. (2016). [Online]. Available: <https://blogs.office.com/2016/12/20/new-to-office-365-in-december-accessibility-updates-and-more/>
- [62] R. R. Varior, B. Shuai, J. Lu, D. Xu, and G. Wang, "A Siamese long short-term memory architecture for human re-identification," in *Proc. European Conf. Computer Vision*, 2016.
- [63] J. Liu, A. Shahroudy, D. Xu, and G. Wang, "Spatio-temporal LSTM with trust gates for 3D human action recognition," in *Proc. European Conf. Computer Vision*, 2016.
- [64] P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, and L. Zhang, "Bottom-up and top-down attention for image captioning and VQA," *arXiv Preprint*, arXiv:1707.07998.
- [65] Z. Ren, X. Wang, N. Zhang, X. Lv, and L.-J. Li, "Deep reinforcement learning-based image captioning with embedding reward," in *Proc. Conf. Computer Vision and Pattern Recognition*, 2017.
- [66] K. Lin, D. Li, X. He, Z. Zhang, and M.-T. Sun, "Adversarial ranking for language generation," *arXiv Preprint*, arXiv:1705.11001
- [67] S. J. Rennie, E. Marcheret, Y. Mroueh, J. Ross, and V. Goel, "Self-critical Sequence Training for Image Captioning," in *Proc. Conf. Computer Vision and Pattern Recognition*, 2017.
- [68] L. Yu, W. Zhang, J. Wang, and Y. Yu, "SeqGAN: Sequence generative adversarial nets with policy gradient," in *Proc. Association Advancement Artificial Intelligence*, 2017.
- [69] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. Cambridge, MA: MIT Press 2016.
- [70] Q. Wu, D. Teney, P. Wang, C. Shen, A. Dick, and A. van den Hengel, "Visual question answering: A survey of methods and data sets," in *Computer Vision and Image Understanding*. Elsevier, 2017.
- [71] Seeing AI. [Online]. Available: <https://www.microsoft.com/en-us/seeing-ai/>
- [72] S. Reed, Z. Akata, X. Yan, L. Logeswaran, H. Lee, and B. Schiel, "Generative adversarial text to image synthesis," in *Proc. Int. Conf. Machine Learning*, 2016.

Hemanth Venkateswara, Shayok Chakraborty,
and Sethuraman Panchanathan

Deep-Learning Systems for Domain Adaptation in Computer Vision

Learning transferable feature representations



©ISTOCKPHOTO.COM/ZAPP2PHOTO

Domain adaptation algorithms address the issue of transferring learning across computational models to adapt them to data from different distributions. In recent years, research in domain adaptation has been making great progress owing to the advancements in deep learning. Deep neural networks have demonstrated unrivaled success across multiple computer vision applications, including transfer learning and domain adaptation. This article outlines the latest research in domain adaptation using deep neural networks. It begins with an introduction to the concept of knowledge transfer in machine learning and the different paradigms of transfer learning. It provides a brief survey of nondeep-learning techniques and organizes the rapidly growing research in domain adaptation based on deep learning. It also highlights some drawbacks with the current state of research in this area and offers directions for future research.

Introduction to domain adaptation

Traditional machine-learning paradigms like supervised learning train statistical models to make predictions on unseen data in the future. These models do not guarantee optimal performance if the test data are vastly different from the training data. To reduce the effort involved in recollecting labeled data and retraining a new model, knowledge transfer between tasks or domains is desirable [1].

The concept of knowledge transfer and the need for adaptive machine-learning models is illustrated in the example of an autonomous-driving car trained with daytime road-traffic data. This car cannot be used to drive autonomously on the roads at night since the light conditions are vastly different from the data with which it was trained. Similarly, consider a car trained with traffic data from sunny days. This car may not show the same level of performance when it's snowing or raining; or a car trained with road-traffic data from the United States will not work as effectively on the streets of London, where the road-traffic rules may vary with different signage, no turns on red, and driving on the left side of the road being some of them. A self-driving car will need to be trained with London street data (signs, traffic rules, etc.) before it can be put to test on the streets of London.

Digital Object Identifier 10.1109/MSP.2017.2740460
Date of publication: 13 November 2017

However, it would be expensive and time-consuming to acquire such training data and retrain a new self-driving car from scratch, especially considering the time and resources needed to train a self-driving car. It is in these situations that domain adaptation algorithms help to transfer the knowledge gained from learning in one environment and reduce the training effort when adapting the model to a new environment.

Variations in vision-based data can be attributed to multiple causes, such as differences in image quality (resolution, brightness, occlusion, and color), changes in camera perspective, dissimilar backgrounds, and an inherent diversity of the samples themselves. All of these can result in distribution mismatch between training and test data. Distribution mismatch can also arise when training and test data are from different modalities; for example, standard color red, blue, green (RGB) image data versus RGB-depth data as in [2]–[4], RGB data versus image sketches [5], or RGB data versus paintings [6]. The authors in [7] perform heterogeneous face recognition across near-infrared images, RGB images, and image sketches. Castrejon et al. [8] introduce a procedure for multimodal domain adaptation across RGB, sketches, clipart, and textual descriptions of indoor scenes. Distribution mismatch can also be introduced when there is a time lag between the capture of image instances [9]. In all of the aforementioned procedures, different domain adaptation techniques are employed to adapt computational models across distributions.

Domain adaptation deals with knowledge transfer, where knowledge from a source domain is transferred to a target domain in the form of learned models and efficient feature representations. The data from the source and the target, although similar, are from different distributions, for, e.g., U.S. street data versus London street data. A machine-learning model trained on the source data set is often adapted to the target data set. The challenge for transfer of knowledge occurs when there are very limited or no labeled data in the target domain, which makes it hard to train models that need some form of supervision. This section defines the problem of knowledge transfer, describes the different transfer learning paradigms along with domain adaptation, and outlines the relevance of research in domain adaptation.

Problem definition

In a standard supervised learning setting, test data are sampled from the same distribution as the training data. Therefore, trained models can guarantee a level of performance. When test data come from a distribution very different from training data, transfer of knowledge from the training domain is necessary to build robust models. At the core of a transfer learning system is a computational model that retains knowledge from one or more tasks, domains, or distributions and applies that knowledge to develop an effective hypothesis for a new one [10]. Transfer learning is often associated with domain adaptation; however, it is more elucidative to understand transfer learning as a broader paradigm that encompasses multiple types of knowledge transfer [1], [10], [11], one of which is domain adaptation. Therefore, domain adaptation can be treated as a special case of transfer learning. To introduce domain adaptation and its relation to other paradigms

of knowledge transfer, a brief outline of various knowledge transfer paradigms is provided. These are multitask learning (MTL), self-taught learning, sample selection bias, lifelong machine learning (LML), zero-shot learning, and domain adaptation.

For the purpose of this discussion, the definitions of *domain* and *task* are outlined in line with [1]. A domain \mathcal{D} is said to consist of two components, a feature space \mathcal{X} and a marginal probability distribution $P(X)$ that governs the feature space, where $\mathbf{X} = \{\mathbf{x}_1, \dots, \mathbf{x}_n\} \subset \mathcal{X}$ is the set of samples from the feature space. For example, if the learning task is audio transcription, the data from different subjects can be treated as different domains. The voice of the subject can be considered to be the feature space \mathcal{X} , and $\mathbf{X} = \{\mathbf{x}_1, \dots, \mathbf{x}_n\}$ is the set of audio signals (words) uttered by the subject, where $P(X)$ is the marginal probability that governs $\mathbf{X} \subset \mathcal{X}$. Two domains are considered different if their feature spaces are different (for example, different users) or their probability distributions are different (for example, casual conversation versus reading a report). If $\mathcal{D} = \{\mathcal{X}, P(X)\}$ is a domain, then a task \mathcal{T} consists of two components, $\mathcal{T} = \{\mathcal{Y}, f(\cdot)\}$, where \mathcal{Y} is the label space and $f(\cdot)$ is the function $f: \mathcal{X} \rightarrow \mathcal{Y}$. The function $f(\cdot)$ is unknown, and, in a supervised setting, it is learned from training data pairs (\mathbf{x}_i, y_i) , where $\mathbf{x}_i \in \mathcal{X}$ and $y_i \in \mathcal{Y}$. The function $f(\mathbf{x})$ can then be used to predict the label of a test instance \mathbf{x} . From a probabilistic perspective, $f(\mathbf{x})$ can be viewed as the posterior probability $p(y | \mathbf{x})$.

MTL

In this setting, labeled training data are available for a set of K tasks $\mathcal{T} = \{\mathcal{T}_1, \mathcal{T}_2, \dots, \mathcal{T}_K\}$, where each task is associated with a different domain, $\mathcal{D} = \{\mathcal{D}_1, \mathcal{D}_2, \dots, \mathcal{D}_K\}$. Given the k th task, it is not possible to estimate the empirical joint distribution $\hat{P}_k(X, Y)$ reliably with data from the k th domain, $\mathcal{D}_k = \{\mathbf{x}_k^i, \mathbf{y}_k^i\}_{i=1}^{n_k}$, $\mathbf{x}_k^i \in \mathcal{X}_k$ and $\mathbf{y}_k^i \in \mathcal{Y}_k$. A good approximation for $\hat{P}_k(X, Y)$ is learned by exploiting the training data from all of the domains $\mathcal{D} = \{\mathcal{D}_1, \mathcal{D}_2, \dots, \mathcal{D}_K\}$ and learning all of the tasks simultaneously [10]. The tasks are different irrespective of the equality of the domains. In terms of availability of labels, all of the domains usually have labels. Even by this definition, $\hat{P}_k(X, Y)$ is inferred by combining the data from all of the tasks and learning all of the tasks simultaneously. Transfer of knowledge between tasks enhances the performance of each individual task. An introduction and a survey of MTL procedures is provided in [12] and [13].

Self-taught learning

The concept of self-taught learning is based on how humans learn in an unsupervised manner from unlabeled data [14]. In this paradigm, the transfer of knowledge is from unrelated domains in the form of learned representations. Given unlabeled data, $\{\mathbf{x}_u^1, \dots, \mathbf{x}_u^k\}$, where $\mathbf{x}_u^i \in \mathbb{R}^d$, the self-taught learning framework estimates a set of K basis vectors that are later used as a basis to represent the target data. Specifically,

$$\begin{aligned} \min \sum_i \left\| \mathbf{x}_u^i - \sum_j a_j^i \mathbf{b}_j \right\|^2 + \beta \left\| \mathbf{a}_i \right\| \\ \text{s.t. } \left\| \mathbf{b}_j \right\| \leq 1, \forall j \in 1, \dots, K, \end{aligned} \quad (1)$$

where, $\{\mathbf{b}_1, \dots, \mathbf{b}_K\}$ are a set of basis vectors that are learned from unlabeled data and $\mathbf{b}_i \in \mathbb{R}^d$. For input data \mathbf{x}_i^j , the corresponding sparse representation is $\mathbf{a}_i = \{a_i^1, \dots, a_i^K\}$, with a_i^j corresponding to the basis vector \mathbf{b}_j . The transfer of learning occurs when the same set of basis vectors $\{\mathbf{b}_1, \dots, \mathbf{b}_K\}$ is used as a basis to represent labeled target data. Some of the prominent machine-learning and computer vision techniques that incorporate self-taught learning are [15]–[18].

Sample selection bias

The concept of sample selection bias was introduced in economics as a Nobel prize-winning work by James Heckman in 1979 [19]. When a distribution of sampled data does not reflect the true distribution of the data set it is sampled from, it is a case of sample selection bias. For example, a financial bank intends to model the profile of a loan defaulter to deny such defaulters a loan from the bank. It therefore builds a model based on the loan defaulters it has in its records. However, this is a small subset and, therefore, does not truthfully model the general public the bank wants to profile but does not have access to. Therefore, the defaulter profile generated by the bank is offset by what is called the *sample selection bias*. In this learning scenario, a data set $\mathbf{X} = \{\mathbf{x}_i, \mathbf{y}_i\}_{i=1}^n$ is made available. This data set is used to estimate the joint distribution $\hat{P}(X, Y)$, which is an approximation for the true joint distribution $P(X, Y)$. However, $\hat{P}(X, Y) \neq P(X, Y)$, where $\hat{P}(X, Y)$ is the estimated distribution and $P(X, Y)$ is the true distribution. This could be because there are very few data samples, which could lead to a poor estimation of the prior distribution, $\hat{P}(X) \neq P(X)$. Other cases when the training data does not represent the target (test) data and introduces a bias in the class prior ($\hat{P}(Y) \neq P(Y)$) eventually lead to incorrect estimation of the conditional ($\hat{P}(Y|X) \neq P(Y|X)$). To correct this discrepancy, knowledge transfer is implemented by weighting the training data samples to reflect the test distribution [20]. When both the marginal ($\hat{P}(X) \neq P(X)$) and the conditionals are different ($\hat{P}(Y|X) \neq P(Y|X)$), the problem is referred to as *sample selection bias* [21]–[23].

LML

The concept of lifelong learning was discussed in the seminal work by Thrun [24]. The concept of transfer in lifelong learning can be formulated as follows. A machine-learning model trained for K tasks $\{\mathcal{T}_1, \mathcal{T}_2, \dots, \mathcal{T}_K\}$ is updated by learning task \mathcal{T}_{K+1} with data \mathcal{D}_{K+1} . The work discussed if learning the $K + 1$ th task was easier than learning the first task. The key characteristics of lifelong learning are 1) a continuous learning process, 2) knowledge accumulation, and 3) use of past knowledge to assist in future learning [25]. LML differs from MTL because it retains knowledge about previous tasks and applies that knowledge to learn new tasks. It also differs from standard domain adaptation, which transfers knowledge to learn only one task (target).

One-shot learning and zero-shot learning

These can be viewed as extreme cases of transfer learning [11]. Both these forms of transfer seek to learn data categories from

minimal data. The key motivation is the ability to transfer knowledge from previously learned categories to recognize new categories. In one-shot learning, the model is trained to recognize a new category of data based on just one labeled example [26]. It relies on the ability of the model to learn representations that cleanly separate the underlying categories. On the other hand, zero-shot learning is the ability to recognize new categories without having seen any example of them. Zero-data learning [27] and zero-shot learning [28], [29] are examples where the model has learned to transfer knowledge from training data not completely related to the categories of interest. For example, a model that has been trained to recognize breeds of dogs can be provided with a description of the categories {fox, wolf, hyena, wild dog}. Without having ever seen an image of any of these categories, the zero-data learning model can be trained to associate the textual description to learn and recognize the new category.

Domain adaptation

In domain adaptation, the source domain \mathcal{D}_S and the target domain \mathcal{D}_T are not the same, and the goal is to solve a common task $\mathcal{T} = \{\mathcal{Y}, f(\cdot)\}$. For example, in an image-recognition task, the source domain could contain labeled images of objects against a white background, and the target domain could consist of unlabeled images of objects against a noisy and cluttered background. Both the domains inherently have the same set of image categories. The difference between the domains is modeled as the variation in their joint probability distributions $P_S(X, Y) \neq P_T(X, Y)$ [10]. Standard domain adaptation assumes that there is plenty of labeled data in the source domain and there is no labeled data (or few samples) in the target domain. Since there are no labeled samples (or very few) of target data, it is difficult to get a good estimate of $\hat{P}_T(X, Y)$. The key task of domain adaptation lies in approximating $\hat{P}_T(X, Y)$ using the source data distribution estimation $\hat{P}_S(X, Y)$. This is possible because the two domains are assumed to be correlated. This correlation is often modeled as covariate shift, where $P_S(X) \neq P_T(X)$ and $P_S(Y|X) \approx P_T(Y|X)$.

Domain adaptation algorithms are evaluated on the basis of minimizing the expected error of prediction on the target data set. A classification model for the target is usually trained using the source data, which has labels along with the target data without labels (or very few labels). Domain adaptation algorithms can be classified largely into two groups based on the availability of labels for the target data. In unsupervised domain adaptation, there are no labels for the target data. Only the source data has labels [30]–[33]. A classifier trained with the labeled source data are adapted to the unlabeled target data. In supervised domain adaptation, a few labeled samples are present in the target domain for all of the categories. However, these are few in number, and a target classifier trained with only these data points could overfit. This paradigm is also referred to as *semisupervised domain adaptation* because there are labels for only a few target samples in each category and the unlabeled target data is also used in a transductive setting to estimate the labels [34]–[36]. The source data, which have much more labeled data, are used along with the target data to estimate the optimal target classifier and prevent overfitting.

A few domain adaptation methods have labeled target data sets but use the source data to augment the number of labeled samples. These methods are also referred to as *supervised domain adaptation* and are used to train classifiers for unseen target data [35], [37], [38]. Tzeng et al. [39] propose a new paradigm for supervised domain adaptation where a few labels are present for only a few of the categories in the target domain. Their reasonable assumption is that, as is often encountered in a real-world setting, it is possible to get a few labeled target samples. Although the popular setting is two-domain adaptation, there are also examples of multisource domain adaptation as in [40]–[42]. In this article, we consider the following definitions:

- supervised domain adaptation: when the target domain has a few labeled samples for all of the categories
- unsupervised domain adaptation: when the target domain does not have any labeled samples.

A majority of the domain adaptation literature cited in this article is unsupervised.

Relevance of domain adaptation

Human intelligence is a competitive benchmark that machine intelligence is seeking to emulate and eventually outperform. One of the hallmarks of human intelligence is the ability to adapt and transfer knowledge across multiple domains. For example, if humans are familiar with a language, they can easily understand almost anyone speaking it, even if they were to hear it for the first time; or, if a person has learned to drive a car, he or she can easily adapt to driving a truck by adapting some previously learned knowledge to the new setting. To enhance machine intelligence to the level of human intelligence and beyond, machine-learning models will have to model knowledge transfer. The ability to transfer knowledge will provide tools to process the vast amounts of unlabeled data available in the form of online video, audio, images, and text. These advances in artificial intelligence and machine-learning will greatly benefit a wide range of signal processing applications, including communication systems, financial markets, medical imaging, robotics, and digital video processing.

The state-of-the-art algorithms for domain adaptation are dominated by deep-learning-based approaches. Deep-learning methods are outperforming standard nondeep-learning techniques for domain adaptive image classification. The success of deep-learning methods has led to a rapid growth in domain adaptation research. It is necessary to categorize the myriad approaches and organize them to get a better understanding of the current research in domain adaptation. This article provides a classification of deep-learning approaches for domain adaptation. It also highlights the drawbacks with current approaches and outlines directions for future research.

Shallow domain adaptation: Survey

Prior to the introduction of deep neural networks for vision (AlexNet [43]), computer vision researchers relied on handcrafted features like scale invariant feature transform (SIFT) [44], histogram of oriented gradients (HoG) [45], etc. to create a bag-of-words-based vector representation for images and videos [46].

Domain adaptation techniques developed and studied using these features are called *shallow methods* (as opposed to deep-learning methods). It is important to understand some of these approaches, as they form the basis of our understanding of domain adaptation. In addition, the first batch of deep-learning methods developed for domain adaptation are based on a few of these shallow domain adaptation techniques.

There are many nondeep-learning (shallow) approaches that address the problem of domain-shift in unsupervised domain adaptation. All of these procedures work at the level of features, i.e., the images are represented as feature vectors, and the domain adaptation algorithm attempts to reduce the domain disparity between the feature vectors of the source and the target. Since the goal is to classify the target data, one straightforward technique is to modify a support vector machine (SVM) classifier trained for the source data and adapt it to classify target data. In [10], Bruzzone and Marconcini introduce the domain adaptive SVM (DASVM), where the source SVM decision boundary is iteratively modified and adapted to classify target data. In [47], Aytar and Zisserman develop a projective model transfer SVM (PMT-SVM), where a transformation matrix is learned to adapt SVM decision boundaries across domains. Hoffman et al. [48], develop the max-margin domain transfer (MMDT) approach, where a linear SVM decision boundary for a source is transformed to classify target data. Their work is an extension of the seminal work by Saenko et al. [49], where a transformation matrix is learned to cluster the source and target data based on category. Both of these methods consider a few labeled samples in the target domain. Other linear procedures project the source and the target data to a common subspace, where domain alignment is bettered. In [50] and [51] the authors estimate a common subspace to align the principal axes of the source and target feature spaces.

When linear feature-based approaches like linear transformations are inadequate in overcoming domain disparity, nonlinear techniques are applied to ameliorate the domain discrepancy between the source and the target. Nonlinear transformations project the data points to a high-dimensional space and align the domains in that space. Reducing domain disparity through nonlinear alignment of data has been made possible with the maximum mean discrepancy (MMD), which is a nonparametric distance estimate designed by embedding the data into a reproducing kernel Hilbert space (RKHS). The data are mapped to a high-dimensional (possibly infinite-dimensional) space defined by $\Phi(\mathbf{X}) = [\phi(\mathbf{x}_1), \dots, \phi(\mathbf{x}_n)]$. $\phi: \mathbb{R}^d \rightarrow \mathcal{H}$ defines a mapping function, and \mathcal{H} is a RKHS. When two data sets belong to the same distribution, their MMD is zero. Gretton et al. in [52], introduced the MMD to estimate the distance between the source and the target data sets, which is given by

$$\text{MMD} = \left\| \frac{1}{n_s} \sum_{i=1}^{n_s} \phi(\mathbf{x}_i^s) - \frac{1}{n_t} \sum_{j=1}^{n_t} \phi(\mathbf{x}_j^t) \right\|_{\mathcal{H}}^2. \quad (2)$$

The distance between the two distributions is the distance between their means in an RKHS. When the RKHS is universal, the MMD measure approaches zero only when the

distributions are the same. Many nonlinear domain adaptation methods apply the MMD in different ways to align the source and target domains.

In [53], the authors apply MMD to introduce the domain transfer SVM (DTSVM) for video concept detection. Some other procedures apply MMD to reweigh the data points in the source domain to select source data that are similar to the target when training a domain adaptive classifier [42], [54], [55]. Spectral methods apply the MMD to achieve nonlinear alignment of domains. Kernel-principal component analysis (Kernel-PCA) is combined with MMD to determine nonlinear projections that transform the source and target to a common subspace, as in [32], [56], and [57]. Manifold-based approaches are also popular in domain adaptation for computer vision, where the subspace of a domain is treated as a point on the manifold. The curve connecting two subspaces is sampled to determine the transformations that are necessary to transform the source subspace into the target subspace [58]. In [30], the authors determine a product of an infinite number of such transformations that projects the source subspace into the target subspace using the geodesic flow kernel. These are some of the popular techniques for domain adaptation without using deep networks. More detailed surveys of shallow domain adaptation approaches can be found in [1], [59], and [60]. Likewise, the bounds for the expected error on the target and the theoretical foundations of domain mismatch are outlined in [61] and [62].

Insights

Recent years have seen deep-learning systems outperform most nondeep-learning techniques across multiple problems in computer vision, including domain adaptation. Does this mean that shallow domain adaptation procedures are obsolete? Not quite. Most of the deep learning domain adaptation procedures are based on shallow domain adaptation techniques as outlined in the following sections [63]–[65]. Objective functions based on shallow domain adaptation procedures guide deep networks to extract highly adaptive representations. Research and advances in shallow adaptation techniques are necessary for progress in domain adaptation. Shallow methods do not require expensive graphics processing unit systems for deep learning. When training data sets are small or real-time performance is needed, shallow domain adaptation techniques are preferred over deep-learning systems.

Deep-learning domain adaptation: Survey

In recent years, deep neural networks have revolutionized the field of machine learning and computer vision. Deep-learning-based domain adaptation has outperformed nondeep-learning algorithms because of the highly discriminatory nature of the features extracted using deep neural networks. The progress of research in computer vision can be directly linked to the advances in feature extraction and representation techniques. Feature representation is the process of representing the spatial (or spatiotemporal) information in an image (or video) as a vector. Feature descriptors like SIFT and HOG are handcrafted techniques for feature representation that are task and data agnostic. Feature

representations determined using deep networks are task and data specific. The loss functions guide the network in determining the best features for a given data set to achieve a specific task. This is the main advantage of using deep neural networks, which becomes more evident in domain adaptation.

Shallow domain adaptation approaches are considered to be fixed representation approaches. In a fixed representation approach, the features are predetermined and fixed, and domain adaptation is performed using these predetermined features. On the other hand, deep-learning-based domain adaptation methods extract transferable feature representations specific to the data and the adaptation task at hand. The unrivaled success of deep-learning methods in domain adaptation can be attributed to this aspect. Feature representations using deep neural networks are highly nonlinear due to multiple levels of nonlinearity in the feature extraction process. They are also termed *hierarchical features* due to the hierarchical nature of the model and the nonlinear multilayer structure of the network. In this section, we categorize the literature in domain adaptation based on these hierarchical feature representations.

Naïve hierarchical methods

Deep convolutional neural networks (CNNs) have been shown to be very good feature extractors. Deep CNNs trained on millions of images are, by themselves, very good feature extractors, not just for the data set they are trained on, but for any generic image. In [66], Razavian et al. have demonstrated how a deep CNN trained on the ILSVRC 2013 ImageNet data set [67] can be used for extracting generic features for any image. Regular SVMs trained on these generic features have shown astounding results across multiple applications like scene recognition, fine-grained recognition, attribute recognition, and image retrieval. A pretrained CNN can be used to extract generic features for the source and the target. This can be termed as a *naïve* form of domain adaptation.

Pretrained deep neural networks can also be fine-tuned to the task at hand. It is well documented that the lower layers of a CNN extract generic features that are common across multiple tasks, and the upper layers extract task-specific features. Features transition from general to specific by the last layer of the CNN. The work by Yosinski et al. in [68] captures the extent of generality and specificity of neurons in each layer. Transferability has been shown to be negatively affected by two issues: 1) the specificity of neurons (to the source task) in the upper layers adversely affects transfer to the target task, and 2) the fragile nature of dependencies between layers that are task specific inhibits the reuse of layers across different tasks. Adding new layers to a pretrained (trained on source data) network and retraining it with target data is another intuitive method to transfer knowledge in a deep-learning setting. When the entire newly adapted network is fine-tuned with target data, it can lead to a very efficient adaptation. This form of adaptation has been explored in [69]. The authors demonstrate a procedure to reuse the layers trained on the ImageNet data set to compute midlevel representations for images. Despite the differences

in image statistics, these features lead to improved results for object and action classification for different data sets.

The authors in [70] study the features extracted from the final layers of a deep neural network for a fixed set of object classification and detection tasks. The generic features from the fifth, sixth, and seventh fully connected layers of an AlexNet [43] show remarkable adaptation properties and outperform state-of-the-art methods in classification and detection. Whereas [70] studied adaptation using CNNs, [71] studied adaptation of features extracted using stacked denoising autoencoders for text-based sentiment classification.

Insights

To boost the performance for a data set using a shallow procedure, naïve methods can be applied using deep networks as feature extractors [69], [70]. Sometimes, there may be constraints to deploy only shallow methods due to data set size, hardware resources, etc. In such situations, naïve methods can be very effective. When domain discrepancy between the source and the target is not very large, pretrained deep networks provide highly adaptive features for the source and target.

Adopted shallow methods

These sets of deep-learning methods adopt shallow (nondeep learning) domain adaptation procedures in a deep neural network. In these approaches, the features extracted by the layers of the deep network are learned to be domain invariant. Domain alignment is achieved either through MMD, moment alignment [64], or a loss function that drives the source and target classifiers to be indistinguishable. In discussing these methods, the central idea is outlined, leaving out the details of network architecture, optimization procedures, loss functions, etc.

In [72], the authors adapt an AlexNet [43] to output domain invariant features in the final, fully connected $fc8$ layer in the

deep domain confusion algorithm. The network has two loss components: 1) softmax classification loss for the source data points and 2) domain confusion loss. The network minimizes an MMD loss over the source and target data outputs of the $fc8$ layer in every minibatch during training. This is termed the *domain confusion loss*. A related idea is studied in [73], where the network has a domain confusion loss along with a domain classification loss. The domain classification loss ensures the output feature representations of the source and target data are distinct. This is in contrast to the goal of the domain confusion loss, which tries to learn domain-invariant representations. The network is trained to alternately minimize the two losses and reach an equilibrium. Both of these methods assume the presence of a few labeled samples in the target domain.

Long et al. introduce the deep adaptation networks (DAN) model [63], which extends the concept of domain confusion by incorporating an MMD loss for all of the fully connected layers ($fc6$, $fc7$, and $fc8$) of the AlexNet. The MMD loss is estimated for the feature representations over every minibatch during training. The work also introduces MMD estimation computed with an efficient linear complexity based on [74]. The linear MMD estimation is also unbiased because the MMD for the entire source and target data can be expressed as the sum of MMD across minibatches. Based on the network architecture of the DAN [63], Venkateswara et al. [65] develop a hashing algorithm for domain adaptation. The architecture of the domain adaptive hash (DAH) network is based on the VGG-F, and domain alignment is achieved using MMD just like in the DAN. Figure 1 depicts the architecture of the DAH, which is similar to the DAN. The network is trained in an iterative manner using batches comprising source and target data. The loss function of the DAH has three components: 1) MMD loss for the fully connected layers $fc6$, $fc7$, and $hash-fc8$, which aligns the features of the two domains; 2) supervised hash loss for the source, which extracts unique hash codes for every category

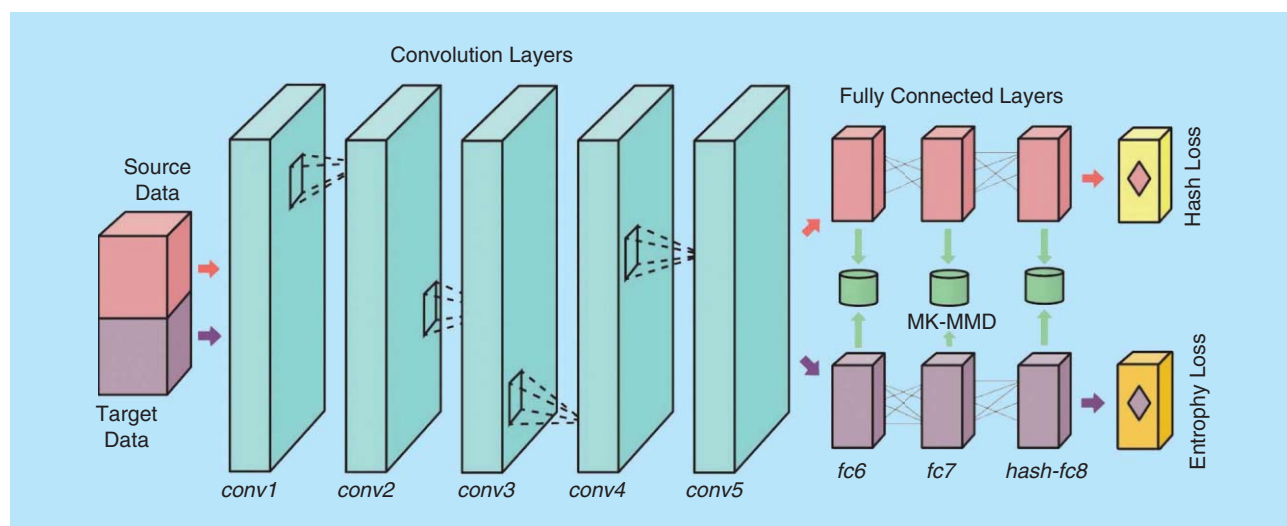


FIGURE 1. The DAH [65] network outputs hash codes of d dimensions for the source and target images. The architecture of the DAH is similar to the domain adaptation network [63]. Adaptive features are extracted by using an MMD loss between the source and target data points in each subset for the fully connected layers $fc6$, $fc7$, and $hash-fc8$. fc : fully connected. $conv$: convolution. (Figure used courtesy of [65].)

in the source and similar hash codes for images belonging to the same category; and 3) unsupervised entropy loss that guides the unlabeled target data to align its hash codes according to the source hash codes. The DAH network solves two important problems: classification with weak supervision or insufficient labels (through domain adaptation) and determining hash codes in an unsupervised setting (hash codes for target data).

An extension to DAN is achieved with the residual transfer network (RTN) in [75], which implements a residual layer as the final layer of the network in addition to a softmax loss. In the RTN, feature adaptation is achieved with MMD loss, and the source and target classifier adaptation is implemented through the residual layer [76]. The source classifier $f_s(\mathbf{x})$ is tightly coupled with the target classifier $f_T(\mathbf{x})$, varying with only a slight perturbation $\Delta f(f_T(\mathbf{x}))$, which is learned by the network, with $f_s(\mathbf{x}) = f_T(\mathbf{x}) + \Delta f(f_T(\mathbf{x}))$. In addition, the source classifier is constrained by the softmax loss over the source data, and the target classifier is constrained with unlabeled entropy loss over the target data.

Compacting deep neural networks and reducing the number of parameters are essential for creating smaller, more manageable networks. These procedures usually replace the larger convolutional layer kernels with kernels of size 1×1 and 3×3 . Although such procedures produce networks that maintain the classification accuracies, Wu et al. [77] note that the adaptability of these networks is adversely affected, resulting in low accuracies for domain adaptation. Wu et al. propose a set of layers called *Conv-M*, which consist of multiscale convolution and deconvolution with kernels larger than 3×3 . The proposed compact network also uses MMD to align the source and target features at multiple layers and produces state-of-the-art results on the standard *Office* and *Office-Caltech* data sets. The network is also guided with a reconstruction loss that reconstructs images using the encoded feature representations. The domain reconstruction and classification network developed by Ghifary et al. [78] is also guided by a reconstruction loss that decodes the feature encoding along with a standard classification loss.

While the MMD is a standard nonparametric measure used to align the features of the domains, Koniusz et al. [4] propose a technique to align the higher-order statistics of the features. The scatter statistics of samples belonging to a class (within-class) are aligned across the two domains. These include the means, scale/shear, and orientation measures of samples belonging to a single class. The procedure also maintains good separation for between-class scatters to enhance classification accuracies. Unlike the popular unsupervised setting, this deep-learning technique is trained using a few labeled data from the target domain.

In all of these deep domain adaptation approaches, the weights are shared between the source and the target network to ensure domain invariant features. The authors in [79] argue that merely ensuring domain invariant features may be detrimental to the discriminative power of the features. Their model is a twin network (one for the source and another for the target) with a loss function over the weights for every source target layer pair. The loss term ensures the weights of the source and

the target are closely related (but not identical). The source network is trained with a softmax loss over the source data, and both the networks also minimize the MMD loss to extract domain invariant features.

Insights

Adopted deep methods are well suited for medium-sized data sets (thousands of images like *Office*). These data sets are large enough to fine-tune a deep network but not too large to fully train a deep network from scratch. A pretrained deep network like Alexnet [43] is often used as a base network and fine-tuned for domain adaptation [63], [65]. One technique to adopt a shallow method is determining a closed-form solution that can then be modeled as an objective function for a deep network [64].

Adversarial learning methods

In recent years, one of the most significant advances to deep learning has been the introduction of generative adversarial networks (GANs) by Goodfellow et al. [80]. GANs are networks that generate data (text, images, audio, etc.) such that the data follow a predetermined distribution $P(X)$. A vanilla GAN implementation has two deep networks, generator $g(\cdot)$ and discriminator $f(\cdot)$, competing against each other. The generator network takes in a noise vector $\mathbf{z} \in \mathbb{R}^d$ sampled from a uniform or normal distribution and generates an output $g(\mathbf{z})$. The discriminator takes in $\mathbf{x} \in P(X)$ and $g(\mathbf{z})$ and tries to discriminate between the two. The generator network tries to fool the discriminator network by generating data that appear to belong to $P(X)$, and the discriminator tries to distinguish between real images and fake images. The equilibrium is a saddle point in the network parameter space. The core concept of the GAN is applied to achieve domain adaptation. Whereas, in a standard GAN, a noise vector $\mathbf{z} \in \mathbb{R}^d$ is converted into a fake image, in a domain adaptive setting, a source image is converted into a fake target image. The pixel-GAN in [33] is a straightforward extension of the GAN for unsupervised domain adaptation. In this model, along with a noise vector input \mathbf{z} , the generator inputs the source image and is trained to convert it into a target image. The discriminator, on the other hand, is trained to distinguish between real target images and generated target images (fake target images generated from the source). In addition, a separate network is trained to classify the generated target images. Along similar lines, Taigman et al. [81] develop an image translation network that converts an image from one domain into an image in another domain using adversarial networks. There have been many recent works applying adversarial training for domain adaptation. A few of the most recent procedures based on the core GAN idea but with subtle variations are [82]–[85].

In the domain adversarial neural network (DaNN) in [86], the authors train a deep neural network in a domain-adversarial manner for image classification-based domain adaptation. The bottom layers of the network act as feature extractors. The features from the bottom layers are fed into two branches of the network. The first branch is a softmax classifier trained with the labeled source data. The second branch is a domain classifier

trained to distinguish between the features of the source and the target. The key to the DaNN is the gradient reversal layer connecting the bottom feature extraction layers and the domain classifier. During back propagation, the gradient from the domain classifier is reversed when learning the feature extractor weights. This technique is popularly called the *gradient reversal*. In this way, the feature extractor is trained to extract domain invariant features. A closely related work is presented in [87].

Liu and Tuzel implement a coupled GAN model (CoGAN) in [88]. The CoGAN trains a coupled network, which shares weights at different layers of the GAN. It is set up so that the lower layers of the generators and the upper layers of the discriminators share weights. A common noise vector, \mathbf{z} , is fed into the two generators $g_1(\cdot)$ and $g_2(\cdot)$ to generate outputs $g_1(\mathbf{z})$ and $g_2(\mathbf{z})$. These outputs are fed into two discriminators $f_1(\cdot)$ and $f_2(\cdot)$. Discriminator $f_1(\cdot)$ is trained to discriminate between $g_1(\mathbf{z})$ and the source \mathbf{x}^s . Likewise, discriminator $f_2(\cdot)$ is trained to discriminate between $g_2(\mathbf{z})$ and the target \mathbf{x}^t . Additionally, the source discriminator has a softmax layer to classify the source data points \mathbf{x}^s . The CoGAN was tested with *MNIST* and *USPS* data to yield impressive unsupervised domain adaptation results. In an extension to the CoGAN, the authors Liu et al. [89] develop an image translation network that combines the CoGAN with a variational autoencoder [90]. This image translation network converts images in the target domain to images in the source domain, which enables efficient classification of target data with a source classifier.

Insights

Adversarial methods are the latest trend in deep learning, and they have shown remarkable performance in domain adaptation. When there is a need to model the source domain distribution, generative models like adversarial networks are beneficial. GANs can be used for image translation [91], converting images from one domain to the other [33], [89]. However, they require large data sets to fully train a deep network since fine-tuning has so far not been implemented with GANs.

Miscellaneous hierarchical methods

One of the earliest procedures for deep-learning domain adaptation was proposed by Chopra et al. [92]. The deep learning for domain adaptation by interpolation between domains learns a cross-domain representation by interpolating the path between the source and target domains along the lines of [58]. Multiple data sets with varying ratios of source and target data points are sampled to create intermediate representations between the two domains. The final cross-domain feature is a concatenation of all of the intermediate feature representations.

Hu et al. [93] develop a metric learning method for supervised transfer learning using clustering. The deep transfer metric learning model trains a deep neural network to minimize intraclass distances and increase interclass distances. Additionally, the features of the last layer of the network are learned to be domain invariant by minimizing the MMD between the source and target feature outputs. Sener et al. [94] develop a deep-learning approach that imputes the labels for the target in a transductive

learning environment. Using these imputed target data labels, the largest margin, nearest-neighbor loss is applied to ensure cyclic consistency of label assignment, and a k-nearest neighbor graph over the target data points is applied to implement structural consistency. The deep network predicts the labels so as to minimize intraclass distances and maximize interclass distances.

Sun et al. [95] develop a domain transfer method called the *localized action frame (LAF)* for fine-grained action localization in temporally untrimmed videos. The LAF motivates domain transfer between weakly labeled web images and videos. The domain transfer works in both directions: the video frames are used to select web images that are relevant and drop nonaction web images, and, in turn, the web images are used to select action-like frames and drop nonaction frames in the video. After the relevant frames and images are selected, a long short-term memory network is used to train a fine-grained action detector to model the temporal evolution of actions and classify the action in the frames. The work also released a data set of sports videos with more than 130,000 videos from 240 categories.

Bousmalis et al. [96] train domain separation networks to extract domain-invariant feature representations and domain-specific representations of source and target data. A shared encoder network $E_c(\mathbf{x})$ is trained to extract domain invariant feature representations for the source and the target data. Private encoder networks $E_p^s(\mathbf{x})$ and $E_p^t(\mathbf{x})$ for the source and target, respectively, are trained to extract feature representations that are distinct from the domain-invariant representations that are the outputs of $E_c(\mathbf{x})$. A shared decoder network is trained to reconstruct the original input data based on the outputs from $E_c(\mathbf{x})$, $E_p^s(\mathbf{x})$, and $E_p^t(\mathbf{x})$. A classifier is trained with the source outputs of $E_c(\mathbf{x})$. The feature representations that are the outputs of $E_c(\mathbf{x})$ can be declared domain-invariant.

Insights

The miscellaneous hierarchical methods do not fall into any of the previous categories. With the success of deep learning, there has been a significant growth in the adoption of deep learning for domain adaptation. So far, there has not been any classification system for the steadily increasing number of deep-learning domain adaptation models. This article provides a preliminary classification system to guide researchers studying deep-learning domain adaptation. With the introduction of new data sets, the modeling of domain-shift, and advances in deep domain adaptation, a more rigorous classification system can be developed over time.

Directions for future research

While the problem of variations in data coming from different distributions is outstanding, the solutions provided by domain adaptation models are not easily applied to real-world applications in computer vision. This can be attributed to the manner in which domain adaptation models are currently being developed and evaluated in the research community. From the early approaches toward general domain adaptation [97] and vision-based domain adaptation [58] to current-day models [4], [33], the standard setup with well defined source and target

data sets and a shared label space has remained constant. Well-defined source and target data sets, discrete and shared label spaces, and constraints on labeled and unlabeled data are possibly some of the reasons for limiting the applicability of domain adaptation to real-world settings. This section discusses a few problems with the way domain adaptation models are developed and evaluated in the research community along with some proposals for changes. Directions for future research in domain adaptation are later outlined. Most of the proposed solutions in domain adaptation are based on models developed in the following environment:

- 1) Two different data sets are used to represent the source and the target domains. Domain adaptation models trained on specific data sets may perform exceedingly well, adapting between the two data sets. These models do not guarantee the same performance when adapting between different data sets. Every new pair of data sets may require training its own adaptation model. A more universal approach would be to have two well-defined domains (rather than data sets) to represent the source and the target. Algorithms developed to address a particular domain shift can guarantee performance across applications that encounter the same domain shift.
- 2) The source data set is labeled, and the target data set is unlabeled in unsupervised domain adaptation. This appears to be a stringent and restrictive constraint because, in a real-world setting, it is possible to have a few labeled samples for the target data set. Most domain adaptation models do not account for labeled target data. Optimal model-parameters and model-design choices, which are usually estimated using a labeled validation set, cannot be applied for unsupervised domain adaptation because there are no labeled target examples. There is no prescribed procedure to validate the model parameters of current domain adaptation methods [32], [56], [98]. A few labeled target examples are essential for validation purposes. Tzeng et al. [39] outline a semisupervised adaptation model, where some of the categories in the target have a few labeled data points. Domain adaptation models that account for a few labeled data points in the target domain can only outperform their unsupervised counterparts.
- 3) The label space of the source and target is exactly the same. Domain adaptation models assume a shared label space between the source and target domains. Most real-world scenarios satisfy such a criteria. However, robust domain adaptation models should account for a relaxed setting, where there is no restriction on the label space of the domains being exactly the same.
- 4) Current domain adaptation approaches are modeled to solve only a specific subset of adaptation problems, which assume closed-world representations with a fixed set of discrete and disjoint labels. However, most real-world problems have generic representations, and they cannot be limited to discrete or disjoint label settings. Newer algorithms in domain adaptation must remodel the basic problem setup and evaluation protocol in step with real-world applications.

To solve real-world domain adaptation problems and eventually address the problem of artificial general intelligence, these changes are suggested in domain adaptation research. Apart from these changes to the basic approach in domain adaptation, the following subsections outline the specific directions for future research in this area.

Modeling domain shift

The concept of a domain has been defined vaguely in computer vision. Images from different data sets are viewed as belonging to different domains. Data sets have an inherent bias, and images from a data set have certain properties that can help identify the data set [99]. However, there has been limited effort in understanding what creates this bias and on modeling the domain shift between data sets. The authors in [100] attempt to identify the domainness—a measure for domain specificity of an image.

The difficult problem of modeling domain shift in computer vision has been rarely addressed. There has been work on identifying domains from a mixture of multiple data sets and then studying transfer of knowledge between the domains [101]. Although this does not necessarily model a domain, it provides some direction toward identifying a domain through analysis. The difficulties of modeling domain shift in computer vision mostly arise due to variations in representation and not merely variations in the data being represented. The very process of imaging (camera perspective and occlusion), storage (resolution and size), and representation (color, brightness, and contrast) can lead to variations. Image background (context) is another cause for variation. Finally, the diversity in the real data itself could also lead to variations in their images.

Most domain adaptation systems create adaptive models that perform generic domain adaptation. The models are often guided by the data sets that are used. On the other hand, it might be beneficial to tailor the adaptation model to a specific variation in the data. This would, however, need a comprehensive understanding of domain shift. It might also lead to task-specific domain adaptation models based on the nature of domain-shift, leading to increased adoption of domain adaptation in real-world applications.

New data sets

Current data sets for domain adaptation are not based on any models of domain shift. They are merely data samples coming from different sources, all with the same categories. The domain difference between these data sets is attributed to the bias between the data sets, without a specific model characterizing the domain shift [99]. The domain adaptation procedures that are developed using these data sets can, therefore, be considered very generic. There is no guarantee on the performance of these procedures when applied to new problems. For example, if a domain adaptation approach were to be developed using the digit data sets *USPS* and *MNIST* [102], there is no guarantee that this procedure would work well for a domain adaptation problem with medical images. On the other hand, if a data set were to be created based on a domain shift

model, then algorithms that are developed using this data set can be applied to any domain adaptation problem where the same domain shift is observed. This is one primary reason highlighting the need for introducing new data sets for domain adaptation based on modeling domain shift.

The standard data sets for computer vision-based domain adaptation are facial expression data sets *CKPlus* [103] and *MMI* [104], digit data sets *SVHN* [105], *USPS*, and *MNIST* [102], head pose recognition data sets *PIE* [56], object recognition data sets *COIL* [56], *Office* [49], and *Office-Caltech* [30]. These data sets were created before deep learning became popular and are insufficient for training and evaluating deep learning-based domain adaptation approaches. A deep-learning model with millions of parameters requires millions of images for training. Current approaches fine-tune pretrained deep networks with these small data sets to avoid overfitting issues. The current data sets are small with a limited number of categories and limited variation. For instance, the most popular object-recognition data set *Office* has 4,110 images across 31 categories. In addition, the image statistics of the three domains in *Office* are nearly identical.

Due to some inconsistencies in the *Office* data set [33], [96], recent approaches evaluate their models using *MNIST*, *modified-MNIST*, and *SVHN* data sets [33], [81], [85], [89], [94]. Recently, a couple of data sets have been introduced for deep-learning-based domain adaptation. *Office-Home* is an object recognition data set that can be used to evaluate deep-learning algorithms for domain adaptation [65]. The *Office-Home* data set consists of four domains, with each domain containing images from 65 categories of everyday objects and a total of around 15,500 images. Castrejon et al. [8] introduce a multimodal domain adaptation data set *CMPlaces* with RGB, sketches, clipart, and textual descriptions of indoor scenes with 205 categories and millions of images. However, these data sets do not address all of the concerns regarding data sets, and newer and larger data sets are necessary based on modeling domain shift. Evolution in data sets and the evolution of models for domain shift need to complement each other.

Cross-domain generative models

Generative models like GANs are currently very popular in the computer vision research community [80]. They have a wide range of applications, including image superresolution [106], text-to-image generation [107], [108], image-to-image translation [91], and conditional image generation [109], [110]. Adversarial methods have been successfully applied in domain adaptation in the form of cross-domain image generation. Cross-domain generative models transform images from one domain into images in another domain [91]. These models can be applied to transform labeled source images into target images. These transformed images are then used to train a target classifier [81], [88], [89]. One can argue that cross-domain generative models learn a transformation, mapping the images from one domain into another. This is a unique procedure to achieve domain adaptation in the space of images. Current procedures in domain adaptation learn to adapt either the classifiers [10], [47], [49], [111] or the features [32], [50], [63].

Cross-domain generative models enable adaptation in the image space itself. Evolution in cross-domain generative models will provide robust mechanisms for accurately mapping the image spaces across domains, thereby alleviating the need for labeled target data. Classifier and feature adaptation can also be applied on top of image translation to enhance domain adaptation. Cross-domain generative modeling is a relatively new frontier in computer vision research with promising results in domain adaptation.

Joint distribution models

Current forms of adaptation merely align the marginal distributions of the source $P_S(X)$ and the target $P_T(X)$. The popular MMD measure from [52] is often applied to align the marginal distributions of the source and target data, as described in the instance selection approach [55]. The goal of domain adaptation is not merely aligning the domains but also being able to use the models trained on the source on the target. In most cases, the domain adaptive models are created for classification. It would, therefore, make more sense to align the joint distributions $P_S(X, Y)$ with $P_T(X, Y)$ rather than merely the marginal distributions. The alignment of joint distributions will make a classifier trained on the source an effective classifier for the target.

The challenge with this approach is that target labels are not available in unsupervised domain adaptation. The workaround is to impute the target data labels and refine them iteratively. There has been work in this regard as in [56], where the joint distributions are aligned in a spectral method using kernel-PCA by imputing the labels and refining them over multiple iterations. A deep-learning approach has also been attempted in this regard in [112], using a transductive approach to learn the target labels while also minimizing the joint domain discrepancy. As discussed in the section “Miscellaneous Hierarchical Methods,” Sener et al. [94] develop a deep-learning approach that imputes the labels for the target in a transductive learning environment. The aforementioned approaches use the predicted target data labels to ensure joint distribution alignment. Conditional generative models along with joint distribution alignment could usher in the next wave of domain adaptation models.

Person-centered models

Very soon, computing is going to become all pervasive. The environment is plugged with computing devices, and an average person carries quite a few smart-devices like a phone, watch, wristband, etc. Can this computing be adapted to every user? Computing that adapts to a user's needs and idiosyncrasies can be called *person-centered computing* [113]. This would mean that personal devices would model their interaction and response based on the user's needs rather than a one-size-fits-all approach where users train themselves to adapt to their devices to effectively interact with them. This paradigm, where the user and the device adapt to each other, is termed *coadaptation*.

These personalized devices will need to be designed to have core functional components making them applicable to a broad range of users. In addition, they must also have coadaptive components that help customize the device at an individual user level. The device must adapt to the user based on patterns gathered

from user interaction with the device. The learning models for coadaptation will be based on unsupervised domain adaptation, which would involve gleaning patterns from unlabeled user interaction data. There has been no work so far in the domain adaptation literature for person-centered device adaptation, and these person-centered adaptive models would make technology more accessible, especially to individuals with disabilities.

Conclusions

The current generation of artificial intelligence systems can outperform humans in a narrow set of tasks like playing chess or GO. Even though deep neural networks have contributed to the unprecedented progress of artificial intelligence research in the last few years, artificial general intelligence has, so far, been elusive. To advance artificial intelligence, computational systems will need the ability to transfer learning and progressively augment knowledge. Transfer learning paradigms like domain adaptation will be key to heralding the next generation of artificial intelligence systems. The current generation of domain adaptation models is dominated by deep-learning systems. Prior to the advent of deep learning, domain adaptation approaches had to develop adaptive computational models based on fixed representations of data. Deep-learning systems have found great success in domain adaptation because of their ability to extract domain aligned features specific to the adaptation task. This has led to a surge in domain adaptation research in recent years, and this article has provided a survey of literature in the area of domain adaptation based on deep learning. In this survey, we have outlined the concept of knowledge transfer across computational models and categorized the different paradigms of transfer and compared them with domain adaptation. This article is meant to provide a clear understanding of the scope of research in domain adaptation and also highlight the promising directions for future research.

Acknowledgments

We are grateful to the reviewers who provided highly valuable comments that have helped to improve the quality of the article. This material is based on work supported by the National Science Foundation under grant 1116360.

Authors

Hemanth Venkateswara (hemanthv@asu.edu) graduated summa cum laude with master's degrees in physics and computer science from the Sri Sathya Sai University, India, in 2005 and 2007, respectively. He is a postdoctoral researcher at the School of Computing Informatics and Decision Systems Engineering at Arizona State University. His areas of research includes machine learning and computer vision, with applications in domain adaptation using deep learning. Prior to earning his Ph.D. degree, he worked as a senior software engineer at Alcatel-Lucent Technologies, India. He is a Member of the IEEE and the Association for Computing Machinery.

Shayok Chakraborty (shayok.chakraborty@asu.edu) received his Ph.D. degree in computer science from Arizona State University (ASU) in 2013. He was an assistant research

professor of computer science at ASU and is now an assistant professor of computer science at Florida State University, Tallahassee. He has worked as a postdoctoral researcher at Intel Labs and also in the Electrical and Computer Engineering Department at Carnegie Mellon University, Pittsburgh, Pennsylvania. His research interests include computer vision and machine learning. He has published his research in premier conferences and journals such as the IEEE Conference on Computer Vision and Pattern Recognition, the Association for Computing Machinery (ACM) SIGKDD, ACM Multimedia, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *IEEE Transactions on Neural Networks and Learning Systems*, and *Pattern Recognition*. He has also extensively served in the program committee and as a reviewer of these conferences and journals. He is a Member of the IEEE and ACM.

Sethuraman Panchanathan (panch@asu.edu) received his bachelor's degree in electronics and communication engineering from the Indian Institute of Science in 1984, his master's degree in electrical engineering from Indian Institute of Technology, Madras, in 1986, and his Ph.D. degree in electrical and computer engineering from the University of Ottawa, Canada, in 1989. He leads the Knowledge Enterprise Development, Arizona State University (ASU), Tempe, which advances research, innovation, strategic partnerships, entrepreneurship, global, and economic development at ASU. In 2014, he was appointed by U.S. President Barack Obama to the U.S. National Science Board. He has also been appointed by U.S. Secretary of Commerce Penny Pritzker to the National Advisory Council on Innovation and Entrepreneurship. He is a fellow of the National Academy of Inventors, the Canadian Academy of Engineering, and the Society of Optical Engineering. He currently serves as the chair-elect in the Council on Research within the Association of Public and Land-Grant Universities. He has authored more than 425 papers in refereed journals and conferences. His research interests include human-centered multimedia computing, haptic user interfaces, person-centered tools, and ubiquitous computing technologies for enhancing the quality of life for individuals with disabilities, machine learning for multimedia applications, medical image processing, and media processor designs.

References

- [1] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Trans. Knowledge Data Eng.*, vol. 22, no. 10, pp. 1345–1359, 2010.
- [2] L. Chen, W. Li, and D. Xu, "Recognizing RGB images by learning from RGB-D data," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2014, pp. 1418–1425.
- [3] J. Hoffman, S. Gupta, J. Leong, S. Guadarrama, and T. Darrell, "Cross-modal adaptation for RGB-D detection," in *Proc. IEEE Robotics and Automation Int. Conf.*, 2016, pp. 5032–5039.
- [4] P. Koniusz, Y. Tas, and F. Porikli, "Domain adaptation by mixture of alignments of second-or higher-order scatter tensors," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2017.
- [5] B. F. Klare, S. S. Bucak, A. K. Jain, and T. Akgul, "Towards automated caricature recognition," in *Proc. IEEE Biometrics 5th IAPR Int. Conf.*, 2012, pp. 139–146.
- [6] E. J. Crowley and A. Zisserman, "In search of art," in *Proc. Workshop on Computer Vision Art Analysis, European Conf. Computer Vision (ECCV)*, 2014, pp. 54–70.
- [7] S. Saxena and J. Verbeek, "Heterogeneous face recognition with CNNs," in *Proc. European Conf. Computer Vision*, 2016.
- [8] L. Castrejon, Y. Aytar, C. Vondrick, H. Pirsiavash, and A. Torralba, "Learning aligned cross-modal representations from weakly aligned data," in

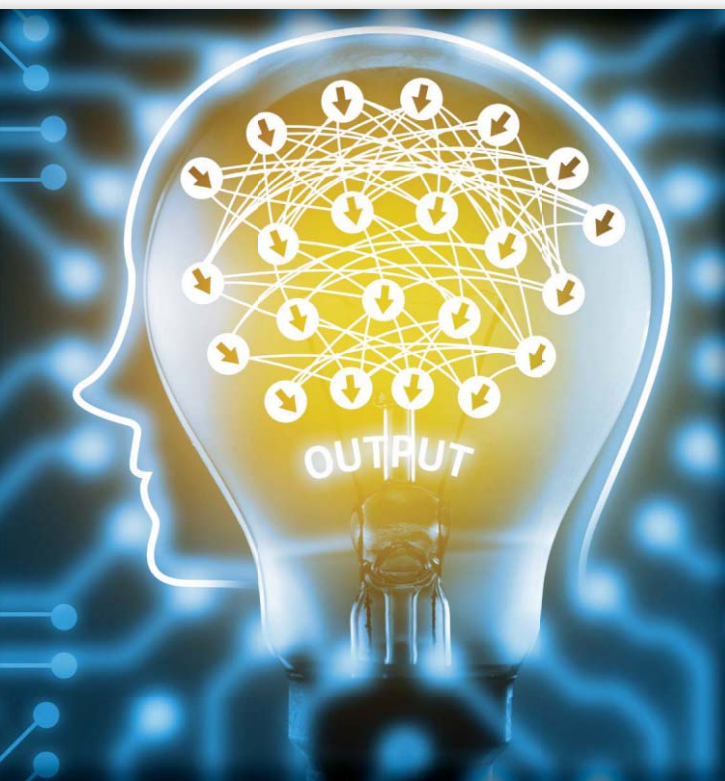
- Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2016, pp. 2940–2949.
- [9] B. Fernando, T. Tommasi, and T. Tuytelaars, “Location recognition over large time lags,” *Comput. Vision Image Understand.*, vol. 139, pp. 21–28, Oct. 2015.
- [10] L. Bruzzone and M. Marconcini, “Domain adaptation problems: A DASVM classification technique and a circular validation strategy,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 5, pp. 770–787, 2010.
- [11] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. Cambridge, MA: MIT Press, 2016.
- [12] R. Caruana, “Multitask learning,” *Mach. Learn.*, vol. 28, no. 1, pp. 41–75, 1997.
- [13] S. Thrun and L. Pratt, *Learning to Learn*. New York: Springer Science & Business Media, 2012.
- [14] R. Raina, A. Battle, H. Lee, B. Packer, and A. Y. Ng, “Self-taught learning: transfer learning from unlabeled data,” in *Proc. ACM Int. Conf. Machine Learning*, 2007, pp. 759–766.
- [15] J. Yang, K. Yu, Y. Gong, and T. Huang, “Linear spatial pyramid matching using sparse coding for image classification,” in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2009, pp. 1794–1801.
- [16] Y. Bengio, “Learning deep architectures for AI,” *Foundations Trends Mach. Learn.*, vol. 2, no. 1, pp. 1–127, 2009.
- [17] H. Lee, R. Grosse, R. Ranganath, and A. Y. Ng, “Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations,” in *Proc. ACM Int. Conf. Machine Learning*, 2009, pp. 609–616.
- [18] J. Mairal, F. Bach, J. Ponce, and G. Sapiro, “Online learning for matrix factorization and sparse coding,” *J. Mach. Learn. Res.*, vol. 11, pp. 19–60, Jan. 2010.
- [19] J. J. Heckman, “Sample selection bias as a specification error,” *Econometrica: J. Economet. Soc.*, vol. 1, no. 47, pp. 153–161, 1979.
- [20] B. Zadrozny, “Learning and evaluating classifiers under sample selection bias,” in *Proc. ACM Int. Conf. Machine Learning*, 2004, p. 114.
- [21] C. Cortes, M. Mohri, M. Riley, and A. Rostamizadeh, “Sample selection bias correction theory,” in *Proc. Int. Conf. Algorithmic Learning Theory*, 2008, pp. 38–53.
- [22] M. Dudík, S. J. Phillips, and R. E. Schapire, “Correcting sample selection bias in maximum entropy density estimation,” in *Proc. Advances in Neural Information Processing Systems Conf.*, 2005, pp. 323–330.
- [23] J. Huang, A. Gretton, K. M. Borgwardt, B. Schölkopf, and A. J. Smola, “Correcting sample selection bias by unlabeled data,” in *Proc. Advances in Neural Information Processing Systems Conf.*, 2006, pp. 601–608.
- [24] S. Thrun, “Is learning the n -th thing any easier than learning the first?” in *Proc. Advances in Neural Information Processing Systems*, 1996, pp. 640–646.
- [25] G. Fei, S. Wang, and B. Liu, “Learning cumulatively to become more knowledgeable,” in *Proc. ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining*, 2016, pp. 1565–1574.
- [26] L. Fei-Fei, R. Fergus, and P. Perona, “One-shot learning of object categories,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 4, pp. 594–611, 2006.
- [27] H. Larochelle, D. Erhan, and Y. Bengio, “Zero-data learning of new tasks,” in *Proc. AAAI Conf. Artificial Intelligence*, 2008, vol. 1, p. 3.
- [28] M. Palatucci, D. Pomerleau, G. E. Hinton, and T. M. Mitchell, “Zero-shot learning with semantic output codes,” in *Proc. Advances in Neural Information Processing Systems Conf.*, 2009, pp. 1410–1418.
- [29] R. Socher, M. Ganjoo, C. D. Manning, and A. Ng, “Zero-shot learning through cross-modal transfer,” in *Proc. Advances in Neural Information Processing Systems Conf.*, 2013 pp. 935–943.
- [30] B. Gong, Y. Shi, F. Sha, and K. Grauman, “Geodesic flow kernel for unsupervised domain adaptation,” in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2012, pp. 2066–2073.
- [31] J. Ni, Q. Qiu, and R. Chellappa, “Subspace interpolation via dictionary learning for unsupervised domain adaptation,” in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2013, pp. 692–699.
- [32] M. Long, J. Wang, G. Ding, J. Sun, and P. S. Yu, “Transfer joint matching for unsupervised domain adaptation,” in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2014, pp. 1410–1417.
- [33] K. Bousmalis, N. Silberman, D. Dohan, D. Erhan, and D. Krishnan, “Unsupervised pixel-level domain adaptation with generative adversarial networks,” in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2017.
- [34] H. Daumé III, A. Kumar, and A. Saha (2010). Frustratingly easy semi-supervised domain adaptation. *Proc. Workshop on Domain Adaptation for NLP*. [Online]. Available: <http://hal3.name/docs/#daume10easyss>
- [35] W. Li, L. Duan, D. Xu, and I. W. Tsang, “Learning with augmented features for supervised and semi-supervised heterogeneous domain adaptation,” *IEEE Trans. Pattern Analysis Machine Intell.*, vol. 36, no. 6, pp. 1134–1148, 2014.
- [36] T. Yao, Y. Pan, C.-W. Ngo, H. Li, and T. Mei, “Semi-supervised domain adaptation with subspace learning for visual recognition,” in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, 2015, pp. 2142–2150.
- [37] J. Yang, R. Yan, and A. G. Hauptmann, “Cross-domain video concept detection using adaptive SVMs,” in *Proc. ACM Int. Conf. Multimedia*, 2007, pp. 188–197.
- [38] S. J. Pan, J. T. Kwok, and Q. Yang, “Transfer learning via dimensionality reduction,” in *Proc. AAAI Conf. Artificial Intelligence*, 2008, vol. 8, pp. 677–682.
- [39] E. Tzeng, J. Hoffman, T. Darrell, and K. Saenko, “Simultaneous deep transfer across domains and tasks,” in *Proc. IEEE Int. Conf. Computer Vision*, 2015, pp. 4068–4076.
- [40] Q. Sun, R. Chattopadhyay, S. Panchanathan, and J. Ye, “A two-stage weighting framework for multi-source domain adaptation,” in *Proc. Advances in Neural Information Processing Systems Conf.*, 2011, pp. 505–513.
- [41] Y. Mansour, M. Mohri, and A. Rostamizadeh, “Domain adaptation with multiple sources,” in *Proc. Advances in Neural Information Processing Systems Conf.*, 2009, pp. 1041–1048.
- [42] R. Chattopadhyay, Q. Sun, W. Fan, I. Davidson, S. Panchanathan, and J. Ye, “Multisource domain adaptation and its application to early detection of fatigue,” *ACM Trans. Knowledge Discovery Data*, vol. 6, no. 4, pp. 18, 2012.
- [43] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “Imagenet classification with deep convolutional neural networks,” in *Proc. Advances in Neural Information Processing Systems Conf.*, 2012, pp. 1097–1105.
- [44] D. G. Lowe, “Object recognition from local scale-invariant features,” in *Proc. IEEE Int. Conf. Computer Vision*, 1999, vol. 2, pp. 1150–1157.
- [45] N. Dalal and B. Triggs, “Histograms of oriented gradients for human detection,” in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2005, vol. 1, pp. 886–893.
- [46] B. Julesz, “Textons, the elements of texture perception, and their interactions,” *Nature*, vol. 290, no. 5802, pp. 91–97, 1981.
- [47] Y. Aytar and A. Zisserman, “Tabula rasa: Model transfer for object category detection,” in *Proc. IEEE Int. Conf. Computer Vision*, 2011, pp. 2252–2259.
- [48] J. Hoffman, E. Rodner, J. Donahue, K. Saenko, and T. Darrell, “Efficient learning of domain-invariant image representations,” in *Proc. Int. Conf. Learning Representations*, 2013.
- [49] K. Saenko, B. Kulis, M. Fritz, and T. Darrell, “Adapting visual category models to new domains,” in *Proc. European Conf. Computer Vision*, 2010, pp. 213–226.
- [50] B. Fernando, A. Habrard, M. Sebban, and T. Tuytelaars, “Unsupervised visual domain adaptation using subspace alignment,” in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2013, pp. 2960–2967.
- [51] B. Sun, J. Feng, and K. Saenko, “Return of frustratingly easy domain adaptation,” in *Proc. Conf. Artificial Intelligence (AAAI)*, 2016.
- [52] A. Gretton, K. M. Borgwardt, M. Rasch, B. Schölkopf, and A. J. Smola, “A kernel method for the two-sample-problem,” in *Proc. Advances in Neural Information Processing Systems Conf.*, 2007, vol. 19, p. 513.
- [53] L. Duan, I. W. Tsang, D. Xu, and S. J. Maybank, “Domain transfer SVM for video concept detection,” in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2009, pp. 1375–1381.
- [54] W.-S. Chu, F. De la Torre, and J. F. Cohn, “Selective transfer machine for personalized facial action unit detection,” in *Proc. IEEE Conf. Comput. Vision and Pattern Recognition*, 2013, pp. 3515–3522.
- [55] B. Gong, K. Grauman, and F. Sha, “Connecting the dots with landmarks: Discriminatively learning domain-invariant features for unsupervised domain adaptation,” in *Proc. ACM Int. Conf. Machine Learning*, 2013, pp. 222–230.
- [56] M. Long, J. Wang, G. Ding, J. Sun, and P. S. Yu, “Transfer feature learning with joint distribution adaptation,” in *Proc. ACM Int. Conf. Machine Learning*, 2013, pp. 2200–2207.
- [57] S. J. Pan, I. W. Tsang, J. T. Kwok, and Q. Yang, “Domain adaptation via transfer component analysis,” *IEEE Trans. Neural Netw.*, vol. 22, no. 2, pp. 199–210, 2011.
- [58] R. Gopalan, R. Li, and R. Chellappa, “Domain adaptation for object recognition: An unsupervised approach,” in *Proc. IEEE Int. Conf. Computer Vision*, 2011, pp. 999–1006.
- [59] O. Beijbom, “Domain adaptations for computer vision applications,” *arXiv Preprint*, arXiv:1211.4860, 2012.
- [60] V. M. Patel, R. Gopalan, R. Li, and R. Chellappa, “Visual domain adaptation: A survey of recent advances,” *IEEE Signal Process. Mag.*, vol. 32, no. 3, pp. 53–69, 2015.
- [61] J. Blitzer, K. Crammer, A. Kulesza, F. Pereira, and J. Wortman, “Learning bounds for domain adaptation,” in *Proc. Advances in Neural Information Processing Systems Conf.*, 2008, pp. 129–136.
- [62] S. Ben-David, J. Blitzer, K. Crammer, A. Kulesza, F. Pereira, and J. W. Vaughan, “A theory of learning from different domains,” *Mach. Learn.*, vol. 79, no. 1–2, pp. 151–175, 2010.
- [63] M. Long, Y. Cao, J. Wang, and M. Jordan, “Learning transferable features with deep adaptation networks,” in *Proc. ACM Int. Conf. Machine Learning*, 2015, pp. 97–105.

- [64] B. Sun and K. Saenko, "Deep Coral: Correlation alignment for deep domain adaptation," in *Proc. European Conf. Computer Vision*, 2016, pp. 443–450.
- [65] H. Venkateswara, J. Eusebio, S. Chakraborty, and S. Panchanathan, "Deep hashing network for unsupervised domain adaptation," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2017.
- [66] A. S. Razavian, H. Azizpour, J. Sullivan, and S. Carlsson, "CNN features off-the-shelf: an astounding baseline for recognition," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2014, pp. 512–519.
- [67] ImageNet large scale visual recognition challenge 2013. [Online]. Available: <http://www.image-net.org/challenges/LSVRC/2013/>
- [68] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson, "How transferable are features in deep neural networks?" in *Proc. Advances in Neural Information Processing Systems Conf.*, 2014, pp. 3320–3328.
- [69] M. Oquab, L. Bottou, I. Laptev, and J. Sivic, "Learning and transferring mid-level image representations using convolutional neural networks," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2014, pp. 1717–1724.
- [70] J. Donahue, Y. Jia, O. Vinyals, J. Hoffman, N. Zhang, E. Tzeng, and T. Darrell, "DeCAF: A deep convolutional activation feature for generic visual recognition," in *Proc. ACM Int. Conf. Machine Learning*, 2014, vol. 32, pp. 647–655.
- [71] X. Glorot, A. Bordes, and Y. Bengio, "Domain adaptation for large-scale sentiment classification: A deep learning approach," in *Proc. ACM Int. Conf. Machine Learning*, 2011, pp. 513–520.
- [72] E. Tzeng, J. Hoffman, N. Zhang, K. Saenko, and T. Darrell, "Deep domain confusion: Maximizing for domain invariance," *arXiv Preprint*, arXiv:1412.3474, 2014. [Online]. Available: <http://arxiv.org/abs/1412.3474>
- [73] E. Tzeng, J. Hoffman, T. Darrell, and K. Saenko, "Simultaneous deep transfer across domains and tasks," in *Proc. IEEE Int. Conf. Computer Vision*, 2015, pp. 4068–4076.
- [74] A. Gretton, D. Sejdinovic, H. Strathmann, S. Balakrishnan, M. Pontil, K. Fukumizu, and B. K. Sriperumbudur, "Optimal kernel choice for large-scale two-sample tests," in *Proc. Advances in Neural Information Processing Systems Conf.*, 2012, pp. 1205–1213.
- [75] M. Long, H. Zhu, J. Wang, and M. I. Jordan, "Unsupervised domain adaptation with residual transfer networks," in *Proc. Advances in Neural Information Processing Systems Conf.*, 2016, pp. 136–144.
- [76] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.
- [77] C. Wu, W. Wen, T. Afzal, Y. Zhang, Y. Chen, and H. Li, "A compact DNN: Approaching googlenet-level accuracy of classification and domain adaptation," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2017.
- [78] M. Ghifary, W. B. Kleijn, M. Zhang, D. Balduzzi, and W. Li, "Deep reconstruction-classification networks for unsupervised domain adaptation," in *Proc. European Conf. Computer Vision*, 2016, pp. 597–613.
- [79] A. Rozantsev, M. Salzmann, and P. Fua, "Beyond sharing weights for deep domain adaptation," *arXiv Preprint*, arXiv:1603.06432, 2016. [Online]. Available: <http://arxiv.org/abs/1603.06432>
- [80] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Proc. Advances in Neural Information Processing Systems Conf.*, 2014, pp. 2672–2680.
- [81] Y. Taigman, A. Polyak, and L. Wolf, "Unsupervised cross-domain image generation," in *Proc. Int. Conf. Learning Representations*, 2017.
- [82] K. Kamnitsas, C. Baumgartner, C. Ledig, V. F. Newcombe, J. P. Simpson, A. D. Kane, D. K. Menon, A. Nori, A. Criminisi, D. Rueckert, et al., "Unsupervised domain adaptation in brain lesion segmentation with adversarial networks," in *Proc. Int. Conf. Information Processing Medical Imaging*, 2016, pp. 597–609.
- [83] E. Tzeng, J. Hoffman, K. Saenko, and T. Darrell, "Adversarial discriminative domain adaptation," *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2017.
- [84] X. Peng and K. Saenko, "Synthetic to real adaptation with deep generative correlation alignment networks," *arXiv Preprint*, arXiv:1701.05524, 2017.
- [85] P. Russo, F. M. Carlucci, T. Tommasi, and B. Caputo, "From source to target and back: symmetric bi-directional adaptive GAN," *arXiv Preprint*, arXiv:1705.08824, 2017.
- [86] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, and V. Lempitsky, "Domain-adversarial training of neural networks," *J. Mach. Learn. Res.*, vol. 17, no. 59, pp. 1–35, 2016.
- [87] H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, and M. Marchand, "Domain-adversarial neural networks," *arXiv Preprint*, arXiv:1412.4446, 2014.
- [88] M.-Y. Liu and O. Tuzel, "Coupled generative adversarial networks," in *Proc. Advances in Neural Information Processing Systems Conf.*, 2016, pp. 469–477.
- [89] M.-Y. Liu, T. Breuel, and J. Kautz, "Unsupervised image-to-image translation networks," *arXiv Preprint*, arXiv:1703.00848, 2017.
- [90] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," *arXiv Preprint*, arXiv:1312.6114, 2013.
- [91] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," *arXiv Preprint*, arXiv:1611.07004, 2016.
- [92] S. Chopra, S. Balakrishnan, and R. Gopalan, "DLID: Deep learning for domain adaptation by interpolating between domains," in *Proc. ICML Workshop on Challenges in Representation Learning*, 2013.
- [93] J. Hu, J. Lu, and Y.-P. Tan, "Deep transfer metric learning," in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, 2015, pp. 325–333.
- [94] O. Sener, H. O. Song, A. Saxena, and S. Savarese, "Learning transferable representations for unsupervised domain adaptation," in *Proc. Advances Neural Information Processing Systems Conf.*, 2016, pp. 2110–2118.
- [95] C. Sun, S. Shetty, R. Sukthankar, and R. Nevatia, "Temporal localization of fine-grained actions in videos by domain transfer from web images," in *Proc. ACM Int. Conf. Multimedia*, 2015, pp. 371–380.
- [96] K. Bousmalis, G. Trigeorgis, N. Silberman, D. Krishnan, and D. Erhan, "Domain separation networks," in *Proc. Advances in Neural Information Processing Systems Conf.*, 2016, pp. 343–351.
- [97] H. Shimodaira, "Improving predictive inference under covariate shift by weighting the log-likelihood function," *J. Statist. Plan. Inference*, vol. 90, no. 2, pp. 227–244, 2000.
- [98] H. Venkateswara, S. Chakraborty, and S. Panchanathan, "Nonlinear embedding transform for unsupervised domain adaptation," in *Proc. European Conf. Computer Vision Workshops*, 2016, pp. 451–457.
- [99] A. Torralba and A. A. Efros, "Unbiased look at data set bias," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2011, pp. 1521–1528.
- [100] T. Tommasi, M. Lanzi, P. Russo, and B. Caputo, "Learning the roots of visual domain shift," in *Proc. IEEE Int. Conf. Computer Vision Workshops*, 2016, pp. 475–482.
- [101] B. Gong, K. Grauman, and F. Sha, "Reshaping visual data sets for domain adaptation," in *Proc. Advances in Neural Information Processing Systems Conf.*, 2013, pp. 1286–1294.
- [102] K. Jarrett, K. Kavukcuoglu, Y. Lecun et al., "What is the best multi-stage architecture for object recognition?" in *Proc. IEEE Int. Conf. Computer Vision*, 2009, pp. 2146–2153.
- [103] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews, "The extended cohn-kanade data set (CK+): A complete data set for action unit and emotion-specified expression," in *Proc. IEEE Conf. Comput. Vision and Pattern Recognition Workshops*, 2010, pp. 94–101.
- [104] M. Pantic, M. Valstar, R. Rademaker, and L. Maat, "Web-based database for facial expression analysis," in *Proc. IEEE Conf. Multimedia and Expo*, 2005.
- [105] Y. Netzer, T. Wang, A. Coates, A. Bissacco, B. Wu, and A. Y. Ng, "Reading digits in natural images with unsupervised feature learning," in *Proc. Workshops Advances in Neural Information Processing Systems Conf.*, 2011.
- [106] C. Ledig, L. Theis, F. Huszár, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, et al., "Photo-realistic single image super-resolution using a generative adversarial network," *arXiv Preprint*, arXiv:1609.04802, 2016. [Online]. Available: <http://arxiv.org/abs/1609.04802>
- [107] S. Reed, Z. Akata, X. Yan, L. Logeswaran, B. Schiele, and H. Lee, "Generative adversarial text to image synthesis," in *Proc. 33rd Int. Conf. Machine Learning*, vol. 3, 2016, pp. 1060–1069.
- [108] H. Zhang, T. Xu, H. Li, S. Zhang, X. Huang, X. Wang, and D. Metaxas, "StackGAN: Text to photo-realistic image synthesis with stacked generative adversarial networks," *arXiv Preprint*, arXiv:1612.03242, 2016. [Online]. Available: <http://arxiv.org/abs/1612.03242>
- [109] A. Nguyen, J. Yosinski, Y. Bengio, A. Dosovitskiy, and J. Clune, "Plug & play generative networks: Conditional iterative generation of images in latent space," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2017.
- [110] X. Chen, Y. Duan, R. Houthoofd, J. Schulman, I. Sutskever, and P. Abbeel, "InfoGAN: Interpretable representation learning by information maximizing generative adversarial nets," in *Proc. Advances in Neural Information Processing Systems Conf.*, 2016, pp. 2172–2180.
- [111] H. Venkateswara, P. Lade, J. Ye, and S. Panchanathan, "Coupled support vector machines for supervised domain adaptation," in *Proc. ACM Int. Conf. Multimedia*, 2015, pp. 1295–1298.
- [112] M. Long, J. Wang, and M. I. Jordan, "Deep transfer learning with joint adaptation networks," in *Proc. 34th Int. Conf. Machine Learning*, 2017, pp. 2208–2217.
- [113] S. Panchanathan, T. McDaniel, and V. Balasubramanian, "Person-centered accessible technologies: Improved usability and adaptation through inspirations from disability research," in *Proc. ACM Workshop User Experience in e-learning and Augmented Technologies Education*, 2012, pp. 1–6.

Jongyoo Kim, Hui Zeng, Deepti Ghadiyaram,
Sanghoon Lee, Lei Zhang, and Alan C. Bovik

Deep Convolutional Neural Models for Picture-Quality Prediction

Challenges and solutions to data-driven image quality assessment



©ISTOCKPHOTO.COM/ZAPP2PHOTO

Convolutional neural networks (CNNs) have been shown to deliver standout performance on a wide variety of visual information processing applications. However, this rapidly developing technology has only recently been applied with systematic energy to the problem of picture-quality prediction, primarily because of limitations imposed by a lack of adequate ground-truth human subjective data. This situation has begun to change with the development of promising data-gathering methods that are driving new approaches to deep-learning-based perceptual picture-quality prediction. Here, we assay progress in this rapidly evolving field, focusing, in particular, on new ways to collect large quantities of ground-truth data and on recent CNN-based picture-quality prediction models that deliver excellent results in a large, real-world, picture-quality database.

Introduction

Recent years have seen significant efforts applied to the development of successful models and algorithms that can automatically and accurately predict the perceptual quality of two-dimensional (2-D) and three-dimensional (3-D) digital images and videos as reported by human viewers [1]. Concurrently, there has been a tremendous surge of work on exploiting large data sets of annotated image data as inputs to deep neural networks (NNs) toward solving such challenging problems as image classification and recognition [2]. These efforts have often produced dramatic improvement relative to the state of the art. It is perhaps unsurprising that very deep models, having universal representation capability, should produce excellent results when trained on massive data sets using fast graphical computing architectures. Nevertheless, the generalization capability of these models is remarkable.

Yet, until recently, there has been limited effort directed toward optimizing picture-quality prediction models using deep networks, although, in principal, this could also lead to greatly improved performance. The practical significance of the problem and the relative ease of implementing algorithms learned on deep architectures make this a compelling topic. The explosive consumption of visual media in recent years, owing to advances in

Digital Object Identifier 10.1109/MSP.2017.2736018
Date of publication: 13 November 2017

digital camera technology, digital television, streaming video services, and social media applications, is driving a critical need for improved picture-quality monitoring. The pipelines from picture content generation to consumption are fraught with numerous sources of distortions, including blur, noise, and artifacts arising from such processes as compression, scaling, format conversion, color modification, and so on. Multiple interacting distortions are often present, which greatly complicates the problem. Picture-quality models that can accurately predict human-quality judgments can be used to greatly improve consumer satisfaction via automatic monitoring of the qualities of massively distributed pictures and videos and to perceptually benchmark picture processing algorithms such as compression engines, denoising algorithms, and superresolution systems that substantially affect viewed picture quality. While many successful picture-quality models have been devised, the problem is hardly solved, and there remains significant scope for improvement [3]. Deep-learning engines offer a potentially powerful framework for achieving sought-after gains in performance; however, as we will explain, progress has been limited by a lack of adequate amounts of distorted picture data and ground-truth subjective quality scores, which are much harder to acquire than other kinds of labeled image data. Furthermore, typical data-augmentation strategies such as those used for machine vision are of little use on this problem.

Perceptual picture-quality prediction

Picture-quality models are generally classified according to whether a pristine reference image is available for comparison. Full-reference and reduced-reference models assume that a reference is available; otherwise, the model is no-reference, or blind. Reference models are generally deployed when a process is applied to an original image, such as compression or enhancement. No-reference models are applied when the quality of an original image is suspect, as in a source inspection process, or when analyzing the output of a digital camera. Generally, no-reference prediction is a more difficult problem.

Both reference and no-reference picture-quality models rely heavily on principles of computational visual neuroscience and/or on highly regular models of natural picture statistics [1]. Heretofore, the most successful no-reference models have relied on powerful but shallow regression engines to achieve results that approach the prediction accuracy of reference-quality predictors.

Deep learning and CNNs

Deep learning has had a transformative impact on such difficult problems as speech recognition and image classification, achieving improvements in performance that are significantly superior to those obtained using conventional model-based methods optimized using shallower networks. In particular, most of the top-ranked image recognition and classification systems have been optimized using CNNs. One of the principal advantages of deep-learning models are the remarkable generalization capabilities that they can acquire when they are trained on large-scale labeled data sets. Models learned using conventional machine-learning methods are heavily dependent on the determination

and discrimination capability of sophisticated training features. By contrast, deep-learning models employ multiple levels of linear and nonlinear transformations to generate highly general data representations, thereby greatly decreasing dependence on the selection of features, which are often reduced simply to raw pixel values [2], [4]. In particular, deep CNNs optimized for image recognition and classification have greatly outperformed conventional methods. Open-source frameworks such as TensorFlow [5] have also greatly increased the accessibility of deep-learning models, and their application to diverse image processing and analysis problems has greatly expanded.

Unlike traditional NNs, CNNs can be adapted to effectively process high-dimensional, raw image data such as red, green, and blue (RGB) pixel values. Two key ideas underlie a convolutional layer: local connectivity and shared weights. Each output neuron of a convolutional layer is computed only on a locally connected subset of the input, called a *local receptive field* (drawing from vision science terminology). However, by stacking multiple convolutional layers, the effective receptive fields may enlarge to capture global picture characteristics. Usually, the parameters in a layer (i.e., filter weights) are shared across the entire visual field to limit their number. A common conception is that CNNs resemble processing by neurons in visual cortex. This idea largely arises from the observation that, in deep convolutional networks deploying many layers of adaptation on images, early layers of processing often resemble the profiles of low-level cortical neurons in visual area V1, i.e., directionally tuned Gabor filters [6], or neurons in visual area V2 implicated in assembling low-level representations of image structure [7]. At early layers of network abstraction, these perceptual attributes make them appealing tools for adaption to the picture-quality prediction problem.

An example of a CNN structure similar to those studied here is shown in Figure 1, which also illustrates the kernels learned and the feature maps obtained when the model is trained for the picture-quality prediction task. Generally, a CNN model consists of several convolutional layers followed by fully connected layers. Some convolutional layers may be followed by pooling layers, which reduce the sizes of the feature maps. The fully connected layers are essentially traditional NNs, where all of the neurons in a previous layer are connected to every neuron in a current layer.

Motivated by the great success of CNNs on numerous image analysis applications, we comprehensively review and analyze the use of deep CNNs on the picture-quality prediction problem.

Overview of the problem

Machine learning has played an important role in the development of modern picture-quality models. Although these models have been largely driven by features drawn from meaningful quantitative perceptual models, mapping them against the wide variety of generally nonlinear, often commingled, and poorly understood distortions that occur in practice is a formidable problem. Sophisticated, yet shallow mapping engines such as support vector regressors (SVR), have produced good prediction results (against human-quality opinions), yet there remains substantial room for improvement [3], which greatly motivates the study of

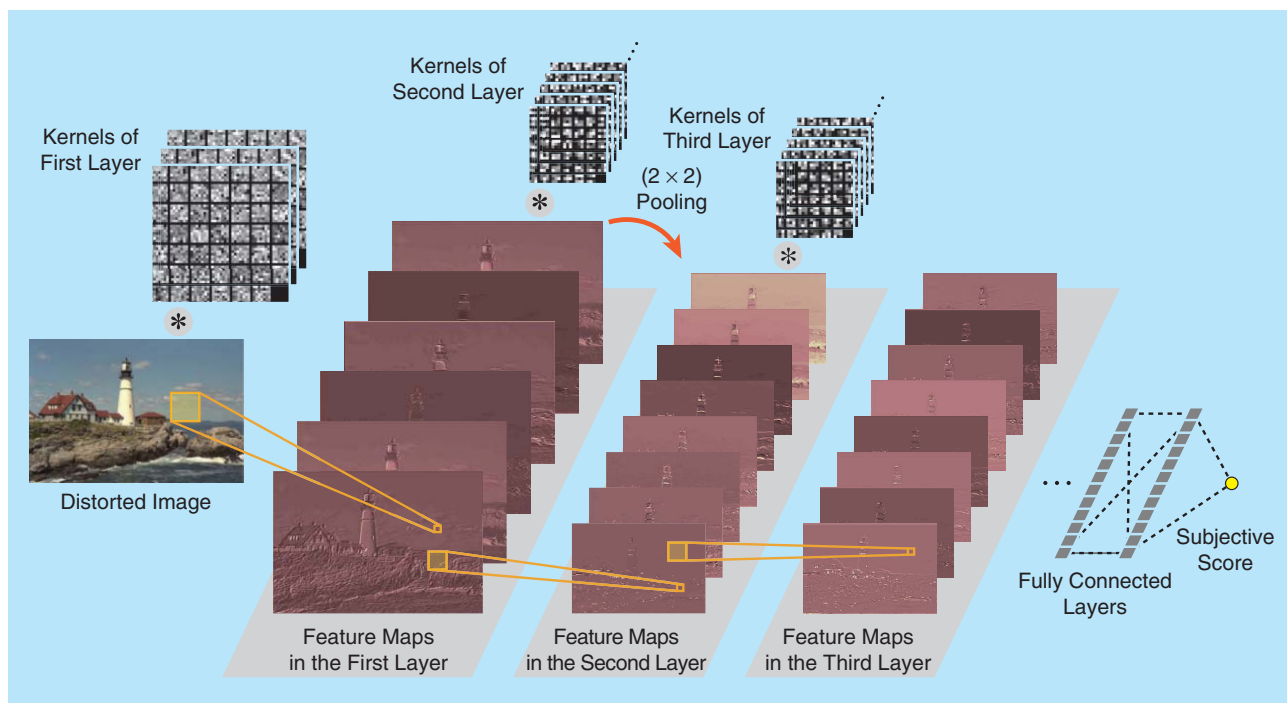


FIGURE 1. An example of a CNN structure for no-reference picture-quality prediction. The model consists of several convolutional layers followed by a few fully connected layers. An activation function is applied at each output of the NN processing flow.

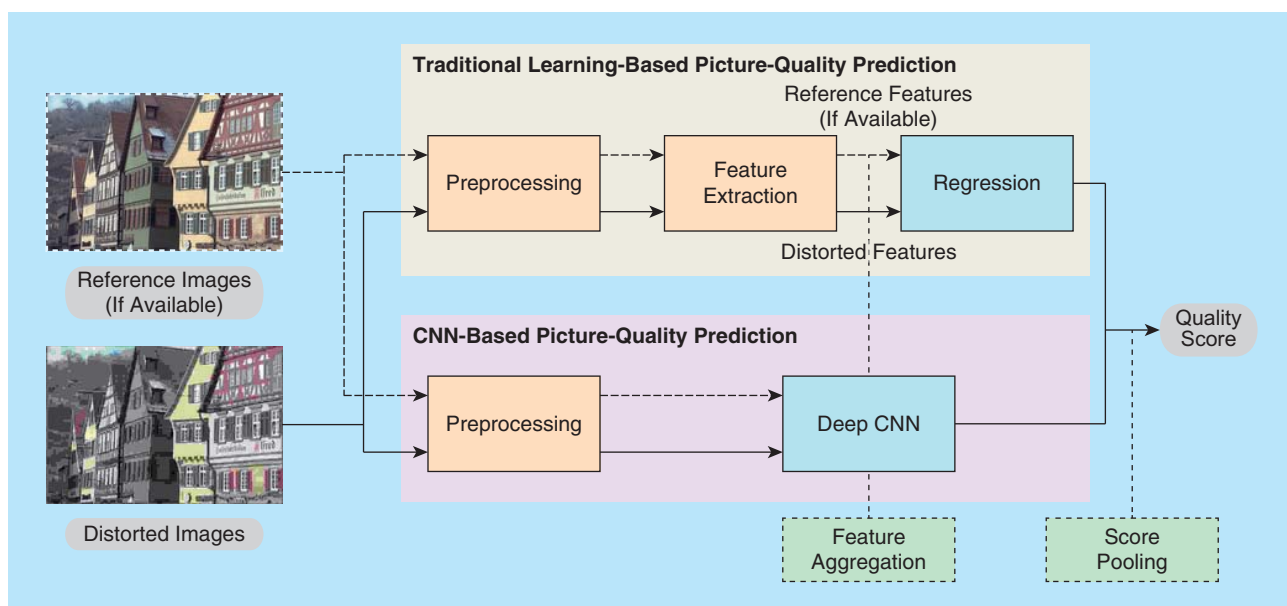


FIGURE 2. Flowchart comparisons of traditional learning-based and CNN-based reference and no-reference picture-quality models. Blue boxes indicate learning processes.

deep learners for this problem. Figure 2 shows conceptual flow diagrams of reference and no-reference learned picture-quality predictors. A major difference of deep CNN models is the lack of a feature extraction stage, although preprocessing steps may still be put to effective use. In a deep CNN, features conducive to effective picture-quality prediction are ostensibly learned by the network during the training process. The preprocessing stages may include, for example, color conversion, local debiasing, local

(divisive) normalization, or a domain transformation to sparsify [8] or reduce redundancy in the data.

Most popular learned picture-quality prediction models operate by regressing an extracted perceptual feature vector onto recorded subjective scores. Typically, shallow regressors such as SVRs, general regression NNs, or random forests have been used [9]–[11]. A deep CNN model can instead alternate feature extraction and regression stages. High-dimensional input data (raw or

preprocessed pixel values) can be fed into the CNN, and, over many iterations or epochs of training on a large data set, useful image representations are learned automatically. In the early layers of a deep CNN, low-level encoding or sparsifying features are learned, possibly followed by intermediate descriptors of feature correlations [7]. In the deeper layers, the learned features contain more abstract information that can capture relationships between image distortions and human perceptions of them. In a CNN, differentiable feature aggregation or pooling stages are interspersed with feature extraction and regression stages, enabling effective end-to-end optimization. However, despite significant successes on a wide array of other image analysis problems, the application of deep learning networks to the picture-quality prediction problem has been complicated by a significant obstacle, which is a lack of an adequate amount of perceptual training data, including accurate local ground-truth scores.

The performance of deep-learning models generally depends heavily on the size of the available training data set(s). Currently available legacy, public-domain, subjective picture-quality databases such as LIVE IQA [12] and TID2013 [13] are far too small to effectively train deep learning models. For example, the LIVE IQA and TID2013 databases each contains fewer than 30 unique image contents and no more than 24 different types of distortions per image, all of which are synthetic [This is as applied to pristine images by a database designer. Algorithm-generated distortions such as Gaussian blur (GB), noise, mean shifts, and so on, contained in these databases are poor models of picture impairments that actually arise in consumer digital photographs. Even JPEG/JPEG2000-coded images are created using much more liberal amounts and spreads of compression (to create perceptual separations) than those produced by real image capture devices.] Even the recent LIVE “In the Wild” Challenge Database (hereafter, LIVE Challenge) [3], the largest available resource in most dimensions (with nearly 1,200 unique pictures, each afflicted by a unique, unknown combination of highly diverse authentic distortions and judged by more than 350,000 unique human subjects) is of insufficient size, although it provides an excellent challenge for any no-reference model. By comparison, image recognition data sets such as ImageNet [14] contain tens of millions of labeled images. Creating larger subjective quality data sets is a formidable problem. Controlled laboratory studies like [12] and [13] are out of the question, and even the crowdsourced study in [3] exhausted the pool of high-quality human subjects available on Amazon Mechanical Turk.

Obtaining adequate quantities of reliable human subjective labels remains a very difficult problem. Unlike the binary (yes/no) confirmations of automatically generated labels that are delivered by online human subjects, as used in the construction of object recognition data sets like ImageNet [2], each of which might be generated in a second or less, collecting human-quality judgments is a complex, time-consuming psychometric task that is as much about assessing each subject’s response, as it is about the quality of the labeling the images. The human subjects determine an internal judgment of the overall quality of each image after holistically scrutinizing it, then record each of their judgments on a continuous, sliding subjective-quality scale, while consciously discounting factors such as image content or photographic aes-

thetics. This highly engaging task requires dozens or even hundreds of human-quality raters to spend 5–10 s on each image. Each subject’s overall session is time-limited, to avoid reductions in attention and performance arising from vision fatigue.

Common strategies for attacking this labeled image paucity are data augmentation techniques, which seek to multiply the effective volume of image data via rotations, cropping, reflections, and so on. Unfortunately, with the likely exception of horizontal reflections, which we use later, applying these kinds of transformations to an image will generally significantly change its perceived quality. While generating a large amount of picture content is simple, ensuring adequate distortion diversity and realism is much harder.

In another common strategy, the images used for training are divided into many small patches. However, this approach produces another problem—distinct local ground-truth subjective labels are not available for each of the patches. In every experimental scenario to date, human subjects supply a single scalar subjective score on each global image. Since images, distortions of images, and human perceptions of both are all highly non-stationary, the scores that subjects would apply to a local image patch will generally differ greatly from those applied to the entire image. Obtaining human judgments of local image patch quality is not practical, as it would greatly increase the overhead of acquiring human scores.

One way to try to overcome the lack of an adequate training data set is to utilize unsupervised learning, e.g., by training a restricted Boltzmann machine or an autoencoder [4] with convolutional layers. With an unsupervised model, it is possible to train deep NN models on very large data sets having no ground-truth labels. However, picture-quality prediction is a subtle problem that involves modeling detailed interactions between distortion and content. Conversely, unsupervised models that are designed to work well on tasks such as image recognition, may succeed in part by learning to promote gross shape-related features, while suppressing small variations. For example, a denoising autoencoder can be trained to reconstruct an original image from a noisy one by enforcing robustness against small corruptions of the input data or adding a regularization term to the objective function. By contrast, the representations learned by a picture-quality predictor must be particularly sensitive to local and global degrees of distortion as well as perceived interactions between content and distortion. Successful, generalizable, deep unsupervised picture-quality prediction models have not yet been reported.

The need for large-scale subjective picture-quality data is underlined by the fact that the perception of picture distortions engages multiple complex processes along the visual pathway, including bandpass, multiscale, and directional decompositions [6]; local nonlinearities; and normalization mechanisms. For example, contrast masking [15], whereby the spatially localized energy of image content can reduce or eliminate the visibility of distortions, is well explained by a local cortical divisive normalization model [16]. Successful reference and no-reference picture-quality models [9], [10], [15], [17] approximate these perceptual mechanisms by various models. However, errors in these approximations, along with a lack of information describing other

relevant, perhaps higher-level processes, still limit their prediction efficacy [3]. Traces of such human response properties exist and are embedded in human subject data. This suggests that they might be unraveled by a deep network served by enough data.

Conventional learning-based picture-quality predictors

The most successful reference picture-quality predictors, such as those deployed by the television industry, such as the Emmy-winning structured similarity (SSIM) model [15] and the visual information fidelity (VIF) index [18] (a core element of the VMAF processing system that quality-controls all Netflix content encodes) are not learned models but instead compute similarity or error measures modulated by perceptual criteria in some manner. Performance is high since a reference error, whether implicit or explicit, is available to be analyzed using perceptual models. No-reference models operate without the benefit of an implied error signal, so their design has relied heavily on machine learning. Broadly, these models deploy perceptually relevant, low-level feature extraction mechanisms based on simple, yet highly regular, parametric models of good-quality pictures. These natural scene statistics (NSS) models are predictably altered by the presence of distortions [18]. Simply stated, high-quality images subjected to bandpass filtering, followed by local energy normalization, become substantially decorrelated and Gaussianized, while distorted images tend not to obey this model (although this is not always the case on authentically distorted pictures, as demonstrated in [3]). Picture-quality prediction models of this type have been developed in the wavelet [18], discrete cosine transform, sparse [8] and spatial domains [9], and have been applied to video signals using natural bandpass space-time video statistics models [19], [20]. The FRIQUEE model [21] achieved state-of-the-art performance on the LIVE Challenge database [3] by regressing on a “bag” of NSS features drawn from diverse color spaces and perceptually motivated transform domains.

There have also been recent attempts to apply other, earlier types of deep-learning models to the no-reference picture-quality prediction problem. For example, Hou et al. trained a deep belief network on wavelet domain NSS features to classify distorted images into five discrete score categories [17], and Li et al. regressed shearlet NSS features onto subjective scores using a stacked autoencoder [22]. These models generally used handcrafted feature inputs, were not trained via end-to-end optimization, and achieved less impressive gains in performance.

CNN-based picture-quality prediction

CNN-based no-reference picture-quality models

As mentioned previously, several CNN-based picture-quality prediction models have attempted to use patch-based labeling to increase the set of informative (ground-truth) training samples. Generally, two types of training approaches have been used: patchwise and imagewise, as depicted in Figure 3. In the former, each image patch is independently regressed onto its target. In the latter, the patch features or predicted scores are aggregated or pooled, then regressed onto a single ground-truth subjective score.

The first application of a spatial CNN model to the picture-quality prediction problem was reported in [23], wherein a high-dimensional input image was directly fed into a shallow CNN model without finding handcrafted features. To obtain more data, each input image was subdivided into small patches as a method of data augmentation, each being assigned the same subjective-quality score during training. Following prior successful NSS-based models [9], [18], this method applies a process of local divisive normalization on each input image and uses both maximum (max) and minimum (min) pooling to reduce the feature maps. Patchwise training was used, and, during application, the predicted patch scores were averaged to obtain a single picture-quality score.

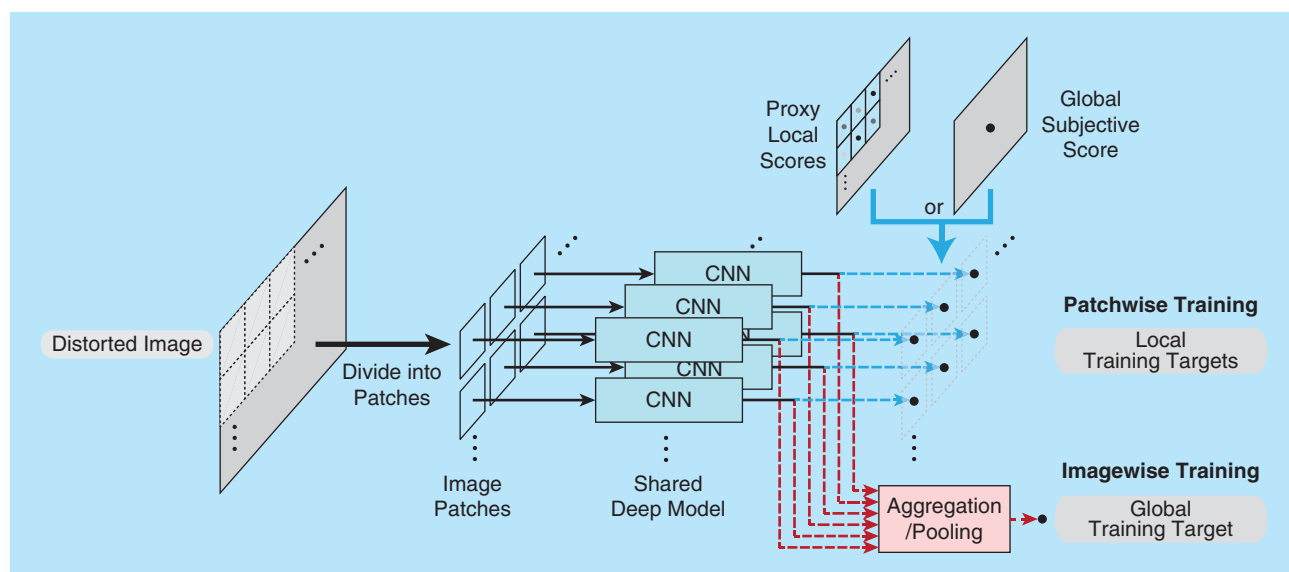


FIGURE 3. Patchwise and imagewise strategies used to train patch-based picture-quality prediction models. First, an input image is partitioned into patches; then, each is fed into the same CNN model. In patchwise training, a proxy local score or global subjective score is used as a training target for each input patch. In imagewise training, extracted features or scores are aggregated, then regressed onto a single, global subjective score.

Li et al. utilized a deep CNN model that was pretrained on the ImageNet data set [24]. A network-in-network (NiN) structure was used to enhance the abstraction ability of the model. The final layer of the pretrained model was replaced by regression layers, which mapped the learned features onto subjective scores. As in [23], image patches were regressed onto identical subjective-quality scores during training.

The labeling of local patches with global subjective-quality scores during training may be problematic. While the reported prediction accuracy of this model was competitive with that of handcrafted feature-based quality prediction models, it is not reasonable to expect local image quality to closely agree with global subjective scores, even when synthetic distortions are applied homogeneously. Picture quality is inevitably space-varying because of the high degree of nonstationarity of picture contents and the complex perceptual interactions that occur between content and distortions (such as masking). A variety of training strategies have been studied as solutions to this problem.

Bosse et al. deployed a deeper, 12-layer CNN model fed only by raw RGB image patches to learn a no-reference picture-quality model [25]. They proposed two training strategies: patchwise training (similar to [23]) and weighted average patch aggregation, whereby the relative importance of each patch was weighted by training on a subnetwork. The overall loss function was optimized in an end-to-end manner. The authors reported state-of-the-art prediction accuracies on the major synthetic-distortion picture-quality databases.

To overcome overfitting problems that can arise from a lack of adequate local ground-truth scores, several authors have suggested training deep CNN models in two separate stages: a pretraining stage, using a large number of algorithm-generated proxy ground-truth quality scores, followed by a stage of regression onto a smaller set of subjective scores. For example, [26] describes a two-stage CNN-based no-reference-quality prediction model, whereby local quality scores generated by a full-reference algorithm are used as proxy patch labels in the first stage of training. In the second stage, the feature vectors obtained from image patches are aggregated using statistical moments, then regressed onto subjective scores. In this instance, the first stage is patchwise training, while the second stage is imagewise training. Since the local proxy scores reflect the nonstationary characteristics of perceived quality, they are reasonable local regression targets, and training of the CNN model is enabled by the abundant training samples. Following the second stage of training on human ground-truth, their model attains highly competitive prediction accuracy on the legacy data sets.

The same authors later developed a two-stage training scheme for no-reference picture-quality prediction called the *deep image quality assessor (DIQA)* [27]. The training process of that model was separated into an objective training stage followed by a subjective training stage. Rather than using a sophisticated picture-quality predictor to produce proxy scores, they computed peak signal-to-noise (PSNR). Using only convolutional layers, feature maps were obtained, which were then regressed onto objective error maps. The second stage aggregated the feature maps by weighted averaging, then

regressed these global features onto ground-truth subjective scores. The weighting maps were also learned during training. The reported prediction accuracy of these models is competitive with state-of-the-art models on the legacy databases.

CNN-based full-reference picture-quality models

While CNNs were first used to model no-reference picture quality, more recently, they have been applied to the reference prediction problem as well.

Liang et al. [28] proposed a dual-path CNN-based full-reference-quality prediction model. They generalized the problem by seeking to predict quality using a nonaligned image of a similar scene as a reference. Locally normalized distorted and reference image patches are fed into a dual-path CNN model, each using the same parameter values. Then the concatenated learned feature vectors are regressed onto the subjective scores of source distorted images. They report state-of-the-art prediction accuracies in both aligned and nonaligned full-reference scenarios.

Gao et al. deployed a deep CNN model pretrained on ImageNet. They used it to conduct full-reference picture-quality prediction [29] by feeding pairs of reference and distorted pictures into the CNN, where each output layer is used as a feature map. Local similarities between the feature maps obtained from the reference and distorted images are then computed and pooled to arrive at global picture-quality scores. The CNN model was not fine-tuned on any picture-quality database.

The deep CNN-based full-reference-quality prediction model in [30], called *DeepQA*, was trained to learn a visual sensitivity weight at each coordinate using measured local spatial characteristics of the distorted image. DeepQA accepts the distorted image and an objective error map (e.g., mean squared error) as inputs. The learned weight map is then used as a multiplier on the objective error map. The authors reported consistent state-of-the-art prediction accuracies as compared to other reference-quality models, on the synthetic-distortion legacy picture-quality databases.

Summary of CNN-based picture-quality models

Table 1 compares the implementations of reported CNN-based no-reference [23]–[27] and full-reference [28]–[30] picture-quality models. For full-reference models, the strategies used to compare distorted and reference features are summarized in the last column. In [28] and [30], this merely amounts to supplying both to the network. Generally, the reviewed models were designed to overcome the lack of training data, which is the most important issue that needs to be resolved to employ deep CNN models successfully. Most of the models used some type of patch-based training to increase the training data volume. Several of the models used proxy ground-truth scores generated by objective-quality prediction models to augment the subjective scores or, alternately, to pretrain the network on a large amount of easily generated proxy data before fine-tuning on subjective scores. Since we have found no serious attempts to use unsupervised deep models, we make no comparisons of this type, although the success of the very simple model [31] suggests this is an interesting research direction. Finding ways to embody models of perception into

Table 1. A comparison of implementations of CNN-based picture-quality prediction models.

Models	Type	Layer Depth	Preprocessing	Feature Aggregation or Score Pooling
[23]	NR	2 Conv and 2 FC	Local normalization	Mean pooling (during testing)
[24]	NR	14 Conv (4 NiN blocks)	Local normalization	Mean pooling (during testing)
[25]	NR	10 Conv and 2 FC	Raw RGB image	Mean or weighted average pooling
[26]	NR	2 Conv and 6 FC	Local normalization	Mean and standard deviation aggregation
[27]	NR	8 Conv and 3 FC	Low-frequency subtraction	Mean or weighted average aggregation
[28]	FR	(2 Conv, 1 FC)×2 and 2 FC	Local normalization	(Not mentioned)
[29]	FR	13 Conv and 3 FC	Raw RGB image	Mean aggregation and pooling
[30]	FR	(2 Conv)×2, 6 Conv and 2 FC	Low-frequency subtraction	Weighted average aggregation

Models	Type	Training Targets		Comments
		First Stage	Second Stage	
[23]	NR	Subjective scores	N/A	(Comparison strategy for FR models) Patchwise training
[24]	NR	Semantic label	Subjective scores	Fine-tuning of pretrained CNN on ImageNet
[25]	NR	Subjective scores	N/A	Weighted average patch aggregation
[26]	NR	Proxy scores	Subjective scores	Uses proxy patch labels
[27]	NR	Objective error map	Subjective scores	Uses proxy patch labels
[28]	FR	Subjective scores	N/A	Concatenation of feature vectors
[29]	FR	Semantic label	N/A	SSIM between feature maps of each layer
[30]	FR	Subjective scores	N/A	Concatenation of feature maps

FR: full-reference, NR: no-reference, Conv: convolutional layers, and FC: fully connected layers.

deep picture-quality models is also an issue. While simpler models often use perceptually relevant bandpass processing and local divisive normalization [23], similar processes may be learned by the network at the early stages. However, it should be possible to impose perceptual weighting or pooling strategies on the network to account for aspects of visual sensitivity, which could accelerate the process of training on subjective scores.

In CNN-based schemes, the process of feature aggregation or score pooling determines the form of a loss function. Examples of aggregation and pooling strategies are shown in Figure 4. The patch-based algorithms described in [23] and [24] did not use aggregation or pooling during training. Instead, each image patch was independently regressed onto the global subjective-quality score. The loss function used is

$$\mathcal{L} = \frac{1}{N} \sum_i \|f(p_i) - S\|, \quad (1)$$

where p_i refers to the i th patch obtained, N is the number of patches, S is the ground-truth score, and $f(\cdot)$ is an NN process. The models were trained via a patchwise optimization, and, during testing, the outputs of multiple patches composing an input image were averaged to obtain a final predicted subjective score. Conversely, imagewise approaches use aggregation or pooling during training. For example, weighted average pooling methods [25] may be used, where the loss function looks like

$$\mathcal{L}' = \|\text{pool}(f(p_1), \dots, f(p_N)) - S\|, \quad (2)$$

where $\text{pool}(\cdot)$ refers to an unspecified pooling method [Figure 4(a)]. In [26] and [27] [Figure 4(b) and (c)], simple feature aggregation was used. A more complicated model, such as a multilayer perception or recurrent NN [4], could also be used for aggregation [Figure 4(d)]. Here, the loss function becomes

$$\mathcal{L}'' = \|g(\text{aggr}(f(p_1), \dots, f(p_N))) - S\|, \quad (3)$$

where $\text{aggr}(\cdot)$ refers to a feature aggregation process and $g(\cdot)$ is a regression NN. The forms (2) and (3) have the advantage that the model can be trained under the same conditions as the actual testing conditions, where the imagewise scores are predicted.

Description of picture-quality databases

The choice and consideration of a database for training is important for deep-learning-based models, since their performance depends highly on the size of the training set. In most picture-quality databases, the distorted images are afflicted by only a single type of synthetically introduced distortion, such as JPEG compression, simulated sensor noise, or simulated blur, as exemplified in Figure 5(a). Since they have played important roles in the development of perceptual picture-quality studies, we briefly describe several popular legacy databases in the following.

The LIVE IQA database [12], which was the first successful public-domain picture-quality database and is still the most widely used, contains 29 reference images and 982 images, each

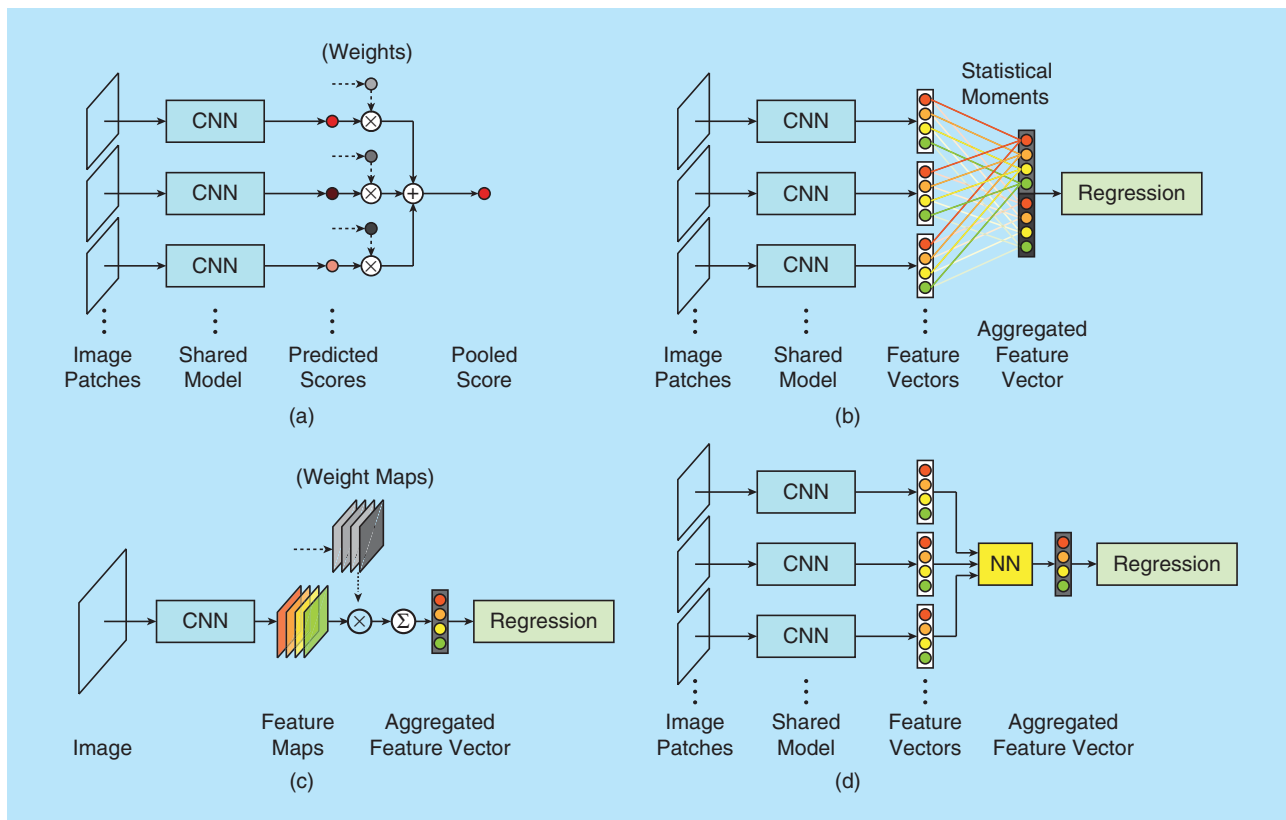


FIGURE 4. Examples of aggregation and pooling strategies in CNN-based picture-quality prediction models. (a) Weighted average pooling, (b) elementwise aggregation, (c) weighted average aggregation, and (d) NN aggregation.

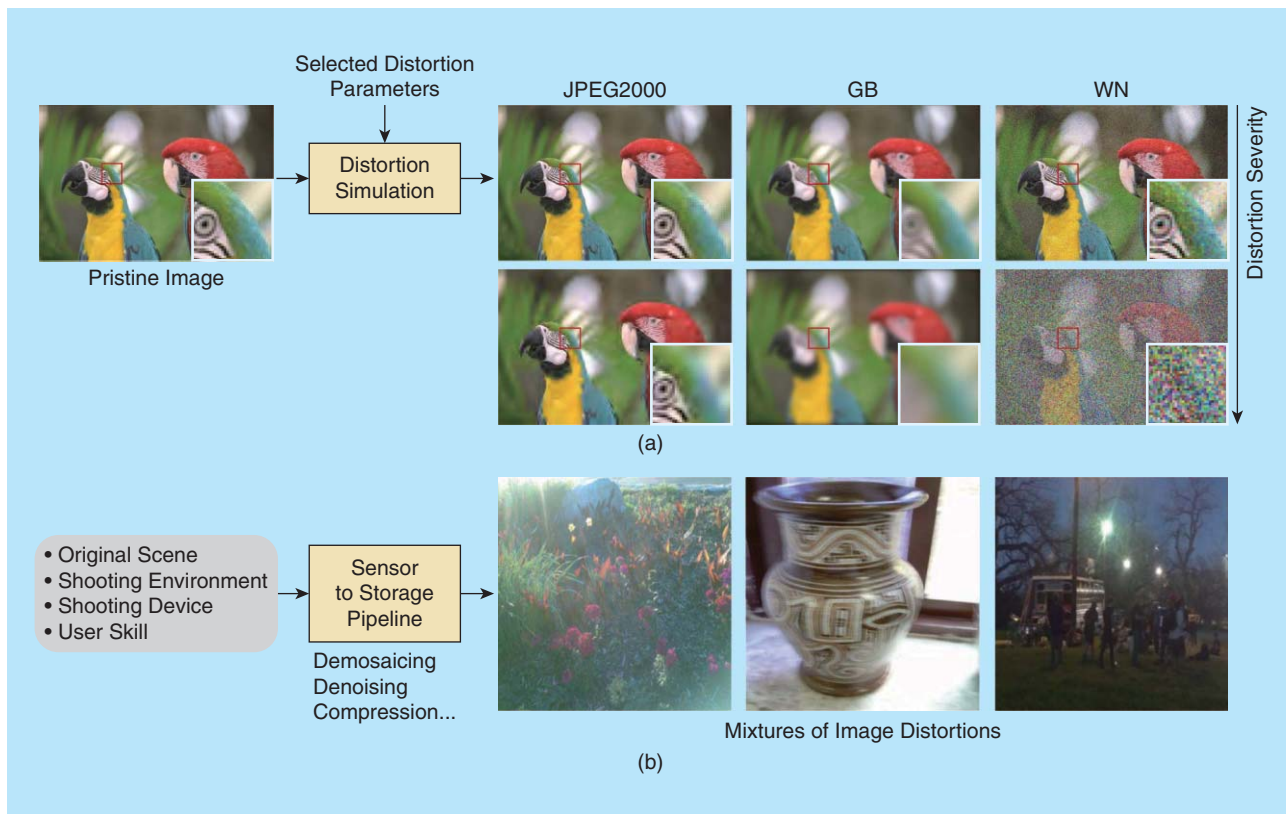


FIGURE 5. (a) Synthetic and (b) authentic image distortions found in picture-quality databases.

impaired by one of five types of synthetic distortions: JPEG and JPEG2000 (JP2K) compression, white Gaussian noise (WN), GB, and Rayleigh fast-fading channel distortion. The differential mean opinion score (DMOS) of each distorted image is provided. The CSIQ database [32] includes 30 reference images and 866 synthetically distorted images of six types: JPEG, JP2K, WN, GB, pink Gaussian noise, and global contrast decrements. The DMOS of the distorted images is also provided. TID2013 [13] contains the largest number of distorted images. It consists of 25 reference images and 3,000 synthetically distorted images with 24 different distortions at five levels of degradation. The database also provides the mean opinion scores (MOS). The LIVE multiply distorted (MD) database [33] was the first to include multiple (synthetically) distorted images. Images in it are distorted by two types of distortion in two combinations: simulated GB followed by JPEG compression and GB followed by additive WN. It contains 15 references and 405 distorted images, and the DMOS of each distorted image is provided.

Finally, the LIVE Challenge database [3] contains nearly 1,200 unique image contents, captured by a wide variety of mobile camera devices under highly diverse conditions. As such, the images were subjected to numerous types of authentic distortions during the capture process, often in complex combinations of multiple interacting impairments, as shown in Figure 5(b). The distortions include, e.g., low-light blur and noise, motion blur, camera shake, overexposure, underexposure, a variety of color errors, compression errors, and many combinations of these and other impairments. There are no reference images in the LIVE Challenge database, since the distorted images are originals, captured by dozens of ordinary photographers. The LIVE Challenge pictures were judged by more than 8,100 human subjects in a tightly monitored crowdsourced study, yielding more than 350,000 human judgments that exhibit excellent internal consistency [3]. A summary of the attributes of these five databases is shown in Table 2.

Performances of CNN picture-quality models

Since only a few CNN-based picture-quality models have been released, we provide the prediction accuracies of baseline models on the five databases as performance references to be compared against. We selected the well-known very deep CNN models AlexNet [2] and ResNet50 [34] as the architectures of the baseline models, where each was pretrained on the ImageNet

classification task. Both of these pretrained models are available for download. The specific network configurations can be found in the original papers. For each pretrained architecture, two types of back-end training strategies were tested: using an SVR to regress the extracted features from the CNN model onto subjective scores and fine-tuning the pretrained networks to conduct picture-quality prediction. We did not test direct training of these models on any of the picture-quality databases, since they are not large enough. Very deep networks easily overfit on insufficient training samples, causing significant decreases in testing accuracy (AlexNet has 62 million and ResNet50 has 26 million parameters). Instead, we tested a smaller CNN network as a baseline model of direct training.

In the first approach, the output of the sixth fully connected layer (4,096 dimensions) from AlexNet and averaged-pooled features (2,048 dimensions) from ResNet50 were used as the input feature vectors to the SVR. From each input image, 25 randomly cropped image patches (the patch size is predefined by the pretrained models: 227×227 for AlexNet, and 224×224 for ResNet50) were used for training and testing. The obtained feature vectors from these 25 image patches were averaged to obtain a single global feature vector.

In the second approach, we randomly cropped 100 image patches from each training image to be used for training (except on the TID2013 database, where 30 cropped patches were used, due to the large number of distorted images in the database). The image patches inherited the quality scores from the source distorted images, which were first normalized to the range [0, 1]. This preprocessing enabled us to use the same parameter settings on all databases. The basic regression loss (1) was used. To alleviate overfitting, one dropout layer with dropout rate 0.5 was added before the last fully connected layer. The learning rate was set to 10^{-3} , and the fine-tuning process iterated for eight and six epochs on AlexNet and ResNet50, respectively. The batch size was fixed at 48 for both models. In the testing stage, the pretrained models were used to predict quality scores on each of 25 random image crops. These were average pooled to produce the final picture-quality scores.

For the direct training approach, we used the following CNN architecture: Conv-48, Conv-48 with stride 2, Conv-64, Conv-64 with stride 2, Conv-64, Conv-64, Conv-128, Conv-128, FC-128, FC-128, and FC-1. Here, “Conv” refers to convolutional layers, “FC” refers to fully connected layers, and the trailing

Table 2. A comparison of IQA databases in terms of numbers of reference images, distorted images, distortion types, authenticity of distortions, type of subjective scores, whether distortions are mixed, and published date.

Database	Number of Reference Images	Number of Distorted Images	Number of Distorted Types	Authenticity of Distortions	Subjective Score Type	Mixtures of Distortions	Published Date
LIVE IQA [12]	29	779	5	Synthetic	DMOS	N/A	2003
CSIQ [32]	30	866	6	Synthetic	DMOS	N/A	2010
TID2013 [13]	25	3,000	24	Synthetic	MOS	N/A	2015
LIVE MD [33]	15	405	2	Synthetic	DMOS	✓	2012
LIVE Challenge [3]	N/A	1,162	Numerous	Authentic	MOS	✓	2016

numbers indicate the number of feature maps (of Conv) or output nodes (of FC). The model accepts 112×112 images as inputs. All of the convolutional layers were configured to use 3×3 filters, using zero-padding to preserve the spatial size. Each layer used a rectified linear unit as the activation function. Following the convolutional layers, each 28×28 feature map (assuming two convolutional layers with a stride of two) was averaged yielding an 128-dimensional feature vector, which is then fed into the fully connected layers. The number of parameters in this model is about 0.4 million, which is much lower than AlexNet or ResNet50. This baseline model was trained using the imagewise L_2 loss in (3). Each input image was partitioned into 112×112 patches when training on the LIVE IQA database, while full-sized images were used on the other databases. On the LIVE IQA database, nonoverlapping patches were used so that overlapped regions could not be accessed multiple times by the CNN model during training and/or testing. The data was augmented by supplementing the training set with horizontally flipped replicas of each image. Each mini-batch contained patches extracted from five images. The training was iterated over 80 epochs.

Two performance metrics were used to benchmark the models: Spearman's rank order correlation coefficient (SRCC), and Pearson's linear correlation coefficient (PLCC). To evaluate the baseline models, we randomly divided each database into two subsets of nonoverlapping content (distorted or otherwise), 80% for training and 20% for testing. Of course, all of the LIVE Challenge pictures contain different contents. The SRCC and PLCC were averaged after ten repetitions of this random process.

The performances of all of the exemplar picture-quality prediction models on the LIVE IQA database are shown in Figure 6. The first five (from left) are no-reference learning-based models, where the last two of these used deep learning. The next seven are CNN-based no-reference-quality prediction models, and the last three are CNN-based full-reference models. The reported SRCC and PLCC scores of the listed models

were taken from the original papers. Overall, the CNN-based full-reference models followed by the CNN-based no-reference models achieved higher prediction accuracies relative to conventional learning-based models on the legacy databases.

Table 3 compares the performance of the various picture-quality prediction models on all of the reviewed databases. The last five rows show results for the baseline models. The three best performing no-reference picture-quality models in each column are boldfaced. Generally, the existing CNN-based models were able to achieve remarkable prediction accuracies on the legacy databases. However, it is much harder to obtain successful results on the LIVE Challenge database. For example, the model proposed in [27], DIQA, achieved an SRCC of 0.687, which is lower than the results attained by a recent successful SVR-based method, FRIQUEE-ALL [21], which achieved an SRCC of 0.72.

However, the baseline models that were pretrained on the ImageNet databases achieved standout accuracies on the LIVE Challenge database. This is likely because the real-world ImageNet pictures are not synthetically distorted. Instead, like the LIVE Challenge pictures, any distortions occurred as a natural consequence of photography, without intervention by the database creator. This further suggests that the pretrained CNNs are, to some degree, already quality-aware, meaning that their learned internal features assist the performance of the task (recognition) by adapting to the presence of authentic distortions.

The baseline models using the first approach achieved very low accuracies on the legacy databases, since they were not exposed to any synthetic distortions during training, and hence the learned features were not very useful to the SVR for quality prediction. Fine-tuning the pretrained baseline deep models significantly improved performance on the legacy synthetic databases, but not enough to make them competitive, since there was not enough data to train them adequately. The exception was the directly trained shallow CNN baseline model, which achieved competitive performance on the legacy databases, but lower accuracies on the LIVE Challenge database.

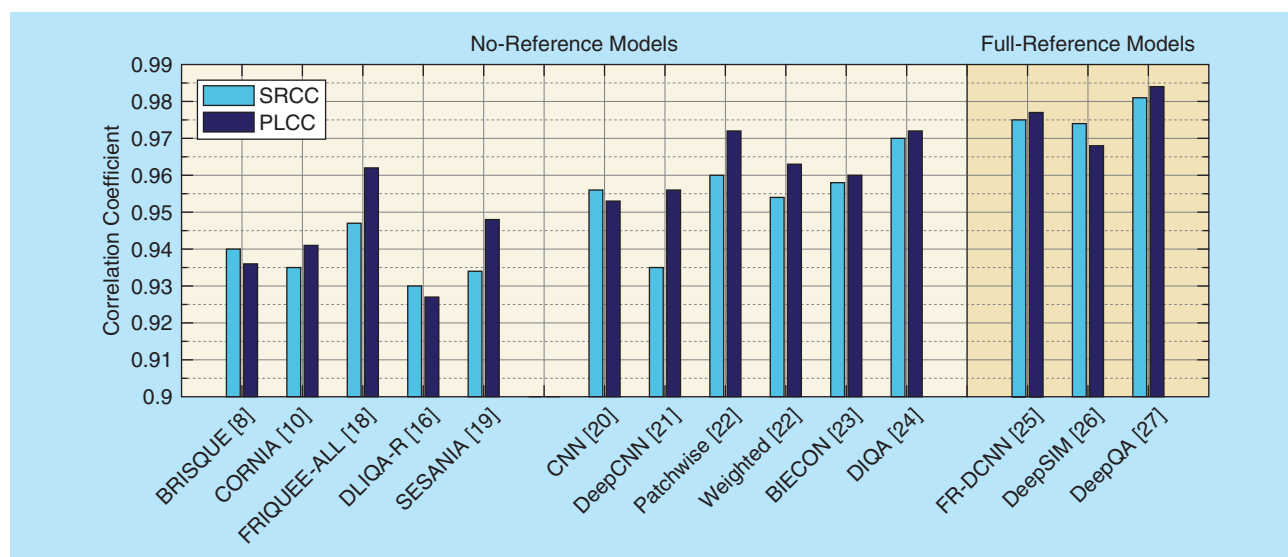


FIGURE 6. A comparison of the SRCC and PLCC of learned picture-quality models on the legacy LIVE IQA database.

Table 3. The SRCC and PLCC comparison on five public-domain subjective picture-quality databases.

Type	Methods	LIVE IQA		CSIQ		TID2013		LIVE MD		LIVE Challenge	
		SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC
FR	PSNR	0.876	0.872	0.806	0.800	0.636	0.706	0.725	0.815	N/A	N/A
	SSIM [15]	0.948	0.945	0.876	0.861	0.775	0.691	0.845	0.882	N/A	N/A
	FSIMc [35]	0.963	0.960	0.931	0.919	0.851	0.877	0.863	0.818	N/A	N/A
NR	DeepQA [30]	0.981	0.982	0.961	0.965	0.939	0.947	0.938	0.942	N/A	N/A
	BRISQUE [9]	0.939	0.942	0.756	0.797	0.572	0.651	0.897	0.921	0.607	0.585
	CORNIA [11]	0.942	0.943	0.714	0.781	0.549	0.613	0.900	0.915	0.618	0.662
	FRIQUEE-ALL [21]	0.948	0.962	0.839	0.863	0.669	0.704	0.925	0.940	0.720	0.720
	BIECON [26]	0.958	0.960	0.815	0.823	0.717	0.762	0.909	0.933	0.595	0.613
	DIQA [27]	0.970	0.972	0.844	0.880	0.843	0.868	0.920	0.933	0.687	0.701
	AlexNet + SVR	0.901	0.908	0.712	0.736	0.263	0.365	0.760	0.803	0.769	0.790
	ResNet50 + SVR	0.925	0.935	0.654	0.700	0.435	0.495	0.797	0.833	0.806	0.825
	AlexNet + fine-tuning	0.947	0.952	0.817	0.840	0.615	0.668	0.899	0.914	0.748	0.779
ResNet50 + fine-tuning	0.950	0.954	0.876	0.905	0.712	0.756	0.909	0.920	0.819	0.849	
Imagewise CNN	0.963	0.964	0.812	0.791	0.800	0.802	0.914	0.929	0.663	0.705	

FR: full reference, NR: no reference. Italics indicate CNN-based methods. Boldface entries indicate the top three performers on each database for each performance metric.

A possible explanation for these results is that the pretrained deep models adapted easily to the authentic distortions in LIVE Challenge as a consequence of having learned image recognition tasks on real-world pictures. Applying them to databases with synthetic distortions, however, like LIVE IQA and TID2013, likely failed to exploit what was learned regarding authentic distortions; hence, significant retraining would be needed to deal with the synthetic distortions. This may help explain the excellent generalization power of pretrained models when applied to other real world image tasks: their ability to handle authentic distortions, by representing them to improve task performance.

Envisioning the future

The sizes of the training sets used is critical to the success of deep NN models. Current public-domain databases have insufficient size as compared to widely used image recognition databases. However, constructing large-scale perceptual-quality databases is a much more difficult problem than image recognition databases. Creating databases for picture-quality assessment requires time-consuming and expensive subjective studies, which must be conducted under controlled laboratory conditions. Even if the number of reference images is small, the required number of subjective tests quickly becomes excessive. Conducting subjective tests using online crowdsourcing is one possible solution (like the LIVE Challenge database), yet even online tests are (probably) prohibitively difficult to scale up to the necessary size, especially while ensuring the aggregate quality of the collected human data. Another possibility would be if a large social media company were to engage their customers to provide picture-quality scores, similar to the Netflix DVD ratings by e-mail of a decade ago. Generally, understanding how to successfully create reliable, very large-scale, and authentic picture-quality databases remains an open question.

Authors

Jongyoo Kim (jongky@yonsei.ac.kr) received his B.S. and M.S. degrees in electrical and electronic engineering from Yonsei University, Seoul, South Korea, in 2011 and 2013, respectively. He is currently working toward his Ph.D. degree in the Department of Electrical and Electronic Engineering, Yonsei University, South Korea. His research interests include two-dimensional (2-D)/three-dimensional (3-D) image and video processing based on the human visual system, quality assessment of 2-D/3-D image and video, 3-D computer vision, and deep learning. He was a recipient of the Global Ph.D. Fellowship by the National Research Foundation of Korea from 2011 to 2016.

Hui Zeng (cshzeng@comp.polyu.edu.hk) received his M.S. degree from the School of Information and Communication Engineering, Dalian University of Technology, China, in 2016. He is currently pursuing his Ph.D. degree in the Department of Computing, The Hong Kong Polytechnic University, under the supervision of Prof. Lei Zhang. His research interests include computer vision, image and video processing, and deep learning.

Deepti Ghadiyaram (deepti@cs.utexas.edu) received her Ph.D. degree from the Department of Computer Science at the University of Texas (UT) at Austin. Her research interests include image and video processing, computer vision, and machine learning. Her Ph.D. work focused on perceptual image and video quality assessment, particularly on building quality-prediction models for pictures and videos captured in the wild and understanding a viewer's time-varying quality of experience while streaming videos. She was a recipient of the UT Austin's Microelectronics and Computer Development Fellowship from 2013 to 2014 and the Graduate Student Fellowship from the Department of Computer Science from 2013 to 2016. She joined Facebook Research in September 2017.

Sanghoon Lee (slee@yonsei.ac.kr) received the B.S. degree from Yonsei University, Seoul, South Korea, in 1989, the M.S.

degree from the Korea Advanced Institute of Science and Technology, Seoul, in 1991, and the Ph.D. degree from the University of Texas at Austin, in 2000. He is a full professor in the Department of Electrical and Electronic Engineering, Yonsei University. His research interests include image/video quality assessment, computer vision, graphics, cloud computing, multimedia communications, and wireless networks. He has been an associate editor of *IEEE Signal Processing Letters* and *Journal of Electronic Imaging* as well as chair of the IEEE P3333.1 Quality Assessment Working Group. He currently serves as a member of the IEEE Multimedia Signal Processing Technical Committee (TC) and the IEEE IVMSIP TC and the APSIPA IVMSIP TC vice chair.

Lei Zhang (cszhang@comp.polyu.edu.hk) received his B.S. degree from Shenyang Institute of Aeronautical Engineering, China, and his M.S. and Ph.D. degrees from Northwestern Polytechnical University, Xi'an, China. He is a chair professor in the Department of Computing, The Hong Kong Polytechnic University. His research interests include computer vision, pattern recognition, image and video analysis, and biometrics. He has published more than 200 papers in those areas, and, as of 2017, his publications have been cited more than 26,000 times in the literature. He is an associate editor of *IEEE Transactions on Image Processing*, *SIAM Journal on Imaging Sciences*, and *Image and Vision Computing* and was selected as a Web of Science Highly Cited Researcher by Thomson Reuters.

Alan C. Bovik (bovik@ece.utexas.edu) received the B.S., M.S., and Ph.D. degrees from the University of Illinois in 1980, 1982, and 1984, respectively. He is a Cockrell Family Regents Endowed Chair Professor at the University of Texas at Austin. He received the 2017 Edwin H. Land Medal from the Optical Society of America, a 2015 Prime-Time Emmy Engineering Award, and the 2013 IEEE Signal Processing Society's Society Award. He has published *The Handbook of Image and Video Processing*, *Modern Image Quality Assessment*, and *The Essential Guide to Image and Video Processing*. He cofounded and was the longest-serving editor-in-chief of *IEEE Transactions on Image Processing*. He also created the IEEE International Conference on Image Processing in Austin, Texas, in 1994. He is a Fellow of the IEEE.

References

- [1] A. C. Bovik, "Automatic prediction of perceptual image and video quality," *Proc. IEEE*, vol. 101, no. 9, pp. 2008–2024, 2013.
- [2] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Advances in Neural Information Processing Systems Conf.* 2012, pp. 1097–1105.
- [3] D. Ghadiyaram and A. C. Bovik, "Massive online crowdsourced study of subjective and objective picture quality," *IEEE Trans. Image Process.*, vol. 25, no. 1, pp. 372–387, 2016.
- [4] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 8, pp. 1798–1828, 2013.
- [5] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, et al., "TensorFlow: Large-scale machine learning on heterogeneous systems." [Online]. Available: <https://www.tensorflow.org/>
- [6] M. Clark and A. C. Bovik, "Experiments in segmenting texton patterns using local-sized spatial filters," *Pattern Recognit.*, vol. 22, no. 6, pp. 707–717, 1989.
- [7] H. Lee, C. Ekanadham, and A. Y. Ng, "Sparse deep belief net model for visual area V2," in *Proc. Advances in Neural Information Processing Systems Conf.*, 2008, pp. 873–880.
- [8] Y. Yuan, Q. Guo, and X. Lu, "Image quality assessment: A sparse learning way," *Neurocomputing*, vol. 159, pp. 227–241, July 2015.
- [9] A. Mittal, A. K. Moorthy, and A. C. Bovik, "No-reference image quality assessment in the spatial domain," *IEEE Trans. Image Process.*, vol. 21, no. 12, pp. 4695–4708, 2012.
- [10] H. Tang, N. Joshi, and A. Kapoor, "Learning a blind measure of perceptual image quality," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2011, pp. 305–312.
- [11] P. Ye, J. Kumar, L. Kang, and D. Doermann, "Unsupervised feature learning framework for no-reference image quality assessment," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2012, pp. 1098–1105.
- [12] H. Sheikh, M. Sabir, and A. Bovik, "A statistical evaluation of recent full reference image quality assessment algorithms," *IEEE Trans. Image Process.*, vol. 15, no. 11, pp. 3440–3451, 2006.
- [13] N. Ponomarenko, L. Jin, O. Ieremeiev, V. Lukin, K. Egiazarian, J. Astola, B. Vozel, K. Chehdi, et al., "Image database TID2013: Peculiarities, results and perspectives," *Signal Process. Image Commun.*, vol. 30, pp. 57–77, Jan. 2015.
- [14] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2009, pp. 248–255.
- [15] Z. Wang, A. C. Bovik, H. Sheikh, and E. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, 2004.
- [16] D. J. Heeger, "Normalization of cell responses in cat striate cortex," *Vis. Neurosci.*, vol. 9, no. 2, pp. 181–197, 1992.
- [17] W. Hou, X. Gao, D. Tao, and X. Li, "Blind image quality assessment via deep learning," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 26, no. 6, pp. 1275–1286, 2015.
- [18] H. R. Sheikh and A. C. Bovik, "Image information and visual quality," *IEEE Trans. Image Process.*, vol. 15, no. 2, pp. 430–444, 2006.
- [19] X. Li, Q. Guo, and X. Lu, "Spatiotemporal statistics for video quality assessment," *IEEE Trans. Image Process.*, vol. 25, no. 7, pp. 3329–3342, 2016.
- [20] M. A. Saad, A. C. Bovik, and C. Charrier, "Blind prediction of natural video quality," *IEEE Trans. Image Process.*, vol. 23, no. 3, pp. 1352–1365, 2014.
- [21] D. Ghadiyaram and A. C. Bovik, "Perceptual quality prediction on authentically distorted images using a bag of features approach," *J. Vision*, vol. 17, no. 1, 2017.
- [22] Y. Li, L.-M. Po, X. Xu, L. Feng, F. Yuan, C.-H. Cheung, and K.-W. Cheung, "No-reference image quality assessment with Shearlet transform and deep neural networks," *Neurocomputing*, vol. 154, pp. 94–109, 2015.
- [23] L. Kang, P. Ye, Y. Li, and D. Doermann, "Convolutional neural networks for no-reference image quality assessment," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2014, pp. 1733–1740.
- [24] Y. Li, L. M. Po, L. Feng, and F. Yuan, "No-reference image quality assessment with deep convolutional neural networks," in *Proc. IEEE Int. Conf. Digital Signal Processing*, 2016, pp. 685–689.
- [25] S. Bosse, D. Maniry, T. Wiegand, and W. Samek, "A deep neural network for image quality assessment," in *Proc. IEEE Int. Conf. Image Processing*, 2016, pp. 3773–3777.
- [26] J. Kim and S. Lee, "Fully deep blind image quality predictor," *IEEE J. Sel. Topics Signal Process.*, vol. 11, no. 1, pp. 206–220, 2017.
- [27] J. Kim and S. Lee, "Deep CNN-based blind image quality predictor," submitted for publication.
- [28] Y. Liang, J. Wang, X. Wan, Y. Gong, and N. Zheng, "Image quality assessment using similar scene as reference," in *Proc. European Conf. Computer Vision*, 2016, pp. 3–18.
- [29] F. Gao, Y. Wang, P. Li, M. Tan, J. Yu, and Y. Zhu, "DeepSim: Deep similarity for image quality assessment," *Neurocomputing*, vol. 257, pp. 104–114, Sept. 2017.
- [30] J. Kim and S. Lee, "Deep learning of human visual sensitivity in FR-IQA framework," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2017, pp. 1676–1684.
- [31] A. Mittal, R. Soundararajan, and A. C. Bovik, "Making a 'completely blind' image quality analyzer," *IEEE Signal Process. Lett.*, vol. 20, no. 3, pp. 209–212, 2013.
- [32] E. C. Larson and D. M. Chandler, "Most apparent distortion: Full-reference image quality assessment and the role of strategy," *J. Electron. Imaging*, vol. 19, no. 1, pp. 19–19–21, 2010.
- [33] D. Jayaraman, A. Mittal, A. K. Moorthy, and A. C. Bovik, "Objective quality assessment of multiply distorted images," in *Proc. Asilomar Conf. Signals, Systems, and Computers*, 2012, pp. 1693–1697.
- [34] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2016, pp. 770–778.
- [35] L. Zhang, L. Zhang, X. Mou, and D. Zhang, "FSIM: A feature similarity index for image quality assessment," *IEEE Trans. Image Process.*, vol. 20, no. 8, pp. 2378–2386, 2011.

Stefano Fortunati, Fulvio Gini, Maria S. Greco,
and Christ D. Richmond

Performance Bounds for Parameter Estimation under Misspecified Models

Fundamental findings and applications

Inferring information from a set of acquired data is the main objective of any signal processing (SP) method. The common problem of estimating the value of a vector of parameters from a set of noisy measurements is at the core of a plethora of scientific and technological advances in recent decades, including wireless communications, radar and sonar, biomedicine, image processing, and seismology.

Developing an estimation algorithm often begins by assuming a statistical model for the measured data, i.e., a probability density function (pdf), which, if correct, fully characterizes the behavior of the collected data/measurements. Experience with real data, however, often exposes the limitations of any assumed data model, since modeling errors at some level are always present. Consequently, the true data model and the model assumed to derive the estimation algorithm could differ. When this happens, the model is said to be mismatched or misspecified. Therefore, understanding the possible performance loss or regret that an estimation algorithm could experience under model misspecification is critical for any SP practitioner. Furthermore, understanding the limits on the performance of any estimator subject to model misspecification is of practical interest.

Motivated by the widespread and practical need to assess the performance of a mismatched estimator, the goal of this article is to help bring attention to the main theoretical find-

ings on estimation theory, and, in particular, on lower bounds under model misspecification, that have been published in the statistical and econometrical literature in the last 50 years. Additionally, some applications are discussed to illustrate the broad range of areas and problems to which this framework extends and, consequently, the numerous opportunities available for SP researchers.

A formal theory of statistical inference under misspecified models

The mathematical basis for a formal theory of statistical inference was presented by Fisher, who introduced the maximum likelihood (ML) method along with its main properties [9]. Since then, ML estimation has been widely used in a variety of applications. One of the main reasons for its popularity is its asymptotic efficiency, i.e., its ability to achieve a minimum value of the error variance as the number of available observations goes to infinity or as the noise power decreases to zero. The concept of efficiency is strictly related to the existence of some lower bounds on the performance of any estimator designed for a specific inference task. Such performance bounds, one of which is the celebrated Cramér–Rao bound (CRB) [8], [33], are fundamentally important in practical applications, as they provide a benchmark of comparison for the performance of any estimator. Specifically, given a particular estimation problem, if the performance of a certain algorithm achieves a relevant performance bound, then no other algorithm can do better. Moreover, evaluating a performance

Digital Object Identifier 10.1109/MSP.2017.2738017
Date of publication: 13 November 2017



©ISTOCKPHOTO.COM/MIKIEVY

bound is often a prerequisite for any feasibility study. The availability of a lower bound for the estimation problem at hand makes the SP practitioner aware of the practical impossibility to achieve better estimation accuracy than the one indicated by the bound itself. Another fundamental feature of a performance bound is its ability to capture and reveal the complex dependences among the various parameters of interest, thus offering the opportunity to more deeply understand the estimation problem at hand and, ultimately, to identify an appropriate design choice of parameters and criteria for an estimator [23].

Before describing specific performance bounds, it is worth mentioning that estimation theory explores two different frameworks: one is deterministic and one is Bayesian. In the classical deterministic approach, the parameters to be estimated are modeled as deterministic but unknown variables. This implies that no a priori information is available that would suggest that one outcome is more or less likely than another. In the Bayesian framework, the parameters of interest are assumed to be random variables, and the goal is to estimate their particular realizations. Unlike the classical deterministic approach, the Bayesian approach exploits this random characterization of the unknown parameters by incorporating a priori information about the unknown parameters in the derivation of an estimation algorithm. The joint pdf of the unknown parameters is assumed to be known and, therefore, can be taken into account in the estimation process through Bayes' theorem [23].

Basics about performance bounds

When discussing lower bounds, the first distinction that needs to be made is between local (small-error) bounds and global (large-error) bounds. A bound can be considered a local-error bound if its calculation relies exclusively on the behavior of the pdf of the data at a single point value of the parameter (or perhaps a very small local neighborhood around this point). If the calculation of a bound requires knowledge of the pdf behavior at multiple (more than one) distinct and well-separated (nonlocal) points, then the bound can be characterized as a global-error bound. Local-error bounds at best determine the limits of the asymptotics of optimal algorithms like ML, whereas the characterization of nonasymptotic performance must somehow take into account the possible influence of parameter values other than the true value.

A bound is said to be tight if it reasonably predicts the performance of the ML estimator. If a bound is only asymptotically tight, then it is reliable only in the presence of a high signal-to-noise ratio (SNR) or a sufficiently large number of measurements. However, if a bound is globally tight, then it is a reliable bound for the error covariance of an ML estimator, irrespective of the SNR level or of the amount of available data. The deterministic bound that can be regarded as the most general representative of the class of global bounds is the Barankin bound (BB) [3]. However, due to its generality, the calculation of the BB is not straightforward, and it usually does not admit a closed-form representation. The most popular local bound is the aforementioned CRB. Unlike the BB, the CRB

is easy to evaluate for many practical problems, but it is only asymptotically reliable. In the nonasymptotic region, which is characterized by a low SNR and/or by a low number of measurements, the CRB can be too optimistic with respect to (w.r.t.) the effective error covariance achievable by an estimator [44].

The second subdivision of the performance bounds is a direct consequence of the dichotomy between the deterministic and the Bayesian estimation frameworks. In particular, we can identify the class of deterministic lower bounds and the class of Bayesian lower bounds [43]. Without any claim of completeness, the class of the deterministic lower bounds includes the (global) BB [3] and two local bounds, the Bhattacharyya bound [5] and the CRB [8], [33]. We stress that the most common forms of these bounds, including the CRB, apply only to unbiased estimators. Versions of these bounds exist, however, that can be applied to biased estimators whose bias function can be determined. Concerning the Bayesian bounds, they can be divided into two classes [34]: the Ziv–Zakai family and the Weiss–Weinstein family, to which the Bayesian version of the CRB belongs. The first family is derived by relating the mean squared error (MSE) to the probability of error in a binary hypothesis testing problem, while the derivation of the latter is based on the covariance inequality. For further details on Bayesian bounds, refer to [43].

An estimation theory under model misspecification: Motivations

Regardless of the differences previously discussed, both the classical deterministic estimation theory and the Bayesian framework are based on the implicit assumption that the assumed data model (the pdf) and the true data model are the same, i.e., the model is correctly specified. However, much evidence from engineering practice shows that this assumption is often violated; the assumed model is different from the true one. There are two main reasons for model misspecification. The first is the imperfect knowledge of the true data model, which leads to an incorrect specification of the data pdf. However, there could be cases where perfect knowledge of the true data model is available, but, due to an intrinsic computational complexity or to a costly hardware implementation, it is not possible nor convenient to pursue the optimal “matched” estimator. In these cases, one may prefer to derive an estimator by assuming a simpler but misspecified data model, e.g., the Gaussian model. Of course, this suboptimal procedure may lead to some degradation in the overall system performance, but it ensures a simple analytical derivation and real-time hardware implementation of the inference algorithm. In such a misspecified estimation framework, the possibility to assess the impact of the model misspecification on the estimation performance is of fundamental importance to guarantee the reliability of the (mismatched) estimator. Misspecified bounds are then the perfect candidates to fulfill this task: they generalize the classical framework by allowing the assumed and true models to differ, yet they

establish performance limits on the estimation error covariance in a way that indicates how the difference between the true and assumed models affects the estimation performance. Having established the main motivations, we can now briefly review the literature on the estimation framework under model misspecification, with a focus on the two classical building blocks, i.e., the ML estimator and the CRB.

Some historical background

The first fundamental result on the behavior of the ML estimator under misspecification appeared in the statistical literature in 1967 and was provided by Huber [20]. In that paper, the consistency and the normality of the ML estimator were proved under very mild regularity conditions. Five years later, Akaike [1] highlighted the link between Huber’s findings and the Kullback–Leibler divergence (KLD) [7]. He noted that the convergence point of the ML estimator under model misspecification could be interpreted as the point that minimizes the KLD between the true and the assumed models. In the early 1980s, these ideas were further developed by White [46], where the term *quasi-ML estimator* was introduced. Some years later, the second fundamental building block of an estimation theory under model misspecification

was established by Vuong [45]. Vuong was the first to derive a generalization of the Cramér–Rao lower bound under misspecified models. The Bayesian misspecified estimation problem has been investigated in [4] and [6].

Quite surprisingly and despite the wide variety of potential applications, the SP community has remained largely unaware of these fundamental results. This topic has only recently been rediscovered and its

applications to well-known SP problems investigated [10]–[12], [14], [18], [19], [22], [28], [32], [35]–[38], [48], [50]. Of course, every SP practitioner was aware of the misspecification problem, but some approaches commonly used within the SP community to address it differed from some of those proposed in the statistical literature. The effect of the misspecification has been modeled by adding into the true data model some random quantities, also called *nuisance parameters*, and by transforming the estimation problem at hand into a higher dimensional hybrid estimation problem. The performance degradation due to the augmented level of uncertainty generated by the nuisance parameters could be assessed by evaluating the true CRB when possible, the hybrid CRB (see, e.g., [16], [29], [31], and [39]), or the modified CRB (see, e.g., [2], [17], and [24]). This approach, although reasonable, is application dependent and not general at all. Other approaches include sensitivity analyses [15], [44].

Finally, the relationship between misspecified estimation theory and robust statistics should also be noted (see [49] for a tutorial on robust statistics). As one would expect, these two frameworks share the same motivations, i.e., an imperfect knowledge of the true data model. The aim of robust

The mathematical basis for a formal theory of statistical inference was presented by Fisher, who introduced the maximum likelihood method along with its main properties.

estimation theory is to develop estimation algorithms that are capable of achieving good performance over a large set of allowable input data models, even if they are suboptimal under any nominal (or true) model. Even though the development of robust estimators is certainly vital in many SP applications, for some of these, the mathematical derivation and consequent implementation may be too involved or too time and hardware intensive. In these cases, as discussed before, one may prefer to apply the classical, nonrobust estimation theory by assuming a simplified, hence, misspecified, statistical model for the data.

The first aim of this article is to summarize the most relevant existing works in the statistical literature using a formalism that is more familiar to the SP community. The second is to show the potential application of misspecified estimation theory, in both the deterministic and Bayesian contexts, for various classical SP problems.

Description of a misspecified model problem

Let $\mathbf{x}_1, \dots, \mathbf{x}_M$ be a set of N -dimensional (generally complex) random vectors representing the outcome of a measurement process. Let $\mathbf{x}_m \in \mathbb{C}^N$ be a single observation vector with pdf $p_X(\mathbf{x}_m)$ belonging to a possibly parametric family, or model, \mathcal{P} that characterizes the observed random experiment. As discussed in the section “A Formal Theory of Statistical Inference Under Misspecified Models,” in almost all practical applications, the true pdf $p_X(\mathbf{x}_m)$ is either not perfectly known, or it does not admit a simple derivation or easy implementation of the estimation algorithm. Thus, instead of $p_X(\mathbf{x}_m)$, in the mismatched estimation framework, we adopt a different parametric pdf, say, $f_X(\mathbf{x}_m|\boldsymbol{\theta})$, with $\boldsymbol{\theta} \in \Theta \subset \mathbb{R}^d$, to characterize the statistical behavior of the data vector \mathbf{x}_m . Potential estimation algorithms may be derived from the misspecified parametric pdf $f_X(\mathbf{x}_m|\boldsymbol{\theta})$, belonging to a parametric model \mathcal{F} , and not from the true pdf $p_X(\mathbf{x}_m)$. Moreover, we assume that $f_X(\mathbf{x}_m|\boldsymbol{\theta})$ could differ from $p_X(\mathbf{x}_m)$ for every $\boldsymbol{\theta} \in \Theta$. Since this assumption represents the division between the classical matched and the misspecified parametric estimation theories, some additional comments are warranted. The matched estimation theory requires the existence of at least a parameter vector $\bar{\boldsymbol{\theta}} \in \Theta$ for which the pdf assumed by the SP practitioner is equal to the true one. Mathematically, we can say that the classical matched theory holds true if, for some $\bar{\boldsymbol{\theta}} \in \Theta$, $p_X(\mathbf{x}_m) = f_X(\mathbf{x}_m|\bar{\boldsymbol{\theta}})$ or, equivalently, if $p_X(\mathbf{x}_m) \in \mathcal{F}$. For example, suppose the collected data, i.e., the outcomes of a random experiment, are distributed according to a univariate Gaussian distribution with the mean value $\bar{\mu}$ and variance $\bar{\sigma}^2$, i.e., $x_m \sim p_X(x_m) = \mathcal{N}(\bar{\mu}, \bar{\sigma}^2)$, $m = 1, \dots, M$. Moreover, suppose that the assumed parametric model for data inference is the Gaussian parametric model, i.e., $\mathcal{F} = \{f_X | f_X(x_m|\boldsymbol{\theta}) = \mathcal{N}(\theta_1, \theta_2) \forall \boldsymbol{\theta} \in \mathbb{R} \times \mathbb{R}^+\}$, where \mathbb{R}^+ is the set of positive real numbers. This situation clearly represents

a matched case, since there exists $\bar{\boldsymbol{\theta}} = (\bar{\mu}, \bar{\sigma}^2) \in \mathbb{R} \times \mathbb{R}^+$ such that $p_X(x_m) = f_X(x_m|\bar{\boldsymbol{\theta}}) = \mathcal{N}(\bar{\mu}, \bar{\sigma}^2)$. Suppose now that the collected data are distributed according to a univariate Laplace distribution with a location parameter $\bar{\gamma}$ and a scale parameter $\bar{\beta}$, i.e., $x_m \sim p_X(x_m) = \mathcal{L}(\bar{\gamma}, \bar{\beta})$. Due to, perhaps, misleading and incomplete information on the experiment at hand or due to the need to derive a simple algorithm, we decide to adopt a parametric Gaussian model \mathcal{F} to characterize the collected data. Unlike the previous example, this is obviously a mismatched case, since there does not exist any $\boldsymbol{\theta} = (\theta_1, \theta_2)$ for which the assumed Gaussian model is equal to the true Laplace model.

Many practical examples of model misspecification can be found in everyday engineering practices. Just to list a few, recent papers have investigated the application of this misspecified model framework to

- the direction-of-arrival (DOA) estimation problem in sensor arrays [22], [36], [37] and multiple-input, multiple-output (MIMO) radars [35]
- the covariance matrix estimation problem in non-Gaussian disturbance [10], [12], [18]
- radar-communication systems coexistence [38]
- waveform parameter estimation in the presence of uncertainty in the propagation model [32]
- the time-of-arrival estimation problem for ultra-wideband signals in the presence of interference [19].

In “The Misspecified CRB” and “The Mismatched ML Estimator” sections, the parameter vector $\boldsymbol{\theta}$ is assumed to be an unknown and deterministic real vector. The extension to the Bayesian case is discussed

in the “Generalization to the Bayesian Setting” section. Suppose, for inference purposes, we collect M independent, identically distributed (i.i.d.) measurement vectors $\mathbf{x} = \{\mathbf{x}_m\}_{m=1}^M$, where $\mathbf{x}_m \sim p_X(\mathbf{x}_m)$. Due to the independence, the true joint pdf of the data set \mathbf{x} can be expressed as the product of the marginal pdf as $p_X(\mathbf{x}) = \prod_{m=1}^M p_X(\mathbf{x}_m)$. The assumed joint pdf of the data set is instead $f_X(\mathbf{x}|\boldsymbol{\theta}) = \prod_{m=1}^M f_X(\mathbf{x}_m|\boldsymbol{\theta})$. This misspecified model framework raises three important questions:

- Is it still possible to derive lower bounds on the error covariance of any mismatched estimator of the parameter vector $\boldsymbol{\theta}$?
- How will the classical statistical properties of an estimator, e.g., unbiasedness, consistency, and efficiency, change in this misspecified model framework?
- How meaningful are the parameter estimates under extreme cases of mismatch?

The misspecified CRB

This section introduces a version of the CRB accounting for possible model misspecification, i.e., the misspecified CRB (MCRB), which can be considered a generalization of the usual CRB. In particular, as we will show later, the CRB is obtained when the model is correctly specified. We start by providing the

The concept of efficiency is strictly related to the existence of some lower bounds on the performance of any estimator designed for a specific inference task.

required regularity conditions and the notion of unbiasedness for mismatched estimators.

Regular models

As with the classical CRB, to guarantee the existence of the MCRB, some regularity conditions on the assumed pdf need to be imposed. Specifically, the assumed parametric model \mathcal{F} must be regular w.r.t. \mathcal{P} , i.e., the family to which the true pdf belongs. The complete list of assumptions that \mathcal{F} must satisfy to be regular w.r.t. \mathcal{P} are given in [45] and briefly recalled in [10]. Most of them are rather technical and facilitate an order reversal of the integral and derivative operators. Nevertheless, there are two assumptions that need to be discussed here due to their importance in the development of the theory. The first condition that must be satisfied is Assumption 1.

Assumption 1

There exists a unique interior point θ_0 of Θ such that

$$\theta_0 = \underset{\theta \in \Theta}{\operatorname{argmin}} \{-E_p\{\ln f_X(\mathbf{x}_m|\theta)\}\} = \underset{\theta \in \Theta}{\operatorname{argmin}} \{D(p_X\|f_X|\theta)\}, \quad (1)$$

where $E_p\{\cdot\}$ indicates the expectation operator of a vector- or scalar-valued function w.r.t. the pdf $p_X(\mathbf{x}_m)$ and $D(p_X\|f_X|\theta) \triangleq \int \ln(p_X(\mathbf{x}_m)/f_X(\mathbf{x}_m|\theta))p_X(\mathbf{x}_m)d\mathbf{x}_m$ is the KLD [7] between the true and the assumed pdfs. As indicated by (1), θ_0 can be interpreted as the point that minimizes the KLD between $p_X(\mathbf{x}_m)$ and $f_X(\mathbf{x}_m|\theta)$, and it is called the *pseudotrue parameter vector* [45], [46].

After having defined the pseudotrue parameter vector θ_0 in this assumption, let \mathbf{A}_{θ_0} be the matrix whose entries are defined as

$$[\mathbf{A}_{\theta_0}]_{ij} \triangleq [E_p\{\nabla_{\theta} \nabla_{\theta}^T \ln f_X(\mathbf{x}_m|\theta_0)\}]_{ij} = E_p\left\{\left.\frac{\partial^2}{\partial \theta_i \partial \theta_j} \ln f_X(\mathbf{x}_m|\theta)\right|_{\theta=\theta_0}\right\}, \quad (2)$$

where $\nabla_{\theta} u(\theta_0)$ and $\nabla_{\theta} \nabla_{\theta}^T u(\theta_0)$ indicate, respectively, the gradient (column) vector and the symmetric Hessian matrix of the scalar function u evaluated at θ_0 . The second fundamental condition that must be satisfied by the assumed model \mathcal{F} to be regular w.r.t. \mathcal{P} is Assumption 2.

Assumption 2

The matrix \mathbf{A}_{θ_0} is nonsingular.

The pseudotrue parameter vector θ_0 plays a fundamental role in estimation theory for misspecified models. Roughly speaking, it identifies the pdf $f_X(\mathbf{x}_m|\theta_0)$ in the assumed parametric model \mathcal{F} that is closest, in the KLD sense, to the true pdf. As the next sections will clarify, it can be interpreted as the counterpart of the true parameter vector of the classical matched theory. Regarding the matrix \mathbf{A}_{θ_0} , its negative represents a generalization of the classical Fisher information matrix (FIM) to the misspecified model framework. To clarify this, we first define the matrix \mathbf{B}_{θ_0} as

$$[\mathbf{B}_{\theta_0}]_{ij} \triangleq [E_p\{\nabla_{\theta} \ln f_X(\mathbf{x}_m|\theta_0) \nabla_{\theta}^T \ln f_X(\mathbf{x}_m|\theta_0)\}]_{ij} = E_p\left\{\left.\frac{\partial \ln f_X(\mathbf{x}_m|\theta)}{\partial \theta_i}\right|_{\theta=\theta_0} \cdot \left.\frac{\partial \ln f_X(\mathbf{x}_m|\theta)}{\partial \theta_j}\right|_{\theta=\theta_0}\right\}. \quad (3)$$

As with matrix \mathbf{A}_{θ_0} , we recognize in \mathbf{B}_{θ_0} the second possible generalization of the FIM. Vuong [45] showed that if $p_X(\mathbf{x}_m) = f_X(\mathbf{x}_m|\bar{\theta})$ for some $\bar{\theta} \in \Theta$, then $\theta_0 = \bar{\theta}$ and $\mathbf{B}_{\theta_0} = -\mathbf{A}_{\theta_0}$, where $\bar{\theta}$ is the true parameter vector of the classical matched theory. The last equation shows that, under the correct model specification, the two expressions of the FIM are equal, as expected [44]. This provides evidence of the fact that the misspecified estimation theory is consistent with the classical one. The reader, however, should note that the equality between the pseudotrue parameter vector and the true one does not imply in any way the equality between the true and the assumed pdfs and, consequently, between the matrices \mathbf{B}_{θ_0} and $-\mathbf{A}_{\theta_0}$. After having established the necessary regularity conditions, we can introduce the class of misspecified-unbiased (MS-unbiased) estimators.

The MS-unbiasedness property

The first generalization of the classical unbiasedness property to mismatched estimators was proposed by Vuong [45]. Specifically, let $\hat{\theta}(\mathbf{x})$ be an estimator of the pseudotrue parameter vector θ_0 , i.e., a function of the M available i.i.d. observation vectors $\mathbf{x} = \{\mathbf{x}_m\}_{m=1}^M$, derived under the misspecified parametric model \mathcal{F} . Then, $\hat{\theta}(\mathbf{x})$ is said to be an MS-unbiased estimator if and only if

$$E_p\{\hat{\theta}(\mathbf{x})\} = \int \hat{\theta}(\mathbf{x})p_X(\mathbf{x})d\mathbf{x} = \theta_0, \quad (4)$$

where θ_0 is the pseudotrue parameter vector defined in (1). The link with the classical matched unbiasedness property is obvious: if the parametric model \mathcal{F} is correctly specified, θ_0 is equal to the vector $\bar{\theta} \in \Theta$ such that $p_X(\mathbf{x}_m) = f_X(\mathbf{x}_m|\bar{\theta})$. Consequently, (4) can be rewritten as $E_p\{\hat{\theta}(\mathbf{x})\} = \int \hat{\theta}(\mathbf{x})f_X(\mathbf{x}|\bar{\theta})d\mathbf{x} = \bar{\theta}$, which is the classical definition of the unbiasedness property. At this point, we are ready to introduce the explicit expression for the MCRB.

A covariance inequality in the presence of misspecified models

In this section, we present the MCRB as introduced by Vuong in his seminal paper [45]. An alternative derivation was proposed by Richmond and Horowitz in [36] and [37]. A comparison between the derivation given in [45] and the one proposed in [36] and [37] has been provided in [10].

Theorem 1

In Theorem 1 [45], let \mathcal{F} be a misspecified parametric model that is regular w.r.t. \mathcal{P} . Let $\hat{\theta}(\mathbf{x})$ be an MS-unbiased estimator derived under the misspecified model \mathcal{F} from a set of M i.i.d. observation vectors $\mathbf{x} = \{\mathbf{x}_m\}_{m=1}^M$. Then, for every possible $p_X(\mathbf{x}_m) \in \mathcal{P}$,

$$C_p(\hat{\theta}(\mathbf{x}), \theta_0) \geq \frac{1}{M} \mathbf{A}_{\theta_0}^{-1} \mathbf{B}_{\theta_0} \mathbf{A}_{\theta_0}^{-1} \triangleq \text{MCRB}(\theta_0), \quad (5)$$

where

$$\mathbf{C}_p(\hat{\boldsymbol{\theta}}(\mathbf{x}), \boldsymbol{\theta}_0) \triangleq E_p\{(\hat{\boldsymbol{\theta}}(\mathbf{x}) - \boldsymbol{\theta}_0)(\hat{\boldsymbol{\theta}}(\mathbf{x}) - \boldsymbol{\theta}_0)^T\} \quad (6)$$

is the error covariance matrix of the mismatched estimator $\hat{\boldsymbol{\theta}}(\mathbf{x})$ where the matrices $\mathbf{A}_{\boldsymbol{\theta}_0}$ and $\mathbf{B}_{\boldsymbol{\theta}_0}$ have been defined in (2) and (3), respectively.

The following comments are in order. The major implication of Theorem 1 is that it is still possible to establish a lower bound on the error covariance matrix of an (MS-unbiased) estimator, even if it is derived under a misspecified data model, i.e., it is derived under a pdf $f_X(\mathbf{x}_m|\boldsymbol{\theta})$ that could differ from the true pdf $p_X(\mathbf{x}_m)$ for every value of $\boldsymbol{\theta}$ in the parameter space Θ . An important question that may arise under a misspecified model framework is which vector in the assumed parameter space Θ should be used to evaluate the effectiveness of a mismatched estimator, particularly when no true parameter vector exists, i.e., $p_X(\mathbf{x}_m) \neq f_X(\mathbf{x}_m|\boldsymbol{\theta})$, for all $\boldsymbol{\theta} \in \Theta$? It is certainly reasonable to use the parameter value that minimizes the distance, in a given sense, between the assumed misspecified pdf $f_X(\mathbf{x}_m|\boldsymbol{\theta})$ and the true pdf $p_X(\mathbf{x}_m)$. Theorem 1 shows that, if one uses the KLD as a measure of distance and by assuming that the misspecified model \mathcal{F} is regular w.r.t. the true model \mathcal{P} , this parameter vector exists, and it is the pseudotrue parameter vector $\boldsymbol{\theta}_0$ defined in (1). Specifically, the MCRB is a lower bound on the error covariance matrix of any MS-unbiased estimator, where the error is defined as the difference between the estimator and the pseudotrue parameter vector. Moreover, if the model \mathcal{F} is correctly specified, then, as stated before, $\boldsymbol{\theta}_0 = \bar{\boldsymbol{\theta}}$, such that $p_X(\mathbf{x}_m) = f_X(\mathbf{x}_m|\bar{\boldsymbol{\theta}})$ and $\mathbf{B}_{\boldsymbol{\theta}_0} = \mathbf{B}_{\bar{\boldsymbol{\theta}}} = -\mathbf{A}_{\bar{\boldsymbol{\theta}}}$. Consequently, the inequality in (5) becomes the classical (matched) CRB inequality for unbiased estimators

$$E_p\{(\hat{\boldsymbol{\theta}}(\mathbf{x}) - \bar{\boldsymbol{\theta}})(\hat{\boldsymbol{\theta}}(\mathbf{x}) - \bar{\boldsymbol{\theta}})^T\} \geq \frac{1}{M} \mathbf{B}_{\bar{\boldsymbol{\theta}}}^{-1} = -\frac{1}{M} \mathbf{A}_{\bar{\boldsymbol{\theta}}}^{-1} \triangleq \text{CRB}(\bar{\boldsymbol{\theta}}). \quad (7)$$

The second point concerns the matter of how to exploit Theorem 1 in practice. The MCRB is a generalization of the classical CRB to the misspecified model framework and can play a similar role. Specifically, the MCRB can be used to assess the performance of any mismatched estimator, and it plays the same key role as the classical CRB in any feasibility study, but with the added flexibility to assess performance under modeling errors. For example, consider the recurring scenario in which the SP practitioner is aware of the true data pdf $p_X(\mathbf{x}_m)$, but, to fulfill some operational constraints, the user is forced to derive the required estimator by exploiting a simpler, but misspecified, model. In this scenario, the MCRB in (5) can be directly applied to assess the potential estimation loss due to the mismatch between the assumed and the true models.

This scenario can be extended to the case in which the SP practitioner is not completely aware of the functional form of the true data pdf, but the user is still able to infer some of its properties, e.g., from empirical data or parameter estimates based on such data. Such knowledge can be

used to motivate surrogate models for the true data pdf, which in turn can be exploited to conduct a system analysis and performance assessment. To clarify this point, consider the case in which the SP practitioner, to derive a simple inference algorithm, decides to assume a Gaussian model to describe the data behavior. However, thanks to a preliminary data analysis, the user knows that the data share a heavy-tailed distribution, e.g., due to the presence of impulsive non-Gaussian noise. Then, the user could choose as true data pdf a heavy-tailed distribution, e.g., the t -distribution, and, consequently, exploit the MCRB to assess how ignoring the heavy-tailed and impulsive nature of the data affects the performance of the estimation algorithm based on a Gaussian model. This explains that, although the chosen “true” pdf (in this example, the t -distribution) may not be the exact true data pdf, it can still serve as a useful surrogate for the purposes of system analysis and design by means of the MCRB.

The MCRB can also be used to predict potential weaknesses (i.e., a breakdown of the estimation performance) of a system. Suppose one has a system/estimator derived under a certain modeling assumption, but it is of practical interest to predict how well this system will react in the presence of different true input data distributions, perhaps characterizing operational scenarios that the system can undergo. Clearly, the MCRB is well suited to address this task.

Another important question may arise analyzing Theorem 1. To evaluate the pseudotrue parameter vector $\boldsymbol{\theta}_0$ in (1) and then the MCRB in (5), we need to know the true data pdf $p_X(\mathbf{x}_m)$, since it is required to evaluate the expectation operators. How can we calculate the MCRB in all the practical cases in which we haven't any a priori knowledge of the functional form of $p_X(\mathbf{x}_m)$? An answer to this fundamental question is given in the section “A Consistent Sample Estimate of the MCRB,” where we show that consistent estimators for both the pseudotrue parameter vector $\boldsymbol{\theta}_0$ and the MCRB can be derived from the acquired data set.

The proposed MCRB can be easily extended to misspecified estimation problems that require equality constraints. We refer the reader to [11] for a comprehensive treatise on this problem. Additionally, with regard to the possibility of generalizing the previously discussed results to the case of complex unknown parameter vectors, the extension to the complex fields can be achieved in two equivalent ways. We can always map a complex parameter vector into a real one simply by stacking its real and imaginary parts, as, e.g., in [35], or we could exploit the so-called Wirtinger calculus, as discussed in [13] and [37].

An interesting case: A lower bound on the MSE via the MCRB

In this section, we focus on a mismatched case that is of great interest in many practical applications. Specifically, we consider the case in which the parameter vector of the assumed model \mathcal{F} is nested in the one of the true parametric model \mathcal{P} , i.e., the assumed parameter space Θ is a subspace of the true

parameter space $\Omega = \Theta \times \Gamma$, where \times indicates the Cartesian product. Under this restriction, the true parametric model can be expressed as

$$\mathcal{P} = \{p_X | p_X(\mathbf{x}_m | \boldsymbol{\theta}, \boldsymbol{\gamma}) \text{ is a pdf } \forall (\boldsymbol{\theta}, \boldsymbol{\gamma}) \in \Theta \times \Gamma\}, \quad (8)$$

while the assumed model is $\mathcal{F} = \{f_X | f_X(\mathbf{x}_m | \boldsymbol{\theta}) \text{ is a pdf } \forall \boldsymbol{\theta} \in \Theta\}$ as before. Note that $f_X(\mathbf{x}_m | \boldsymbol{\theta})$ could differ from the true $p_X(\mathbf{x}_m | \boldsymbol{\theta}, \boldsymbol{\gamma}) \forall \boldsymbol{\theta} \in \Theta$ and $\forall \boldsymbol{\gamma} \in \Gamma$. Moreover, the nested parameter vector assumption includes, as a special

case, the scenario in which the true parameter space and the assumed one are equal, i.e., $\Omega \equiv \Theta$. This case arises, for example, in array processing applications in which both the true and the assumed pdfs of the acquired data vectors can be parameterized by the angles of arrival of a certain number of sources [37]. A practical example of the more general nested model assumption is the estimation of the disturbance covariance matrix in adaptive radar detection [10]. In this misspecified estimation problem, both the unknown true data pdf and the assumed one can be parameterized by a scaled version of the covariance matrix and by the disturbance power. Both of

Variance Estimation

Here is an illustrative example to clarify the use and derivation of the misspecified Cramér–Rao bound (MCRB). Building upon the examples discussed in [10], we investigate the problem of estimating the variance of a Gaussian-distributed data set under the misspecification of the mean value. Let $\mathbf{x} = \{x_m\}_{m=1}^M$ be a set of M independent, identically distributed (i.i.d.) univariate data sampled from a Gaussian probability density function (pdf) with the mean value $\bar{\mu}$ and variance $\bar{\sigma}^2$, i.e., $p_X(x_m) \equiv \mathcal{N}(\bar{\mu}, \bar{\sigma}^2) \in \mathcal{P}$ with $\bar{\mu} \neq 0$. Due to perhaps an imperfect knowledge about the data generation process, the user assumes a zero-mean parametric Gaussian model $\mathcal{F} = \{f_X | f_X(x_m | \theta) = \mathcal{N}(0, \theta) \forall \theta \in \Theta \subseteq \mathbb{R}^+\}$, i.e., the user misspecifies the mean value. Note that, as long as $\bar{\mu} \neq 0$, the true but unknown pdf $p_X(x_m)$ does not belong to the assumed model \mathcal{F} . Moreover, the reader can easily recognize this mismatched scenario as a simple instance of the particular case discussed in the section “An Interesting Case: A Lower Bound on the MSE via the MCRB.” In fact, it is immediately verified that the parameter space $\Theta \subseteq \mathbb{R}^+$ that characterizes the assumed model is a subset of the true parameter space, i.e., $[\bar{\mu}, \bar{\sigma}^2] \in \Omega = \mathbb{R}_0 \times \Theta$, where \mathbb{R}_0 indicates the set of all the real numbers excluding 0.

According to the theory presented in the section “The Misspecified CRB,” we first must check whether the assumed model \mathcal{F} is regular with respect to (w.r.t.) \mathcal{P} ; in other words, we have to prove the existence of the pseudotrue parameter θ_0 (Assumption 1) and the nonsingularity of the matrix \mathbf{A}_{θ_0} defined in (2) (Assumption 2). Note that, for the problem at hand, \mathbf{A}_{θ_0} is a scalar quantity, so we have to prove that $A_{\theta_0} \neq 0$. The pseudotrue parameter θ_0 is defined in (1). Following [7], the Kullback–Leibler divergence (KLD) can be expressed as

$$D(p_X || f_X | \theta) = \frac{\bar{\mu}^2}{2\theta} + \frac{1}{2} \left(\frac{\bar{\sigma}^2}{\theta} - 1 - \ln \frac{\bar{\sigma}^2}{\theta} \right). \quad (S1)$$

The minimum is obtained for $\theta_0 = \bar{\sigma}^2 + \bar{\mu}^2$, which, according to (1), represents the pseudotrue parameter. Since the pseudotrue parameter exists and is unique, Assumption 1 is satisfied. We can now check Assumption 2. To this end, from (2), A_{θ_0} can be evaluated as

$$A_{\theta_0} \triangleq E_P \left\{ \frac{\partial^2 \ln f_X(x_m | \theta)}{\partial \theta^2} \Big|_{\theta=\theta_0} \right\} = \frac{1}{2\theta_0^2} - \frac{1}{\theta_0^3} E_P \{x_m^2\} = -\frac{1}{2\theta_0^2}, \quad (S2)$$

yielding a denominator different from zero since $\bar{\sigma}^2 \in \mathbb{R}^+$; consequently, Assumption 2 is verified as well. Now we can evaluate the MCRB in (5) for the estimation problem at hand. First, the scalar B_{θ_0} can be easily evaluated from (3) as

$$B_{\theta_0} \triangleq E_P \left\{ \left(\frac{\partial \ln f_X(x_m | \theta)}{\partial \theta} \right)^2 \Big|_{\theta=\theta_0} \right\} = \frac{\theta_0^2 + E_P \{x_m^4\} - 2\theta_0 E_P \{x_m^2\}}{4\theta_0^4} = \frac{2\bar{\sigma}^4 + 4\bar{\sigma}^2 \bar{\mu}^2}{4\theta_0^4}. \quad (S3)$$

Finally, from (5), we get

$$\text{MCRB}(\theta_0) = \frac{2\bar{\sigma}^4}{M} + \frac{4\bar{\sigma}^2 \bar{\mu}^2}{M}. \quad (S4)$$

Since this misspecified scenario belongs to the particular class of nested parametric models, as discussed in the section “An Interesting Case: A Lower Bound on the MSE via the MCRB,” we can also rewrite the MCRB in (S4) as a function of the true variance $\bar{\sigma}^2$. This can be easily done by introducing the (scalar) $r \triangleq \bar{\sigma}^2 - \theta_0 = -\bar{\mu}^2$ and, consequently, according to (9), by evaluating the $LB(\bar{\sigma}^2)$ as

$$LB(\bar{\sigma}^2) = \frac{2\bar{\sigma}^4}{M} + \frac{4\bar{\sigma}^2 \bar{\mu}^2}{M} + \bar{\mu}^4. \quad (S5)$$

It can be noted that the lower bound in (S5) is always greater than the classical CRB given by $\text{CRB}(\bar{\sigma}^2) = 2\bar{\sigma}^4/M$

these applications will be discussed in the “Examples of Applications” section, while here we focus on the theoretical implications of the condition in (8). The first immediate consequence of (8) is the fact that if the pseudotrue parameter vector θ_0 and the true parameter subvector $\tilde{\theta}$ belong to the same parameter space Θ , then the difference vector $\mathbf{r} \triangleq \tilde{\theta} - \theta_0$ is well defined, but, in general, it is different from a zero-vector. As shown in [10, Sect. II.D] or in [37, eq. (70)], using \mathbf{r} , a bound on the MSE of the estimate of the true parameter vector $\tilde{\theta}$ under model misspecification can be easily established as

$$\begin{aligned} \text{MSE}_p(\hat{\theta}(\mathbf{x}), \tilde{\theta}) &\triangleq E_p \{ (\hat{\theta}(\mathbf{x}) - \tilde{\theta})(\hat{\theta}(\mathbf{x}) - \tilde{\theta})^T \} \\ &= C_p(\hat{\theta}(\mathbf{x}), \theta_0) + \mathbf{r}\mathbf{r}^T \geq \frac{1}{M} \mathbf{A}_{\theta_0}^{-1} \mathbf{B}_{\theta_0} \mathbf{A}_{\theta_0}^{-1} + \mathbf{r}\mathbf{r}^T \\ &\triangleq \text{LB}(\tilde{\theta}). \end{aligned} \tag{9}$$

Note that, here, the lower bound (denoted as LB) $\text{LB}(\tilde{\theta}) = \text{MCRB}(\theta_0) + \mathbf{r}\mathbf{r}^T$ is considered as a function of the true parameter vector $\tilde{\theta}$. A simple example that clarifies the role of the inequality (9) as lower bound on the MSE is reported in “Variance Estimation.”

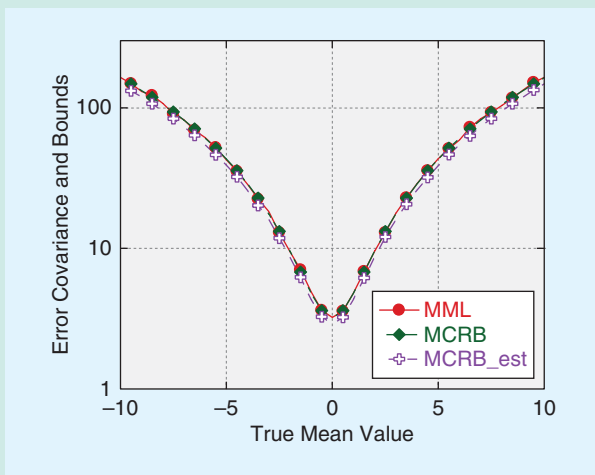


FIGURE S1. The error covariance of the MML estimator, the $\text{MCRB}(\theta_0)$, and the estimated $\text{MCRB}(\theta_0)$ as a function of $\bar{\mu}$. Simulation parameters are set as $M = 10$ and $\bar{\sigma}^2 = 4$.

and they are equal only in the case of perfect model specification, i.e., when the true mean is equal to the assumed mean, i.e., $\bar{\mu} = 0$.

After having established a lower bound on the mean square error (MSE), we now investigate the properties of the mismatched maximum likelihood (MML) estimator for the estimation problem at hand. In particular, we can say that the MML estimator is not consistent since, from (11), it converges to θ_0 , which is different from the true variance $\bar{\sigma}^2$. More formally, we have that

$$\hat{\theta}_{\text{MML}} \triangleq \hat{\theta}_{\text{MML}}(\mathbf{x}) = M^{-1} \sum_{m=1}^M x_m^2 \xrightarrow[M \rightarrow \infty]{a.s.} \theta_0 = \bar{\sigma}^2 + \bar{\mu}^2 \neq \bar{\sigma}^2. \tag{S6}$$

However, according to (4), the MML estimator is misspecified (MS)-unbiased, since

$$E_p \{ \hat{\theta}_{\text{MML}} \} = E_p \left\{ M^{-1} \sum_{m=1}^M x_m^2 \right\} = \bar{\sigma}^2 + \bar{\mu}^2 = \theta_0. \tag{S7}$$

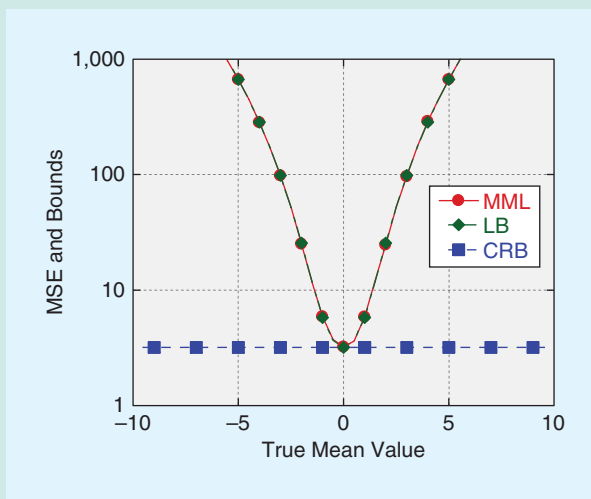


FIGURE S2. The MSE of the MML estimator, the $\text{LB}(\bar{\sigma}^2)$, and the $\text{CRB}(\bar{\sigma}^2)$, as a function of $\bar{\mu}$. Simulation parameters are set as $M = 10$ and $\bar{\sigma}^2 = 4$.

Hence, according to Theorem 1, its error covariance w.r.t. θ_0 , i.e., $C_p(\hat{\theta}_{\text{MML}}, \theta_0)$, is lower bounded by the MCRB in (S4). Figure S1 shows the error covariance of the MML estimator, the $\text{MCRB}(\theta_0)$, and the sample estimate of $\text{MCRB}(\theta_0)$ obtained according to (13)–(15). As we can see, $\text{MCRB}(\theta_0)$ is a tight bound for the error variance of the MML estimator, and the sample $\text{MCRB}(\theta_0)$ accurately predicts it. Due to the particular nested structure of the true and assumed parameter spaces of this example, we can also evaluate the MSE of the MML estimator w.r.t. the true variance, i.e., $\text{MSE}_p(\hat{\theta}_{\text{MML}}, \bar{\sigma}^2)$, and the related $\text{LB}(\bar{\sigma}^2)$ obtained as shown in (9). Note that the lower bound is denoted as LB.

In Figure S2, we report the MSE of the MML estimator, the $\text{LB}(\bar{\sigma}^2)$, and the classical CRB on the estimation of the variance, $\text{CRB}(\bar{\sigma}^2)$, as function of the value of the true mean value $\bar{\mu}$. As expected from (9), $\text{LB}(\bar{\sigma}^2)$ is a tight bound for the MSE of the MML estimator. Finally, it can be noted that the $\text{LB}(\bar{\sigma}^2)$ is equal to the $\text{CRB}(\bar{\sigma}^2)$ only when $\bar{\mu} = 0$, i.e., when the assumed mean value is equal to the true one.

The mismatched ML estimator

The aim of this section is to present the second milestone of the estimation theory under model misspecification: the mismatched ML (MML) estimator. As discussed in the section “A Formal Theory of Statistical Inference Under Misspecified Models,” the theoretical framework supporting the existence and the convergence properties of the MML estimator were developed by Huber [20] and later by White [46]. Here, our goal is to summarize their main findings from an SP standpoint. As detailed in the “Description of a Misspecified Model Problem” section, assume we have a set $\mathbf{x} = \{\mathbf{x}_m\}_{m=1}^M$ of M i.i.d. measurement vectors distributed according to a true, but unknown or inaccessible, pdf $p_X(\mathbf{x}_m)$. So the log-likelihood function for the data \mathbf{x} under a generally misspecified parametric pdf $f_X(\mathbf{x}_m|\boldsymbol{\theta}) \in \mathcal{F}$ is given by $l_M(\boldsymbol{\theta}) \triangleq M^{-1} \sum_{m=1}^M \ln f_X(\mathbf{x}_m|\boldsymbol{\theta})$. Following the classical definition, the MML estimate is the vector that maximizes the (misspecified) log-likelihood function

$$\hat{\boldsymbol{\theta}}_{\text{MML}}(\mathbf{x}) \triangleq \underset{\boldsymbol{\theta} \in \Theta}{\operatorname{argmax}} l_M(\boldsymbol{\theta}) = \underset{\boldsymbol{\theta} \in \Theta}{\operatorname{argmax}} \sum_{m=1}^M \ln f_X(\mathbf{x}_m|\boldsymbol{\theta}), \quad (10)$$

where $\mathbf{x}_m \sim p_X(\mathbf{x}_m)$. The definition of the MML estimator given in (10) is clear and self-explanatory. Furthermore, it is consistent with the classical “matched” ML estimator. But what is the convergence point of $\hat{\boldsymbol{\theta}}_{\text{MML}}(\mathbf{x})$? As proved in [20] and [46], under suitable regularity conditions, the MML estimator converges [almost surely (a.s.)] to the pseudotrue parameter vector $\boldsymbol{\theta}_0$ defined in (1). This is a desirable result since it shows that the MML estimator converges to the parameter vector that minimizes the distance, in the KLD sense, between the misspecified and the true pdfs (see “Variance Estimation” and “Power Estimation in Correlated Data”). In addition, Huber and White investigated the asymptotic behavior of the MML estimator, and their valuable findings can be summarized in the following theorem.

Theorem 2

For Theorem 2 [20], [46], under suitable regularity conditions, it can be shown that

$$\hat{\boldsymbol{\theta}}_{\text{MML}}(\mathbf{x}) \xrightarrow[M \rightarrow \infty]{a.s.} \boldsymbol{\theta}_0. \quad (11)$$

Moreover,

$$\sqrt{M}(\hat{\boldsymbol{\theta}}_{\text{MML}}(\mathbf{x}) - \boldsymbol{\theta}_0) \xrightarrow[M \rightarrow \infty]{d.} \mathcal{N}(\mathbf{0}, \mathbf{C}_{\boldsymbol{\theta}_0}), \quad (12)$$

where $\xrightarrow[M \rightarrow \infty]{d.}$ indicates the convergence in distribution and $\mathbf{C}_{\boldsymbol{\theta}_0} \triangleq \mathbf{A}_{\boldsymbol{\theta}_0}^{-1} \mathbf{B}_{\boldsymbol{\theta}_0} \mathbf{A}_{\boldsymbol{\theta}_0}^{-1}$, where the matrices $\mathbf{A}_{\boldsymbol{\theta}_0}$ and $\mathbf{B}_{\boldsymbol{\theta}_0}$ have been defined in (2) and (3), respectively. Matrix $\mathbf{C}_{\boldsymbol{\theta}_0}$ is sometimes referred to as *Huber’s sandwich covariance*. Two comments are in order:

1) The MML estimator is asymptotically MS-unbiased, and its asymptotic error covariance is equal to the MCRB, i.e., it is an efficient estimator w.r.t. the MCRB. The analogy with the classical matched ML estimator is completely transparent. In particular, if the model \mathcal{F} is correctly specified, i.e., there exists a parameter vector $\tilde{\boldsymbol{\theta}} \in \Theta$ such that $p_X(\mathbf{x}_m) = f_X(\mathbf{x}_m|\tilde{\boldsymbol{\theta}})$, then $\hat{\boldsymbol{\theta}}_{\text{MML}}(\mathbf{x}) \xrightarrow[M \rightarrow \infty]{a.s.} \tilde{\boldsymbol{\theta}}$ with an asymp-

totic error covariance matrix given by the classical CRB, which is the inverse of the FIM $\mathbf{B}_{\tilde{\boldsymbol{\theta}}} = -\mathbf{A}_{\tilde{\boldsymbol{\theta}}}$.

2) Theorem 2 represents a very useful result for practical applications. In fact, it tells us that, when we do not have any a priori information about the true data model, the ML estimator derived under a possibly misspecified model is still a reasonable choice among other MS-unbiased mismatched estimators, since it converges to the parameter vector that minimizes the KLD between the true and the assumed model and it has the lowest possible error covariance (at least asymptotically).

A consistent sample estimate of the MCRB

In this section, we go back to an issue raised before, i.e., the calculation of the MCRB when the true model is completely unknown. In fact, from (5), to obtain a closed form expression of the MCRB, we need to analytically evaluate $\boldsymbol{\theta}_0$, $\mathbf{A}_{\boldsymbol{\theta}_0}$, and $\mathbf{B}_{\boldsymbol{\theta}_0}$. As shown in (1)–(3), these quantities involve the analysis of the expectation operator taken w.r.t. the true pdf $p_X(\mathbf{x}_m)$. If $p_X(\mathbf{x}_m)$ is completely unknown, we will not be able to evaluate these expectations in closed form, but, as an alternative, we could obtain sample estimates of them. More formally, we define the matrices [46]:

$$[\mathbf{A}_M(\boldsymbol{\theta})]_{ij} \triangleq M^{-1} \sum_{m=1}^M \frac{\partial^2 \ln f_X(\mathbf{x}_m|\boldsymbol{\theta})}{\partial \theta_i \partial \theta_j}, \quad (13)$$

$$[\mathbf{B}_M(\boldsymbol{\theta})]_{ij} \triangleq M^{-1} \sum_{m=1}^M \frac{\partial \ln f_X(\mathbf{x}_m|\boldsymbol{\theta})}{\partial \theta_i} \cdot \frac{\partial \ln f_X(\mathbf{x}_m|\boldsymbol{\theta})}{\partial \theta_j}, \quad (14)$$

$$\mathbf{C}_M(\boldsymbol{\theta}) \triangleq [\mathbf{A}_M(\boldsymbol{\theta})]^{-1} \mathbf{B}_M(\boldsymbol{\theta}) [\mathbf{A}_M(\boldsymbol{\theta})]^{-1}. \quad (15)$$

Remarkably, it can be shown (see the proof in [46, Theorem 3.2]) that

$$\mathbf{C}_M(\hat{\boldsymbol{\theta}}_{\text{MML}}) \xrightarrow[M \rightarrow \infty]{a.s.} \mathbf{C}_{\boldsymbol{\theta}_0} = \text{MCRB}(\boldsymbol{\theta}_0). \quad (16)$$

In other words, (16) assures us that we can obtain a strongly consistent estimate of the MCRB by evaluating the sample counterpart of $\mathbf{A}_{\boldsymbol{\theta}_0}$ and $\mathbf{B}_{\boldsymbol{\theta}_0}$, i.e., $\mathbf{A}_M(\boldsymbol{\theta})$ and $\mathbf{B}_M(\boldsymbol{\theta})$, at the value of the MML estimator. This result has strong practical implications, since it provides an estimate of the MCRB when we do not have any prior knowledge of the true pdf $p_X(\mathbf{x}_m)$. Hence, it widens areas of applicability of the MCRB. This, of course, requires the data to be stationary over some reasonable period to allow sufficient averaging (as is required in numerous SP applications). This result can also be used to design statistical tests to detect model misspecification [46], [47, p. 218].

Generalization to the Bayesian setting

The Bayesian philosophy adopts the notion that one has some prior knowledge (a belief or perhaps a guess) about the values a desired parameter will assume before an experiment. Once data are observed, then one can update that prior knowledge based on the information provided by the data measurements. Thus, the

Power Estimation in Correlated Data

Another example that clarifies the theory concerns the estimation of the statistical power of a set of zero mean Gaussian vectors. Let $\mathbf{x} = \{\mathbf{x}_m\}_{m=1}^M$ be a set of M i.i.d. real N -dimensional random vectors sampled from a multivariate Gaussian pdf with a zero mean value and covariance matrix given by $\mathbf{M} = \sigma^2 \mathbf{\Sigma}$, i.e., $p_X(\mathbf{x}_m) \equiv \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{\Sigma}) \in \mathcal{P}$, where σ^2 is the statistical power and $\mathbf{\Sigma}$ is a symmetric, positive definite matrix whose trace is equal to N , i.e., $\text{tr}(\mathbf{\Sigma}) = N$. For simplicity, we assume that $[\mathbf{\Sigma}]_{ij} = \rho^{|i-j|}$, $i, j = 1, \dots, N$, where $|\rho| < 1$ is the one-lag correlation coefficient (this is the typical correlation matrix of an autoregressive process of order 1). Suppose now that the user is not aware of the data correlation structure and decides to assume the following parametric Gaussian model: $\mathcal{F} = \{f_X | f_X(\mathbf{x}_m | \theta) = \mathcal{N}(\mathbf{0}, \theta \mathbf{I}_N) \forall \theta \in \mathbb{R}^+\}$, where \mathbf{I}_N is the identity matrix of dimension N . Note that, as long as $\rho \neq 0$, the true pdf $p_X(\mathbf{x}_m)$ does not belong to the assumed model \mathcal{F} . We will proceed exactly as in "Variance Estimation" by checking the Assumptions 1 and 2 and then by evaluating the MML estimator and the relative MCRB.

To evaluate the pseudotrue parameter θ_0 , we need to find the minimum of the KLD between the true and assumed model. Following [7] again, the KLD between $\mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{\Sigma})$ and $\mathcal{N}(\mathbf{0}, \theta \mathbf{I}_N)$ is given by

$$D(p_X \| f_{X|\theta}) = \frac{1}{2} [\text{tr}(\theta^{-1} \sigma^2 \mathbf{\Sigma}) - N + \ln \theta - \ln \det(\sigma^2 \mathbf{\Sigma})]. \quad (S8)$$

Keeping in mind that $\text{tr}(\mathbf{\Sigma}) = N$, it is immediately verified that the minimum is given by $\theta_0 = \sigma^2$, i.e., the pseudotrue parameter is equal to the true power. After some basic calculus, the terms A_{θ_0} and B_{θ_0} are obtained as

$$A_{\theta_0} \triangleq E_p \left\{ \frac{\partial^2 \ln f_X(\mathbf{x}_m | \theta)}{\partial \theta^2} \Big|_{\theta=\theta_0} \right\} = \frac{N}{2\theta_0^2} - \frac{1}{\theta_0^3} E_p \{ \mathbf{x}_m^T \mathbf{x}_m \} = -\frac{N}{2\sigma^4}, \quad (S9)$$

$$B_{\theta_0} \triangleq E_p \left\{ \left(\frac{\partial \ln f_X(\mathbf{x}_m | \theta)}{\partial \theta} \right)^2 \Big|_{\theta=\theta_0} \right\} = \frac{N\theta_0^2 + E_p \{ (\mathbf{x}_m^T \mathbf{x}_m)^2 \} - 2N\theta_0 E_p \{ \mathbf{x}_m^T \mathbf{x}_m \}}{4\theta_0^4} = \frac{\text{tr}(\mathbf{\Sigma}^2)}{2\sigma^4}. \quad (S10)$$

Finally, from (5), we get

$$\text{MCRB}(\theta_0) = \text{MCRB}(\sigma^2) = \frac{2\sigma^4}{MN^2} \text{tr}(\mathbf{\Sigma}^2). \quad (S11)$$

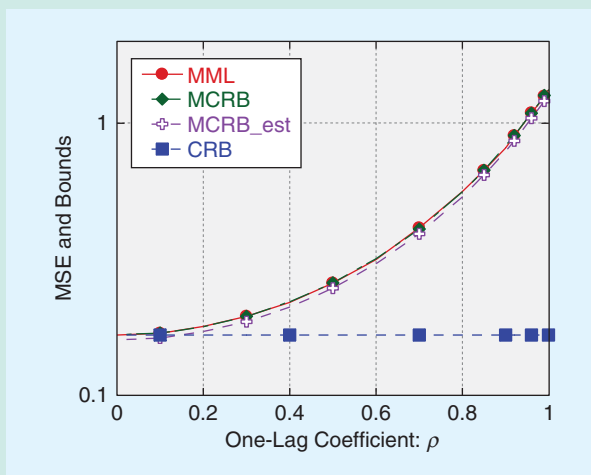


FIGURE S3. The MSE of the MML estimator, the MCRB, the estimated MCRB, and the CRB as functions of ρ . Simulation parameters are set as $N = 8$, $M = 3N$, and $\sigma^2 = 4$.

The CRB for the estimation of the statistical power of the true model can be easily obtained as $\text{CRB}(\sigma^2) = 2\sigma^4/MN$. As expected, the CRB is always greater than the MCRB on σ^2 , and they are equal if, and only if, $\mathbf{\Sigma} = \mathbf{I}_N$, i.e., when the model is correctly specified. We can go on to investigate the properties of the MML estimator. Unlike the example in "Variance Estimation," the MML estimator of the statistical power is consistent since, from (11), it converges to θ_0 that is equal to the true power σ^2 :

$$\hat{\theta}_{\text{MML}}(\mathbf{x}) = (MN)^{-1} \sum_{m=1}^M \mathbf{x}_m^T \mathbf{x}_m \xrightarrow[M \rightarrow \infty]{a.s.} \theta_0 = \sigma^2. \quad (S12)$$

Moreover, the MML estimator is MS-unbiased, since

$$E_p \{ \hat{\theta}_{\text{MML}}(\mathbf{x}) \} = (MN)^{-1} \sum_{m=1}^M E_p \{ \mathbf{x}_m^T \mathbf{x}_m \} = N^{-1} \sigma^2 \text{tr}(\mathbf{\Sigma}) = \sigma^2 = \theta_0. \quad (S13)$$

Then, according to Theorem 1, its MSE is lower bounded by the MCRB in (5). Figure S3 shows the MSE of the MML estimator, the MCRB, the sample estimate of the MCRB, and the CRB as a function of the one-lag coefficient ρ . The MCRB is a tight bound for the MSE of the MML estimator, and the sample MCRB accurately predicts it. Finally, we note that the MCRB is equal to the CRB only when $\rho = 0$, i.e., when $\mathbf{\Sigma} = \mathbf{I}_N$.

Bayesian framework is designed to allow prior knowledge to influence the estimation process in an optimal fashion. Specifically, within a Bayesian framework, estimation of the parameter vector θ is derived from the joint pdf $f_{X,\theta}(\mathbf{x}, \theta)$ instead of solely the conditional (non-Bayesian) pdf $f_{X|\theta}(\mathbf{x}|\theta)$. From basic probability theory, the joint density can be expressed as $f_{X,\theta}(\mathbf{x}, \theta) = f_{\theta|\mathbf{x}}(\theta|\mathbf{x})f_X(\mathbf{x})$, where clearly the posterior density $f_{\theta|\mathbf{x}}(\theta|\mathbf{x})$ summarizes all the information needed to make any inference on θ based on the data $\mathbf{x} = \{\mathbf{x}_m\}_{m=1}^M$. The joint density can likewise be related to the conditional density that models the parameter's influence on data measurements, i.e., $f_{X,\theta}(\mathbf{x}, \theta) = f_{X|\theta}(\mathbf{x}|\theta)f_{\theta}(\theta)$. Prior knowledge about parameter vector θ is reflected in the prior pdf $f_{\theta}(\theta)$. When there is no prior knowledge, all outcomes for the parameter vector can be assumed to be equally likely. Such a noninformative prior pdf often leads to results consistent with standard non-Bayesian approaches, i.e., it yields algorithms and bounds that rely primarily on $f_{X|\theta}(\mathbf{x}|\theta)$. Thus, the Bayesian framework in a sense can be considered as a generalization of the non-Bayesian framework [27], [43], [44].

When the model is perfectly specified, the optimal Bayesian estimator under cost metrics, such as the squared error and the uniform cost, depends primarily on the posterior distribution $f_{\theta|\mathbf{x}}(\theta|\mathbf{x})$. Indeed, the squared error cost is minimized by the conditional mean estimator $\hat{\theta}_{\text{MSE}}(\mathbf{x}) = E_{f_{\theta|\mathbf{x}}} \{\theta|\mathbf{x}\}$, and the uniform cost is minimized by the maximum a posteriori (MAP) estimator $\hat{\theta}_{\text{MAP}}(\mathbf{x}) = \arg\max_{\theta} f_{\theta|\mathbf{x}}(\theta|\mathbf{x})$ [27], [44]. Under a perfect model specification, the asymptotic properties of Bayes estimators and of the posterior distribution have been investigated extensively. Under suitable conditions, as the number of data samples increases, the Bayes estimator tends to become independent of the prior distribution [27, Ch. 4]. Thus, the influence of the prior distribution on a posteriori inferences decreases, and asymptotic behavior similar to the non-Bayesian ML estimator emerges. Indeed, strong consistency, efficiency, and normality properties of Bayes estimators have been established for a large class of prior distributions [41]. This asymptotic behavior has some intuitive appeal, since the prior pdf represents a statistical summary of one's best guess (prior to an actual experiment) of the likelihood the desired parameter will assume any particular value. As actual data measurements become available, however, it makes sense that one will eventually abandon the guidance provided by the prior pdf in light of the valuable information carried by the data measurements obtained from the actual experiment. This phenomenon is well established and has been observed in SP applications. When the prior $f_{\theta}(\theta)$ is incorrect but the model $f_{X|\theta}(\mathbf{x}|\theta)$ is correct, then it is possible that a significantly larger number of data observations (or higher SNR) may be required before the Bayes estimator becomes independent of the influence of the incorrect prior pdf [21, p. 4737].

Misspecification within a Bayesian framework explores the possibility that the assumed joint pdf $f_{X,\theta}(\mathbf{x}, \theta)$ may be incorrect. This, of course, includes the prior pdf $f_{\theta}(\theta)$ as well as the model $f_{X|\theta}(\mathbf{x}|\theta)$. Under model misspecification, the asymptotic properties of the posterior distribution also have been investigated extensively. The following discussion attempts to summarize

some key results on this topic, although no claims are made here that the summary is complete or exhaustive. The goal here is to identify results of potential interest to the SP community in the authors' viewpoint. The first discussion to follow will focus on published results that detail the asymptotic behavior and properties of the Bayesian posterior distribution under model misspecification, i.e., the asymptotic behavior of $f_{\theta|\mathbf{x}}(\theta|\mathbf{x})$ as the amount of data increases. These results can be considered the Bayesian counterparts in the spirit of the contributions of Huber [20] and White [46] that detail ML estimator performance under misspecification, as discussed earlier. Second, a discussion of results on misspecified Bayesian bounds is given. As this remains a relatively new area of research, there appear to be very few published results on this topic; hence, a brief discussion of some of the topic's inherent issues is also provided.

Bayesian estimation under misspecified models

Since Bayes estimators are derived from the posterior density $f_{\theta|\mathbf{x}}(\theta|\mathbf{x})$, considering its asymptotic behavior yields insights into the convergence properties of the associated estimators. Berk [4] was the first to investigate the asymptotic behavior of the posterior distribution under misspecification as the number of data observations becomes arbitrarily large. Specifically, consider a set of i.i.d. data measurements $\mathbf{x} = \{\mathbf{x}_m\}_{m=1}^M$ according to joint pdf $p_X(\mathbf{x}) = \prod_{m=1}^M p_X(\mathbf{x}_m)$. Let the assumed pdf of \mathbf{x} be $f_X(\mathbf{x}|\theta) = \prod_{m=1}^M f_X(\mathbf{x}_m|\theta)$ and the assumed prior pdf be $f_{\theta}(\theta)$. Define the set Θ_A such that

$$\Theta_A \triangleq \left\{ \theta \in \Theta: \underset{\theta \in \Theta}{\operatorname{argmin}} \{ -E_p \{ \ln f_X(\mathbf{x}|\theta) \} \} \right\}. \quad (17)$$

For a large class of unimodal and well-behaved distributions, the set Θ_A consists of a single unique point, i.e., $\Theta_A = \{\theta_0\}$, but the definition clearly allows for the possibility that this set contains more than one point. It is also noteworthy [see also (1)] that the set Θ_A is simply the set of all points/vectors $\theta \in \Theta$ that minimize the KLD $D(p_X \| f_{X|\theta})$ between the true and assumed distributions. Berk noted this relation to the KLD in [4], i.e., prior to the Akaike [1] reference to Huber's work [20]. Berk proved that, if $\Theta_A = \{\theta_0\}$, i.e., it consists of a single unique point θ_0 , then the following convergence in distribution holds:

$$f_{\theta|\mathbf{x}}(\theta|\mathbf{x}) \triangleq f_{\theta|\mathbf{x}}(\theta|\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_M) \xrightarrow[M \rightarrow \infty]{d} \delta(\theta - \theta_0), \quad (18)$$

where $\delta(\mathbf{a}) = \delta(a_1)\delta(a_2) \cdots \delta(a_d)$ and $\delta(a)$ is a Dirac delta function.

From (18), one can presume that θ_0 is the counterpart for the misspecified Bayesian estimation framework of the pseudotrue parameter vector introduced in (1). This conjecture is validated by the fundamental results of Bunke and Milhaud [6] that provide strong consistency arguments for a class of mismatched (or pseudo) Bayesian (MB) estimators. Specifically, let $L(\cdot, \cdot)$ be a nonnegative, real-valued loss function such that $L(\theta, \theta) = 0$. A familiar example of this type of functions is the one leading to the MSE between a given estimate $\hat{\theta}$ and

a given vector θ , i.e., $L_{\text{MSE}}(\hat{\theta}, \theta) = (\hat{\theta} - \theta)^T (\hat{\theta} - \theta)$. Consider now the class of (possibly mismatched) Bayesian estimates defined as

$$\begin{aligned} \hat{\theta}_{\text{MB}}(\mathbf{x}) &\triangleq \underset{\vartheta \in \Theta}{\text{argmin}} E_{f_{\theta|\mathbf{x}}} \{L(\vartheta, \theta)\} \\ &= \underset{\vartheta \in \Theta}{\text{argmin}} \int_{\Theta} L(\vartheta, \theta) f_{\theta|\mathbf{x}}(\theta | \mathbf{x}) d\theta. \end{aligned} \quad (19)$$

Bunke and Milhaud [6] investigated the asymptotic behavior of the class of estimators in (19) and their results can be recast as follows.

Theorem 3

For Theorem 3 [6], under certain regularity conditions (see [6, Assumptions A1–A11]) and provided that $\Theta_A = \{\theta_0\}$, it can be shown that

$$\hat{\theta}_{\text{MB}}(\mathbf{x}) \xrightarrow[M \rightarrow \infty]{a.s.} \theta_0. \quad (20)$$

Moreover,

$$\sqrt{M}(\hat{\theta}_{\text{MB}}(\mathbf{x}) - \theta_0) \xrightarrow[M \rightarrow \infty]{d.} \mathcal{N}(\mathbf{0}, \mathbf{\Lambda}_{\theta_0}), \quad (21)$$

where

$$\mathbf{\Lambda}_{\theta_0} \triangleq \bar{\mathbf{L}}_2^{-1} \bar{\mathbf{L}}_1 \mathbf{A}_{\theta_0}^{-1} \mathbf{B}_{\theta_0} \mathbf{A}_{\theta_0}^{-1} (\bar{\mathbf{L}}_2^{-1} \bar{\mathbf{L}}_1)^T, \quad (22)$$

$$[\bar{\mathbf{L}}_1]_{i,j} = \left. \frac{\partial^2 L(\alpha, \beta)}{\partial \alpha_i \partial \beta_j} \right|_{\substack{\alpha = \theta_0 \\ \beta = \theta_0}}, \quad [\bar{\mathbf{L}}_2]_{i,j} = \left. \frac{\partial^2 L(\alpha, \theta_0)}{\partial \alpha_i \partial \alpha_j} \right|_{\alpha = \theta_0}, \quad (23)$$

and the matrices \mathbf{A}_{θ_0} and \mathbf{B}_{θ_0} have been defined in (2) and (3), respectively. Two comments are in order:

- 1) The similarity between the results given in Theorem 1 for the MML estimator and the ones given in Theorem 2 for the MB estimator is now clear: under model misspecification (and under suitable regularity conditions), both the MML and the MB estimators *a.s.* converge to the point θ_0 that minimizes the KLD between the true and the assumed distributions. Moreover, they are both asymptotically normal-distributed with covariance matrices that are related to the matrices \mathbf{A}_{θ_0} and \mathbf{B}_{θ_0} .
- 2) If, in (19), the squared error loss function $L_{\text{MSE}}(\alpha, \beta)$ is used, then $\bar{\mathbf{L}}_1 = -\bar{\mathbf{L}}_2 = 2\mathbf{I}$, and, consequently, the asymptotic covariance matrices of the MB estimator and the MML estimator are the same, i.e., $\mathbf{\Lambda}_{\theta_0} = \mathbf{C}_{\theta_0} = \mathbf{A}_{\theta_0}^{-1} \mathbf{B}_{\theta_0} \mathbf{A}_{\theta_0}^{-1}$.

While identifying key results from [4] and [6] in this article, reference has been made to several assumptions (e.g., see [6, Assumptions A1–A11]) whose details were omitted here. While important (in particular, the uniqueness of the KLD minimizer is critical in Theorem 3), the inclusion of these details would unnecessarily clutter the discussion. However, the regularity conditions described by [6] characterize a wide spectrum of problems relevant to the SP community.

To conclude, the results discussed in this section are based on a parametric model $f_{\mathbf{x}}(\mathbf{x}|\theta)$ for the data. A similar conver-

gence persists in the nonparametric case. Specifically, Kleijn and van der Vaart [25] address convergence properties of the posterior distribution in the nonparametric case as well as the rate of convergence.

Bayesian bounds under misspecified models

As outlined in the section “A Formal Theory of Statistical Inference Under Misspecified Models,” when the model is correctly specified, a wide family of Bayesian bounds can be derived from the covariance inequality [43]. As is well detailed in [34] and [43], this family includes the Bayesian CRB, the Bayesian Bhattacharyya bound, the Bobrovsky–Zakai bound, and the Weiss–Weinstein bound, among others. Establishing Bayesian bounds under model misspecification appears to have received very limited attention and represents an area of open research. The only results on the topic to the authors’ knowledge are given in [22] and [38]. The approach taken therein differs from the classical approach adopted in [43] with some loss in generality. In fact, the Bayesian bounds obtained in [22] and [38] attempt to build on the non-Bayesian results in [37]. Specifically, it is required that the true conditional pdf $p_{\mathbf{x}|\theta}(\mathbf{x}|\theta)$ and the assumed model $f_{\mathbf{x}|\theta}(\mathbf{x}|\theta)$ share the same parameter space Θ ; thus, any misspecification is exclusively due to the functional form of the assumed distribution. This is essentially the particular case discussed in the non-Bayesian context in the section “An Interesting Case: A Lower Bound on the MSE via the MCRB,” and the bound that we are going to derive has a form similar to the non-Bayesian bound in (9).

Let the conditional mean of the estimator be $E_{p_{\mathbf{x}|\theta}}\{\hat{\theta}(\mathbf{x})\} = \mu(\theta)$, and define the error vector and the bias vector as $\zeta(\mathbf{x}, \theta) \triangleq \hat{\theta}(\mathbf{x}) - \theta$ and $\mathbf{r}(\theta) \triangleq \mu(\theta) - \theta$, respectively. As in (9), the total MSE is given by the sum of the covariance and squared bias. Thus, by use of the covariance inequality [43], a lower bound on MSE under model misspecification is given by

$$\begin{aligned} \text{MSE}_{p_{\mathbf{x},\theta}}(\hat{\theta}(\mathbf{x})) &\triangleq E_{p_{\mathbf{x},\theta}}\{\zeta\zeta^T\} \\ &\geq \frac{1}{M} E_{p_{\mathbf{x},\theta}}\{\zeta\eta^T\} E_{p_{\mathbf{x},\theta}}^{-1}\{\eta\eta^T\} E_{p_{\mathbf{x},\theta}}\{\eta\zeta^T\} \\ &\quad + E_{p_{\theta}}\{\mathbf{r}\mathbf{r}^T\}, \end{aligned} \quad (24)$$

where we dropped the dependences on \mathbf{x} and θ for notation simplicity. The vector function $\eta(\mathbf{x}, \theta)$ represents the score function [43], and a judicious choice of it leads to tight bounds. In [22] and [38], the following score function is considered with the aim of obtaining a bound for the Bayes MAP estimator and ML estimator in mind:

$$\eta(\mathbf{x}, \theta) = \nabla_{\theta} \ln f_{\mathbf{x}|\theta}(\mathbf{x}|\theta) - E_{p_{\mathbf{x}|\theta}}\{\nabla_{\theta} \ln f_{\mathbf{x}|\theta}(\mathbf{x}|\theta)\}. \quad (25)$$

This score function is the same as the one used for the MCRB in [37], and it leads to a version of the misspecified Bayesian CRB (MBCRB). To demonstrate this fact, we define the following two matrices based on the conditional expectation: $E_{p_{\mathbf{x}|\theta}}\{\eta\zeta^T\} = \Xi(\theta)$ and $E_{p_{\mathbf{x}|\theta}}\{\eta\eta^T\} = \mathbf{J}(\theta)$. Closed-form

expressions can be found in [37] for the case where both the true and the assumed conditional distributions are complex Gaussian, for example. The resulting lower bound on the MSE follows from (24) and is given by

$$\text{MSE}_{p_{\mathbf{x},\theta}}(\hat{\boldsymbol{\theta}}(\mathbf{x})) \geq \frac{1}{M} E_{p_\theta} \{ \boldsymbol{\Xi}^T(\boldsymbol{\theta}) \} E_{p_\theta}^{-1} \{ \mathbf{J}(\boldsymbol{\theta}) \} E_{p_\theta} \{ \boldsymbol{\Xi}(\boldsymbol{\theta}) \} + E_{p_\theta} \{ \mathbf{r}\mathbf{r}^T \}. \quad (26)$$

The class of estimators to which the above MCRB applies is that which has a mean and an estimator-score function correlation that, respectively, satisfy the following constraints:

$$E_{p_{\mathbf{x},\theta}} \{ \hat{\boldsymbol{\theta}}(\mathbf{x}) \} = E_{p_\theta} \{ \boldsymbol{\mu}(\boldsymbol{\theta}) \}, \\ E_{p_{\mathbf{x},\theta}} \{ \boldsymbol{\eta}(\mathbf{x}, \boldsymbol{\theta}) [\hat{\boldsymbol{\theta}}(\mathbf{x}) - \boldsymbol{\mu}(\boldsymbol{\theta})]^T \} = E_{p_\theta} \{ \boldsymbol{\Xi}(\boldsymbol{\theta}) \}. \quad (27)$$

These constraints follow from the covariance inequality [37, Sect. III-C] and the choice of score function. This limits the applicability of the bound in contrast to bounds obtained when the model is perfectly specified. Thus, an obvious area of future effort is the development of Bayesian bounds under misspecified models with fewer constraints and broader applicability. To conclude, we note that an example demonstrating the applicability of this Bayesian bound to a DOA estimation for sparse arrays is given in [22].

Examples of applications

In this section, we describe some examples related to the problems of DOA estimation and data covariance/scatter matrix estimation. These problems are relevant in many array processing and adaptive radar applications.

DOA estimation under model misspecification

The estimation of the DOAs of plane-wave signals by means of an array of sensors has been the core research area within the SP array community for years [42]. The fundamental prerequisite for any DOA estimation algorithm is that the positions of the sensors in the array are known exactly, i.e., known geometry. Many authors have investigated the impact of imperfect knowledge of the sensor positions on the DOA estimation performance or of the miscalibration of the array itself (see, e.g., [15] and [42], just to name two). Other authors have proposed hybrid or modified CRBs with the aim to predict the MSE of the DOA estimators in the presence of the position uncertainties [31], [39]. The goal of this section is to show that the misspecified estimation framework presented in this article is a valuable and general tool to deal with modeling errors in the array manifold. The application of the MCRB and the MML estimator to the DOA estimation problem has recently been investigated in [35] for MIMO radar systems and in [37] for uniform linear arrays (ULAs).

Following [37], consider a ULA of N sensors and a single plane-wave signal impinging on the array from a conic angle $\bar{\theta}$. Moreover, suppose that, due to an array miscalibration, the true position vector \mathbf{p}_n of the n th sensor, defined in a three-dimensional Cartesian coordinate frame, is known up to an error term modeled as a zero-mean, Gaussian random vector,

i.e., $\mathbf{e}_n \sim \mathcal{N}(0, \sigma_e^2 \mathbf{I}_3)$. Then, the received data can be expressed as $x_n = s[\mathbf{d}(\bar{\theta})]_n + [\mathbf{c}]_n$, where $[\mathbf{d}(\bar{\theta})]_n = \exp(j\mathbf{k}_{\bar{\theta}}^T(\mathbf{p}_n + \mathbf{e}_n))$ is the n th element of the true (perturbed) steering vector and $\mathbf{k}_{\bar{\theta}} = (2\pi/\lambda)\mathbf{u}(\bar{\theta})$, where $\mathbf{u}(\bar{\theta})$ is a unit vector pointing at the direction defined by $\bar{\theta}$ and λ is the wavelength of the transmitted signal. Moreover, s is an unknown deterministic complex scalar that accounts for the transmitted power, the source scattering characteristics, and the two-way path loss, while $\mathbf{c} = \mathbf{n} + \mathbf{j}$ is the disturbance noise term composed of white Gaussian noise \mathbf{n} and of interference signal (or jammer) \mathbf{j} . Given particular realizations of the position errors \mathbf{e}_n , the clutter vector is usually modeled as a zero-mean complex Gaussian random vector $\mathbf{c} \sim \mathcal{N}(0, \mathbf{I}_N + \sigma_j^2 \mathbf{d}(\theta_j) \mathbf{d}(\theta_j)^H)$, where σ_j^2 and θ_j represent the power and the DOA of the jamming signal. The DOA estimation problem is clearly the estimation of $\bar{\theta}$, given the complex data vector \mathbf{x} . Since, in practice, it is impossible to be aware of the particular realizations of the position error vectors \mathbf{e}_n , the user may decide to derive a DOA estimator starting from the nominal steering vector $\mathbf{v}(\bar{\theta})$, whose components are $[\mathbf{v}(\bar{\theta})]_n = \exp(j\mathbf{k}_{\bar{\theta}}^T \mathbf{p}_n)$, i.e., the user neglects the sensor position errors. The true (unknown) data model is given by the pdf $p_X(\mathbf{x}) = \mathcal{N}(s\mathbf{d}(\bar{\theta}), \mathbf{I}_N + \sigma_j^2 \mathbf{d}(\theta_j) \mathbf{d}(\theta_j)^H) \in \mathcal{P}$, while the assumed parametric model is

$$\mathcal{F} = \{ f_X | f_X(\mathbf{x} | s, \theta) \\ = \mathcal{N}(s\mathbf{v}(\theta), \mathbf{I}_N + \sigma_j^2 \mathbf{v}(\theta_j) \mathbf{v}(\theta_j)^H), \forall s \in \mathbb{C}, \theta \in [0, 2\pi) \}. \quad (28)$$

The true pdf $p_X(\mathbf{x})$ does not belong to \mathcal{F} ; in other words, the assumed parametric pdf $f_X(\mathbf{x} | s, \theta)$ differs from $p_X(\mathbf{x})$ for every value of $\theta \in [0, 2\pi)$. This is because, even if both the true and the assumed pdfs are complex Gaussian, by neglecting the position errors in the assumed steering vector, we are choosing the wrong parameterization for the mean value and the covariance matrix of the assumed Gaussian model. Therefore, how large is the performance loss due to this model mismatch? The MCRB presented in the section “The Misspecified CRB” answers this question. We omit the details of the calculation of the MCRB and the derivation of the joint MML estimator of the DOA and of the scalar s , and we refer readers to [37]. However, to provide some insights about this mismatched estimation problem, Figure 1 illustrates the matched CRB in the estimation of $\bar{\theta}$, i.e., the CRB on the DOA estimation evaluated by considering the true data pdf $p_X(\mathbf{x})$, the MCRB, and the MSE of the MML estimator obtained from the assumed and misspecified pdf $f_X(\mathbf{x} | s, \theta)$. Figure 1 plots the square roots of the bounds and of the MSE (RMSE) in units of beamwidths as a function of element-level SNR. The MCRB accurately predicts the performance of the MML estimator. If the system goal is a ten-to-one beamsplit ratio, i.e., -10 dB RMSE in beamwidths, then this could be accomplished with an SNR of 9.28 dB when the model is perfectly known, but not precisely knowing the true sensor positions requires an additional ~ 10 dB of SNR to achieve the same goal (MCRB $\cong -10$ dB for SNR $\cong 19.4$ dB). However, if the system receives an SNR $\cong 9.3$ dB, then the minimum achievable beamsplit ratio in the presence of array errors is three to one, i.e., the MCRB $\cong -5$ dB RMSE in beamwidths. This

information can be quite valuable in determining where to focus efforts to improve system performance.

Scatter matrix estimation under model misspecification

Another widely encountered inference problem is the estimation of the correlation structure, i.e., the scatter or covariance matrix, of a data set. The estimation of the covariance/scatter matrix is a central component of a wide variety of SP applications [30]: adaptive detection and DOA estimation in array processing, principal component analysis, signal separation, interference cancellation, and the portfolio optimization in finance, just to name a few. Even if the data may come from disparate applications, they usually share a non-Gaussian, heavy-tailed statistical nature, as discussed in [49]. Estimating the covariance matrix of a set of non-Gaussian data, however, is not a trivial task. In fact, non-Gaussian distribution characterization typically requires additional parameters that must be jointly estimated along with the scatter matrix. Think, for example, of the (complex) t -distribution that has been widely adopted as a suitable and flexible model able to characterize the non-Gaussian, heavy-tailed data behavior [26], [30], [40]. A complex, zero-mean, random vector $\mathbf{x}_m \in \mathbb{C}^N$ is said to be t -distributed if its pdf can be expressed as

$$p_X(\mathbf{x}_m | \bar{\Sigma}, \lambda, \eta) \triangleq \frac{1}{\pi^N |\bar{\Sigma}|} \frac{\Gamma(N + \lambda)}{\Gamma(\lambda)} \left(\frac{\lambda}{\eta}\right)^\lambda \left(\frac{\lambda}{\eta} + \mathbf{x}_m^H \bar{\Sigma}^{-1} \mathbf{x}_m\right)^{-(N+\lambda)}, \text{tr}(\bar{\Sigma}) = N, \quad (29)$$

where $\Gamma(\cdot)$ indicates the gamma function while λ and η are the so-called shape and scale parameters, and $\bar{\Sigma}$ is the scatter matrix. This multidimensional pdf is obtained by assuming that vector \mathbf{x}_m follows the compound-Gaussian model with Gaussian speckle and inverse-Gamma distributed texture [40]. For proper identifiability, a constraint on $\bar{\Sigma}$, e.g., $\text{tr}(\bar{\Sigma}) = N$, needs to be imposed. The complex t -distribution has tails heavier than the Gaussian for every $\lambda \in (0, \infty)$, and it becomes the complex Gaussian distribution for $\lambda \rightarrow \infty$. As can be clearly seen from (29), to perform some inference on a t -distributed data set, we must jointly estimate the shape and scale parameters along with the scatter matrix. Unfortunately, as pointed out in [26], a joint ML estimator of these three quantities presents convergence and even existence issues. Moreover, as discussed in the section “A Covariance Inequality in the Presence of Misspecified Models,” the t -distribution may be only an approximation of the true heavy-tailed data model. To overcome these problems, the SP practitioner has fundamentally two choices: 1) to apply some robust covariance matrix estimator (see [30] and [49] for further details) or 2) to assume a simpler, but generally misspecified, model for characterizing the data, gaining the possibility to derive a closed-form estimator at the cost of a loss in the estimation performance [10], [12]. If option 2) is adopted, the most reasonable choice for the simplified data model is the complex Gaussian distribution:

$$f_X(\mathbf{x}_m | \boldsymbol{\theta}) \triangleq f_X(\mathbf{x}_m | \Sigma, \sigma^2) = \frac{1}{(\pi\sigma^2)^N |\Sigma|} \exp\left(-\frac{\mathbf{x}_m^H \Sigma^{-1} \mathbf{x}_m}{\sigma^2}\right), \text{tr}(\Sigma) = N. \quad (30)$$

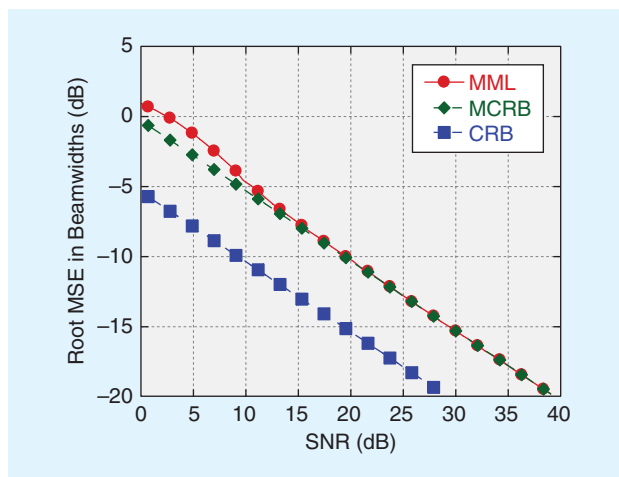


FIGURE 1. The MSE of the MML estimator, the MCRB, for the DOA estimation problem. Simulation parameters are set as $M = 18$ element ULA, the array position errors of $\sigma_e = 0.01\lambda$ of standard deviation, $\theta_1 = 90^\circ, \theta_j = 87^\circ$, and $\sigma_j^2 = 10^3$ (see [37]).

In fact, the joint (constrained) MML estimator of the scatter matrix and of the data power can be derived as

$$\hat{\Sigma}_{\text{CMML}} = \frac{N}{\sum_{m=1}^M \mathbf{x}_m^H \mathbf{x}_m} \sum_{m=1}^M \mathbf{x}_m \mathbf{x}_m^H, \quad \hat{\sigma}_{\text{CMML}}^2 = \frac{1}{NM} \sum_{m=1}^M \mathbf{x}_m^H \hat{\Sigma}_{\text{CMML}}^{-1} \mathbf{x}_m. \quad (31)$$

Two comments are in order:

- 1) It can be shown that $\hat{\Sigma}_{\text{CMML}}$ converges to the true scatter matrix, i.e., $\hat{\Sigma}_{\text{CMML}} \xrightarrow[M \rightarrow \infty]{a.s.} \bar{\Sigma}$; thus, it can be successfully applied to estimate it [10], [12].
- 2) It is computationally inexpensive and easy to implement, which makes the use of $\hat{\Sigma}_{\text{CMML}}$ feasible in real-time applications, e.g., in adaptive radar detection.

Along with knowledge of the MML estimator convergence point, the performance loss that has resulted from model mismatch should also be assessed. To this purpose, since the Gaussian model is nested in heavy-tailed t -distributed model (see the section “An Interesting Case: A Lower Bound on the MSE via the MCRB”), we can evaluate the MCRB for the problem at hand and compare it with the CRB. As an example, in Figure 2, we compare the curves relative to the constrained CRB (CCRB) for the estimation of the scatter matrix under matched conditions (i.e., when the true t -distribution is assumed), the constrained MCRB (CMCRB) [11] (i.e., when the misspecified Gaussian model is assumed), and the MSE of the constrained MML estimator of (31) (details of the calculations can be found in [12]). The distance between the CCRB and the CMCRB curves provides a measure of the performance loss due to model mismatch. As expected, the loss increases when the shape parameter λ reaches zero, i.e., when the data have an extremely heavy-tailed behavior. However, when $\lambda \rightarrow \infty$, i.e., when the t -distribution tends to the Gaussian one, the CCRB and the CMCRB tend to coincide. We note that the constrained MML estimator of the scatter matrix

is an efficient estimator w.r.t. the CMCRB, as predicted by the theory in the section “The Mismatched ML Estimator.”

Concluding remarks

The objective of this article is to provide an accessible and, at the same time, comprehensive treatment of the fundamental concepts about CRBs and efficient estimators in the presence of model misspecification. Every SP practitioner is aware of the fact that, in almost all practical applications, a certain amount of mismatch between the true and the assumed statistical data models is inevitable. Despite its ubiquity, the assessment of performance bounds under model misspecification appears to have received limited attention from the SP community, while it has been deeply investigated by the statistical community. The first aim of this tutorial is to propose to a wide SP audience a comprehensive review of the main contributions to the mismatched estimation theory, both for the deterministic and Bayesian frameworks, with a particular focus on the derivation of CRB under model mismatching. Specifically, we have described how the classical tools of the estimation theory can be generalized to address a mismatched scenario. First, the MCRB has been introduced and the behavior of the MML estimator investigated. Second, results related to the deterministic estimation framework have been extended to the Bayesian one. The existence and the asymptotic properties of a MB estimator have been discussed. Moreover, some general ideas about the possibility to derive MCBs have been provided. In the last part of the article, we showed how to apply the theoretical findings to two well-known relevant problems: the DOA estimation in array processing and the estimation of the disturbance covariance matrix for adaptive radar detection.

Of course, much work remains to be done. A question that naturally arises is whether it is possible to derive a more general class of misspecified bounds. The first step toward this direction has been outlined by Richmond and Horowitz in

[37], where a generalization of the theory to the Bhattacharyya bound, to the BB, and to the Bobrovsky–Mayer–Wolf–Zakai bound has been proposed. Next, as discussed in the “Generalization to the Bayesian Setting” section, a future area of research is the derivation of general Bayesian lower bounds that could be obtained by relaxing or, hopefully, removing the constraints given in (27). Finally, a systematic and deep investigation of a general decision theory under model misspecification is required since it could lead great advantages in a huge number of SP applications.

Acknowledgments

The work of Stefano Fortunati has been partially supported by the Air Force Office of Scientific Research under award number FA9550-17-1-0065.

Authors

Stefano Fortunati (stefano.fortunati@iet.unipi.it) received his Ph.D. degree in telecommunication engineering from the University of Pisa, Italy, in 2012. He then joined the Department of Ingegneria dell’Informazione at the University of Pisa, where he is currently working as a postdoctoral researcher. His professional expertise encompasses different areas of statistical signal processing: estimation and detection theory, statistical methods for data analysis, non-Gaussian signal detection and estimation, robust signal estimation and detection, performance bounds, and compressed sensing theory with applications in radar and sonar systems. He is a Member of the IEEE.

Fulvio Gini (f.gini@iet.unipi.it) received his Dr.-Ing. and Ph.D. degrees in electronic engineering from the University of Pisa, Italy, in 1990 and 1995, respectively. Currently, he is a full professor in the Department of Information Engineering at the University of Pisa. He has received several awards, including the 2001 and 2012 IEEE Aerospace and Electronic Systems (AES) Society’s Barry Carlton Award for Best Paper published in *IEEE Transactions on Aerospace and Electronic Systems*, the 2003 IEEE Achievement Award, and the 2003 IEEE AES Society Nathanson Award to the Young Engineer of the Year. He has authored or coauthored 11 book chapters, approximately 125 journal papers, and 160 conference papers. He is a Fellow of the IEEE.

Maria S. Greco (m.greco@iet.unipi.it) is a full professor in the Department of Information Engineering at the University of Pisa, Italy. She is a corecipient of the 2001 and 2012 IEEE Aerospace and Electronic Systems (AES) Society’s Barry Carlton Awards for Best Paper and recipient of the IEEE AES Society 2008 Fred Nathanson Young Engineer of the Year Award. Her research interests include statistical signal processing, estimation, and detection theory. She has coauthored many book chapters and more than 190 journal and conference papers. She is a Fellow of the IEEE.

Christ D. Richmond (christ.richmond@asu.edu) received his Ph.D. degree in electrical engineering from the Massachusetts Institute of Technology (MIT). He is currently an associate professor at Arizona State University, Tempe, and a former senior staff member at the MIT Lincoln Laboratory. He

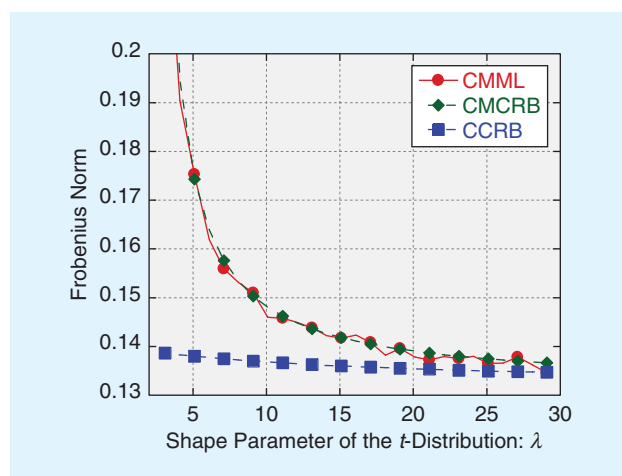


FIGURE 2. The Frobenius norms of the MSE matrix of the CMML estimator, the CMCRB, and the CCRB for the scatter matrix estimation problem. Simulation parameters are set as $N = 16$, $M = 10N$, and the scale parameter of the true t -distribution is $\eta = 1$.

is the recipient of the 1994 Alan Berman Research Publication Award (Naval Research Lab) and the 1999 IEEE Young Author Best Paper Award in the area of sensor array and multichannel signal processing. His research interests include statistical signal processing, detection, estimation, and information theory. He is a Senior Member of the IEEE.

References

- [1] H. Akaike, "Information theory and an extension of the likelihood principle," in *Proc. 2nd Int. Symp. of Information Theory*, 1972, pp. 267–281.
- [2] A. N. D'Andrea, U. Mengali, and R. Reggiannini, "The modified Cramér–Rao bound and its application to synchronization problems," *IEEE Trans. Commun.*, vol. 42, no. 234, pp. 1391–1399, 1994.
- [3] E. W. Barankin, "Locally best unbiased estimates," *Ann. Math. Stat.*, vol. 20, no. 4, pp. 477–501, 1949.
- [4] R. H. Berk, "Limiting behavior of posterior distributions when the model is incorrect," *Ann. Math. Stat.*, vol. 37, no. 3, pp. 51–58, 1966.
- [5] A. Bhattacharyya, "On some analogues of the amount of information and their use in statistical estimation," *Sankhya Indian J. Statist.*, vol. 8, pp. 1–14, 1946.
- [6] O. Bunke and X. Milhaud, "Asymptotic behavior of Bayes estimates under possibly incorrect models," *Ann. Stat.*, vol. 26, no. 2, pp. 617–644, 1998.
- [7] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, 2nd ed. Hoboken, NJ: Wiley, 2006.
- [8] H. Cramér, *Mathematical Methods of Statistics*. Princeton, NJ: Princeton Univ. Press, 1946.
- [9] R. A. Fisher, "Theory of statistical estimation," *Math. Proc. Cambridge*, vol. 22, no. 5, pp. 700–725, 1925.
- [10] S. Fortunati, F. Gini, and M. S. Greco, "The misspecified Cramér–Rao bound and its application to the scatter matrix estimation in complex elliptically symmetric distributions," *IEEE Trans. Signal Process.*, vol. 64, no. 9, pp. 2387–2399, 2016.
- [11] S. Fortunati, F. Gini, and M. S. Greco, "The constrained misspecified Cramér–Rao bound," *IEEE Signal Process. Lett.*, vol. 23, no. 5, pp. 718–721, May 2016.
- [12] S. Fortunati, F. Gini, and M. S. Greco, "Matched, mismatched and robust scatter matrix estimation and hypothesis testing in complex t -distributed data," *EURASIP J. Adv. Signal Process.*, vol. 2016, p. 123, Dec. 2016.
- [13] S. Fortunati, "Misspecified Cramér–Rao bounds for complex unconstrained and constrained parameters," in *Proc. Eur. Signal Process. Conf. (EUSIPCO) 2017*, Kos, Greece, 28 Aug.–2 Sept. 2017.
- [14] C. Fritsche, U. Orguner, E. Ozkan, F. Gustafsson, "On the Cramér–Rao lower bound under model mismatch," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, 19–24 Apr. 2015, pp. 3986–3990.
- [15] B. Friedlander, "Sensitivity analysis of the maximum likelihood direction-finding algorithm," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 26, no. 6, pp. 953–968, Nov. 1990.
- [16] F. Gini and R. Reggiannini, "On the use of Cramér–Rao-like bounds in the presence of random nuisance parameters," *IEEE Trans. Commun.*, vol. 48, no. 12, pp. 2120–2126, Dec. 2000.
- [17] F. Gini, R. Reggiannini, and U. Mengali, "The modified Cramér–Rao bound in vector parameter estimation," *IEEE Trans. Commun.*, vol. 46, no. 1, pp. 52–60, Jan. 1998.
- [18] M. S. Greco, S. Fortunati, and F. Gini, "Maximum likelihood covariance matrix estimation for complex elliptically symmetric distributions under mismatched conditions," *Signal Process.*, vol. 104, pp. 381–386, Nov. 2014.
- [19] A. Gusi-Amigó, P. Closas, A. Mallat and L. Vandendorpe, "Ziv–Zakai lower bound for UWB based TOA estimation with unknown interference," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, Florence, Italy, 2014, pp. 6504–6508.
- [20] P. J. Huber, "The behavior of maximum likelihood estimates under nonstandard conditions," in *Proc. Fifth Berkeley Symp. Mathematical Statistics and Probability*, 1967, pp. 221–233.
- [21] J. M. Kantor, C. D. Richmond, D. W. Bliss, and B. Correll, Jr., "Mean-squared-error prediction for Bayesian direction-of-arrival estimation," *IEEE Trans. Signal Process.*, vol. 61, no. 19, pp. 4729–4739, 2013.
- [22] J. M. Kantor, C. D. Richmond, B. Correll, D. W. Bliss, "Prior mismatch in Bayesian direction of arrival for sparse arrays," in *Proc. IEEE Radar Conf.*, Philadelphia, PA, May 2015, pp. 811–816.
- [23] S. Kay, *Fundamentals of Statistical Signal Processing: Estimation Theory*. Englewood Cliffs, NJ: Prentice-Hall, 1998.
- [24] N. Kbayer, J. Galy, E. Chaumette, F. Vincent, A. Renaux, and P. Larzabal, "On lower bounds for non-standard deterministic estimation," *IEEE Trans. Signal Process.*, vol. 65, no. 6, pp. 1538–1553, 2017.
- [25] J. K. Kleijn and A. W. van der Vaart, "Misspecification in infinite-dimensional Bayesian statistics," *Ann. Stat.*, vol. 34, no. 2, pp. 837–877, 2006.
- [26] K. L. Lange, R. J. A. Little, and J. M. G. Taylor, "Robust statistical modeling using the t distribution," *J. Amer. Stat. Assoc.*, vol. 84, no. 408, pp. 881–896, Dec. 1989.
- [27] E. L. Lehmann and G. Casella, Eds., *Theory of Point Estimation*, 2nd ed. New York: Springer, 1998.
- [28] Y. Noam and J. Tabrikian, "Marginal likelihood for estimation and detection theory," *IEEE Trans. Signal Process.*, vol. 55, no. 8, pp. 3963–3974, Aug. 2007.
- [29] Y. Noam and H. Messer, "Notes on the tightness of the hybrid Cramér–Rao lower bound," *IEEE Trans. Signal Process.*, vol. 57, no. 6, pp. 2074–2084, June 2009.
- [30] E. Ollila, D. E. Tyler, V. Koivunen, and V. H. Poor, "Complex elliptically symmetric distributions: Survey, new results and applications," *IEEE Trans. Signal Process.*, vol. 60, no. 11, pp. 5597–5625, Nov. 2012.
- [31] M. Pardini, F. Lombardini, and F. Gini, "The hybrid Cramér–Rao bound on broadside DOA estimation of extended sources in presence of array errors," *IEEE Trans. Signal Process.*, vol. 56, no. 4, pp. 1726–1730, Apr. 2008.
- [32] P. A. Parker and C. D. Richmond, "Methods and bounds for waveform parameter estimation with a misspecified model," in *Proc. Conf. Signals, Systems, and Computers (Asilomar)*, Pacific Grove, CA, Nov. 2015, pp. 1702–1706.
- [33] C. R. Rao, "Information and the accuracy attainable in the estimation of statistical parameters," *Bull. Calcutta Math. Soc.*, vol. 37, pp. 81–89, 1945.
- [34] A. Renaux, P. Forster, P. Larzabal, C. D. Richmond, and A. Nehorai, "A fresh look at the Bayesian bounds of the Weiss–Weinstein family," *IEEE Trans. Signal Process.*, vol. 56, no. 11, pp. 5334–5352, Nov. 2008.
- [35] C. Ren, M. N. El Korso, J. Galy, E. Chaumette, P. Larzabal, and A. Renaux, "Performances bounds under misspecification model for MIMO radar application," in *Proc. Eur. Signal Process. Conf. (EUSIPCO)*, Nice, France, 2015, pp. 514–518.
- [36] C. D. Richmond and L. L. Horowitz, "Parameter bounds under misspecified models," in *Proc. Conf. Signals, Systems and Computers (Asilomar)*, 3–6 Nov. 2013, pp. 176–180.
- [37] C. D. Richmond and L. L. Horowitz, "Parameter bounds on estimation accuracy under model misspecification," *IEEE Trans. Signal Process.*, vol. 63, no. 9, pp. 2263–2278, 2015.
- [38] C. D. Richmond, P. Basu, "Bayesian framework and radar: On misspecified bounds and radar-communication cooperation," in *Proc. IEEE Statistical Signal Processing Workshop*, Palma de Mallorca, Spain, 26–29 June 2016, pp. 1–4.
- [39] Y. Rockah and P. Schultheiss, "Array shape calibration using sources in unknown locations-part I: Far-field sources," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 35, no. 3, pp. 286–299, Mar. 1987.
- [40] K. J. Sangston, F. Gini, and M. Greco, "Coherent radar detection in heavy-tailed compound-Gaussian clutter," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 48, no. 1, pp. 64–77, Jan. 2012.
- [41] H. Strasser, "Consistency of maximum likelihood and Bayes estimates," *Ann. Stat.*, vol. 9, no. 5, pp. 1107–1113, 1981.
- [42] H. L. Van Trees, *Optimum Array Processing: Part IV of Detection, Estimation, and Modulation Theory*. Hoboken, NJ: Wiley, 2002.
- [43] H. L. Van Trees and K. L. Bell, *Bayesian Bounds for Parameter Estimation and Nonlinear Filtering/Tracking*. Hoboken, NJ: Wiley, 2007.
- [44] H. L. Van Trees, K. L. Bell, and Z. Tian, *Detection, Estimation and Modulation Theory*, vol. 1, 2nd ed. Hoboken, NJ: Wiley, 2013.
- [45] Q. H. Vuong. (1986, Oct. 1). Cramér–Rao bounds for misspecified models, division of the humanities and social sciences. Caltech. [Online]. Available: <https://www.hss.caltech.edu/content/cramer-rao-bounds-misspecified-models>
- [46] H. White, "Maximum likelihood estimation of misspecified models," *Econometrica*, vol. 50, pp. 1–25, Jan. 1982.
- [47] H. White, *Estimation, Inference, and Specification Analysis* (Econometric Society Monograph, no. 22). Cambridge, U.K.: Cambridge Univ. Press, 1996.
- [48] W. Xu, A. B. Gaggeroer, and K. L. Bell, "A bound on mean-square estimation error with background parameter mismatch," *IEEE Trans. Inf. Theory*, vol. 50, no. 4, pp. 621–632, Apr. 2004.
- [49] A. M. Zoubir, V. Koivunen, Y. Chakhchoukh, and M. Muma, "Robust estimation in signal processing: A tutorial-style treatment of fundamental concepts," *IEEE Signal Process. Mag.*, vol. 29, no. 4, pp. 61–80, July 2012.
- [50] A. Mennad, S. Fortunati, M. N. El Korso, A. Younsi, A. M. Zoubir, and A. Renaux, "Stein–Bangs-type formulas and the related misspecified Cramér–Rao bounds for complex elliptically symmetric distributions," *Signal Processing*, vol. 142, pp. 320–329, Jan. 2018.

LECTURE NOTES

Soo-Chang Pei and Kuo-Wei Chang

The Mystery Curve: A Signal Processing Point of View

In the first chapter of a recently published book on artful mathematics [1], a linear combination of harmonic signals called *mystery curves* were introduced. Although their Fourier-based analysis brings interesting results, this lecture note provides a different and important perspective, especially useful for our signal processing community. Based on polar coordinate systems and low-pass filtering approaches, the patterns of the curves can be designed by locally curve tracing instead of trial and error, including not only two-dimensional (2-D) but also three-dimensional (3-D) modeling. Concrete examples and online MATLAB codes are provided so that applications from art and logo design to amplitude modulation (AM)-frequency modulation (FM) signal analysis are realizable.

Relevance

In [1], the first example given is a summation of three rotating circles $\mu(t)$

$$\mu(t) = \left(\cos(t) + \frac{\cos(6t)}{2} + \frac{\sin(14t)}{3}, \sin(t) + \frac{\sin(6t)}{2} + \frac{\cos(14t)}{3} \right) \quad (1)$$

which is illustrated in Figure 1. This fivefold symmetric curve seems incomprehensible by the frequency components one, six, and 14.

Digital Object Identifier 10.1109/MSP.2017.2740457
Date of publication: 13 November 2017

However, it can be shown that the symmetry comes from the difference of frequency and can be easily proved by Fourier series analysis. In this example, if we put μ on the complex plane

$$\begin{aligned} c(t) &= x(t) + iy(t) \\ &= \left(\cos(t) + \frac{\cos(6t)}{2} + \frac{\sin(14t)}{3} \right) \\ &\quad + i \left(\sin(t) + \frac{\sin(6t)}{2} + \frac{\cos(14t)}{3} \right) \\ &= e^{it} + \frac{e^{i6t}}{2} + \frac{ie^{-14t}}{3}. \end{aligned} \quad (2)$$

As we can see, the differences of frequency (1, 6, -14) are

$$\begin{aligned} 1 - 6 &= -5 \\ 6 - (-14) &= 20 \\ 1 - (-14) &= 15 \end{aligned}$$

and the greatest common divisor (gcd) of all the aforementioned frequency differences is five, which is also the number of folds. An equivalent statement in [1] is that (1, 6, -14) are all congruent to one modulo five. In addition, if the frequencies are chosen randomly, as the code shown in *mysterycurve_rand_nofold.m* [3], we can observe that the curves in Figure 2 are not always N -fold symmetry. More examples can be found in [4]. As a result, the relationship between the frequencies and N -fold symmetry can be deduced.

Although the author of [1] solved the mystery and one could generate random mystery curves by setting parameters, it is still hard to synthesize a desired

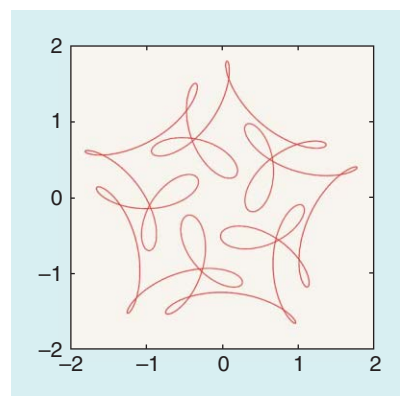


FIGURE 1. The mystery curve. An animated version can be found in [2].

N -fold pattern without trial and error. In this article, we propose another way to generate the mystery curve. Instead of finding a_j and w_j from

$$c(t) = \sum_j a_j e^{iw_j t}, \quad (3)$$

we use the polar coordinate system

$$c(t) = r(t)e^{i\theta(t)}. \quad (4)$$

After deriving the properties and constraints of $r(t)$ and $\theta(t)$, we can create a pattern by locally tracing a curve on the polar coordinate system. This is quite useful for an art or commercial application such as logo design.

The beautiful N -fold symmetry curve can be extended to 3-D cases by similar definition in two dimensions. Today, 3-D printing is popular

and affordable, and modern sculptures or artwork created by the 3-D mystery curve might decorate our living rooms in the near future.

Prerequisites

The prerequisites for this lecture note consist of basic modular arithmetic, Fourier series analysis [5], and geometric representation of complex numbers. Some notations in this article are given as follows. A curve on the 2-D plane is defined as $(x(t), y(t))$ and in the 3-D space it is defined as $(x(t), y(t), z(t))$. Since the symbol z is already used for the z -axis, the complex representation of $(x(t), y(t))$ on the xy plane is denoted by $c(t) = x(t) + iy(t)$, where $i = \sqrt{-1}$. A mystery curve is a periodic curve with N -fold symmetry. For convenience, the period is set to 2π .

Problem statement

In this lecture note we will 1) investigate how the mysterious curve creates the N -fold rotation symmetry; 2) introduce the polar form description of the mysterious curve (It is much easier than the Cartesian coordinate to specify and trace the curve of the shape of the mysterious curve); 3) extend the 3-D mysterious curve; and 4) provide MATLAB code implementations for 2-D/3-D mysterious curves generation [3].

Solution

Review of mystery curve and Fourier series

In this section, we will review how to generate the mystery curve by Fourier series analysis. Recall that an N -fold symmetric curve $c(t)$ must satisfy the following equation:

$$e^{\frac{i2\pi k}{N}} c(t) = c\left(t + \frac{2\pi}{N}\right), \quad (5)$$

where the integers k and N are coprime. The physical meaning of (5) is that the time delay (right-hand side) is equal to some rotation of the past. Note that if k and N are not coprime, $\text{gcd}(k, N) = d > 1$, then a N/d fold symmetry curve will be generated instead. To derive the constraints of $c(t)$, we expand it by Fourier series

$$c(t) = \sum_{n=-\infty}^{\infty} C_n e^{int}. \quad (6)$$

Rewrite (5) as

$$\sum_{n=-\infty}^{\infty} C_n e^{int} \left(1 - e^{\frac{i2\pi(n-k)}{N}}\right) = 0. \quad (7)$$

Equation (7) implies that C_n can be non-zero only when n is congruent to k modulo N . In other words, let $n = Nn' + k$ and $c(t)$ is in the form of

$$\begin{aligned} c(t) &= \sum_{n'=-\infty}^{\infty} C_{n'} e^{i(Nn'+k)t} \\ &= e^{ikt} \sum_{n'=-\infty}^{\infty} C_{n'} e^{iNn't}. \end{aligned} \quad (8)$$

For instance, (2) can be rewritten as

$$c(t) = e^{it} \left(1 + \frac{1}{2}e^{i5t} + \frac{i}{3}e^{-i15t}\right)$$

with $N = 5, k = 1, C_0 = 1, C_1 = 1/2$, and $C_{-3} = i/3$.

We provide some random N -fold mystery curves in Figure 3. The complex representation of all the curves is in the form of

$$c(t) = e^{ikt} (C_1 e^{in_1 Nt} + C_2 e^{in_2 Nt} + C_3 e^{in_3 Nt}), \quad (9)$$

where N is chosen from three, five, or seven. More results can be found in [1], and the code is given in `mysterycurve_rand.m` [3]. Codes in other programming languages such as Python and Mathematica can be found on the Internet [6], [7].

This method, however, needs trial and error to find good mystery curves. If we already have some drafts, scratches, or contours in our mind, this method will not help much. Note that every change on C_n will affect the curve globally. To design a good pattern easily, one might want to draw a curve locally and then rotate it N times. This can be done by considering polar coordinate system instead of Cartesian coordinate system, as we will discuss in the next section.

Mystery curve in polar form

Recall that any complex number $c = x + iy$ can be expressed in polar form

$$c = x + iy = re^{i\theta}.$$

To derive the properties of mystery curve, we rewrite $c(t)$ as

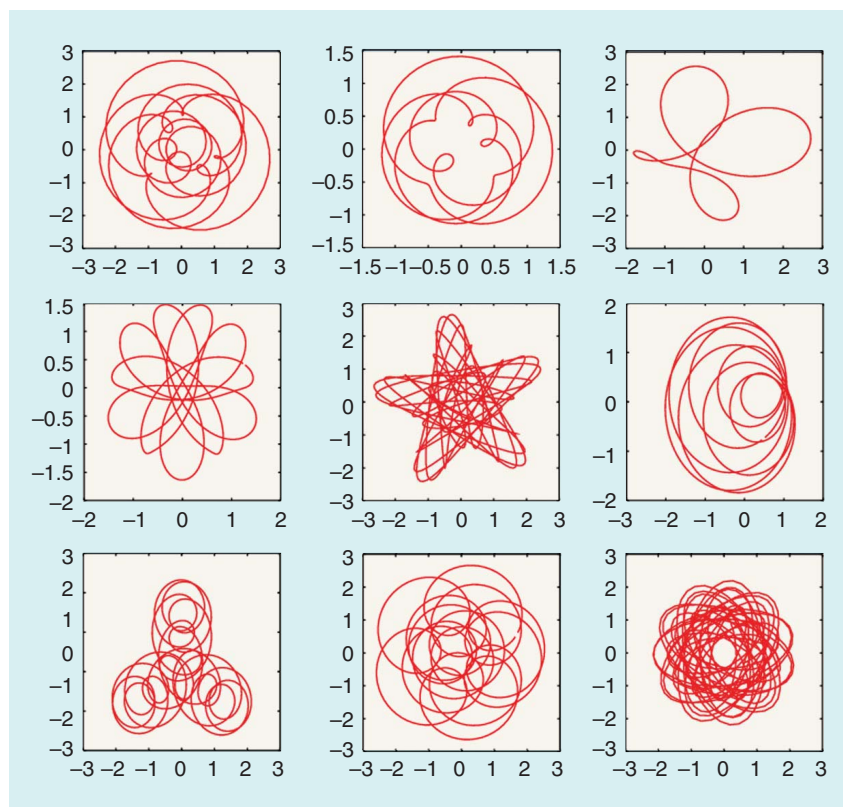


FIGURE 2. The random curves generated by `mysterycurve_rand_nofold.m` [3].

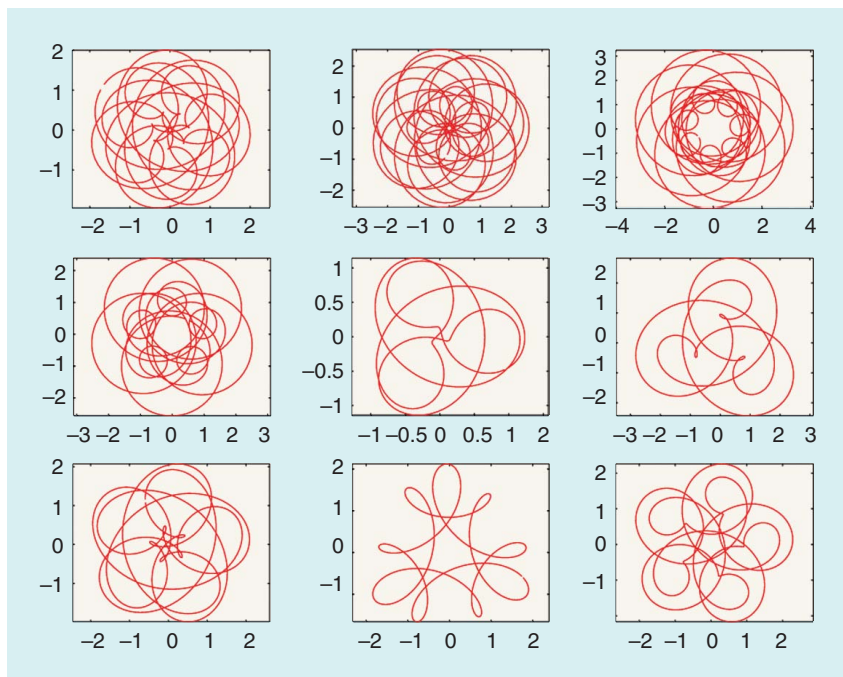


FIGURE 3. Random N -fold mystery curves with $N = 3, 5,$ or 7 , in the form of (9). The parameters $C_1, C_2, C_3, n_1, n_2, n_3,$ and k are chosen randomly.

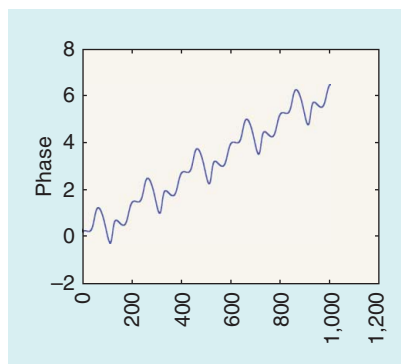


FIGURE 4. The phase $\theta(t)$ of (2). Clearly, it is a linear function plus a periodic function.

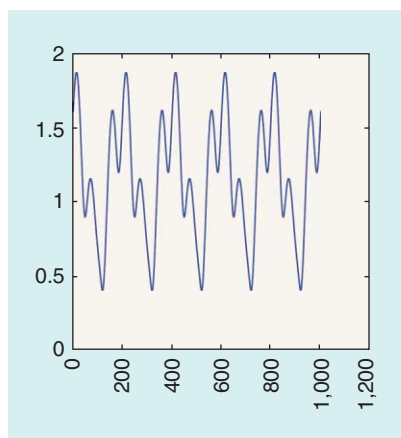


FIGURE 5. The amplitude $r(t)$ of (2). Clearly, it is a periodic function.

$$c(t) = x(t) + iy(t) = r(t)e^{i\theta(t)}.$$

Then (5) becomes

$$c\left(t + \frac{2\pi}{N}\right) = e^{i\frac{2\pi k}{N}} c(t) = r(t)e^{i\left(\theta(t) + \frac{2\pi k}{N}\right)}. \quad (10)$$

Compare (10) to

$$c\left(t + \frac{2\pi}{N}\right) = r\left(t + \frac{2\pi}{N}\right)e^{i\theta\left(t + \frac{2\pi}{N}\right)}. \quad (11)$$

We now get

$$r\left(t + \frac{2\pi}{N}\right) = r(t) \quad (12)$$

$$\theta\left(t + \frac{2\pi}{N}\right) = \theta(t) + \frac{2\pi k}{N}. \quad (13)$$

Equation (12) shows that $r(t)$ is a periodic function with period $2\pi/N$, which is not surprising because the mystery curve is N -fold symmetric. On the other hand, (13) is not so trivial. It actually implies that $\theta(t)$ is a linear function plus a periodic function. More precisely, we can prove

$$\theta(t) = kt + p(t), \quad (14)$$

where $p(t) = p(t + 2\pi/N)$ is a periodic function with period $2\pi/N$.

Proof

Let $p(t) = \theta(t) - kt$ then by definition and (13)

$$\begin{aligned} p\left(t + \frac{2\pi}{N}\right) &= \theta\left(t + \frac{2\pi}{N}\right) - k\left(t + \frac{2\pi}{N}\right) \\ &= \theta(t) + \frac{2\pi k}{N} - kt - \frac{2\pi k}{N} \\ &= \theta(t) - kt = p(t). \end{aligned} \quad (15)$$

□

To illustrate the previous discussion, we take (2) as example. The phase of that mystery curve is given in Figure 4. The linear trend $kt, k = 1$ here can be easily detected. The amplitude $r(t)$ is given in Figure 5, and the code is given in `mysterycurve_exp.m` [3].

The advantage of polar-form design is that the mystery curves can be easily generated and discovered by replacing the periodic functions, due to decoupling the amplitude $r(t)$ and phase $\theta(t)$. Instead of ordinary triangular function \sin or \cos , the absolute value or fractional form of triangular functions such as

$$r(t) = \frac{1}{|\cos(t)| + |\sin(t)|} \quad (16)$$

$$r(t) = \frac{c_1}{c_2 + \cos(Nt)} \quad (17)$$

can be used since they are also periodic. The $r(t)$ in (16) together with $\theta(t) = t$ give us a diamond as shown in Figure 6(a). In Figure 6(b), we draw a periodic function

$$r(t) = \frac{1}{1.3 + \sin(7t)}.$$

Note the constant 1.3 is slightly greater than one to avoid division by zero. In Figure 6(c), we illustrate the mystery curve with this $r(t)$ and $\theta(t) = t + \sin(7t)$. In Figure 6(d) and (e), we simply use the same $\theta(t)$ and replace the $r(t)$ by

$$r(t) = \frac{2 \cos(7t) - 1}{1.1 + \sin(7t)}$$

and

$$r(t) = \frac{2 \cos(7t) + \sin(14t) - 1}{2 + \sin(7t)}.$$

This replacement works only in polar form. For example, the $\sin(14t)$ of the fivefold mystery curve in (1) cannot be easily substituted into any atypical periodic function with the same period such

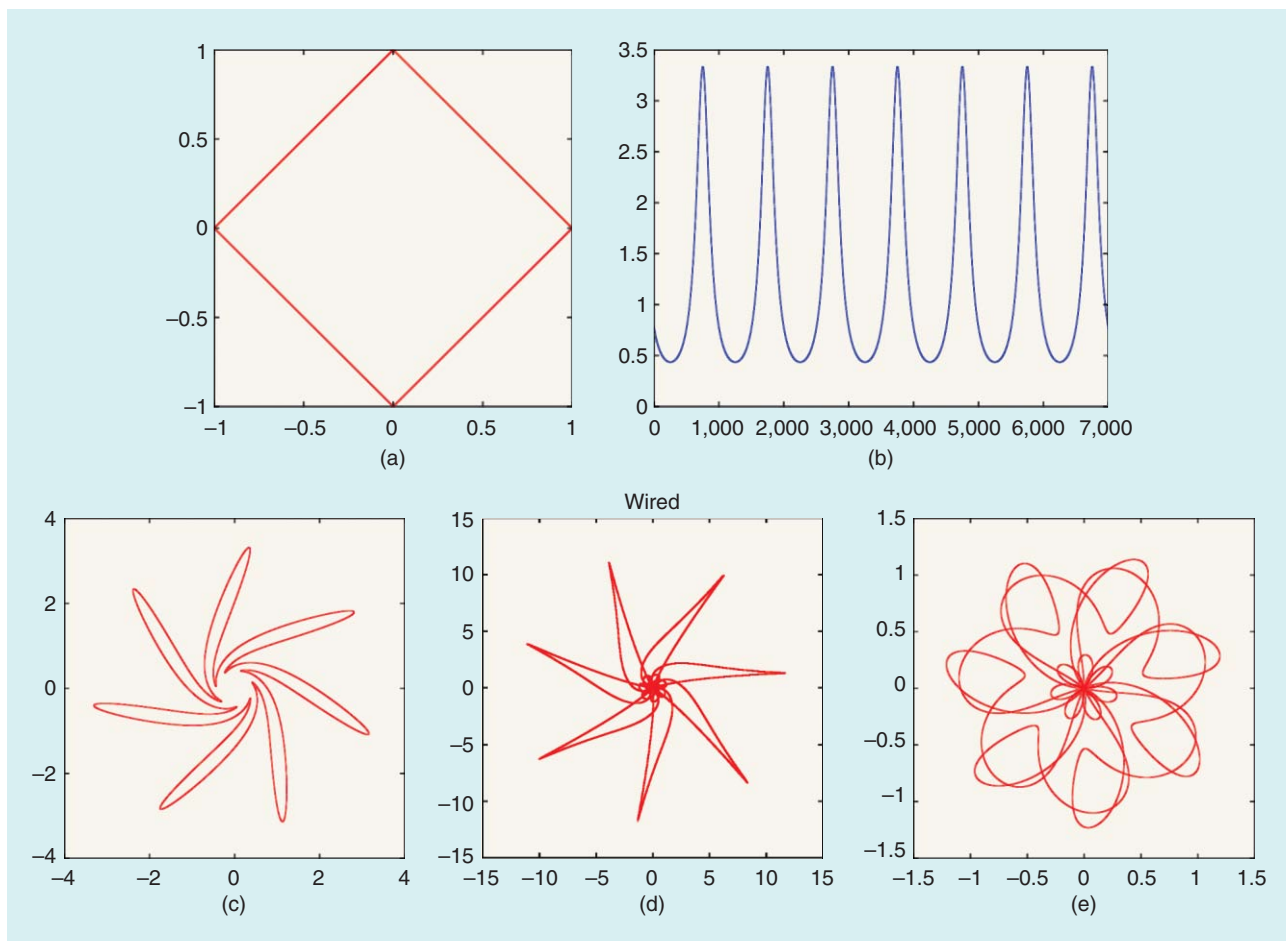


FIGURE 6. (a) The diamond mystery curve $r(t) = 1/|\cos(t)| + |\sin(t)|$ and $\theta(t) = t$. (b) A fractional periodic signal $r(t) = 1/1.3 + \sin(7t)$. (c) The mystery curve corresponding to the same $r(t)$ in (b) and $\theta(t) = t + \sin(7t)$. (d) The mystery curve with the same $\theta(t)$ in (c) but replacing $r(t)$ by $(2\cos(7t) - 1)/(1.1 + \sin(7t))$. (e) Replacing $r(t)$ by $(\cos(7t) + \sin(14t) - 1)/(2 + \sin(7t))$.

as $1/(1.1 + \sin(14t))$ because $v(t)$ must be changed as well. More importantly, $1/(1.1 + \sin(14t))$ introduces some high harmonic frequency terms such as $\cos(28t)$, $\cos(42t)$, etc. Those terms contradict the condition that the difference of each frequency component is a multiple of five.

To illustrate how to design an N -fold symmetric pattern by polar form, assume we can trace the $r(t)$ and $\theta(t)$ in $[0, 2\pi/N]$, as shown by the blue curve in Figure 7. To satisfy conditions (12) and (13), the $r(t)$ must repeat in $[2\pi/N, 2\pi]$ while $\theta(t)$ repeats and adds a linear term. In Figure 8, we present another mystery curve made from the polar coordinate system. The code of using polar form is given in *mysterycurve_made2.m* [3].

The main property of the polar-form mystery curve is that its Fourier series expansion may not have finite terms. In

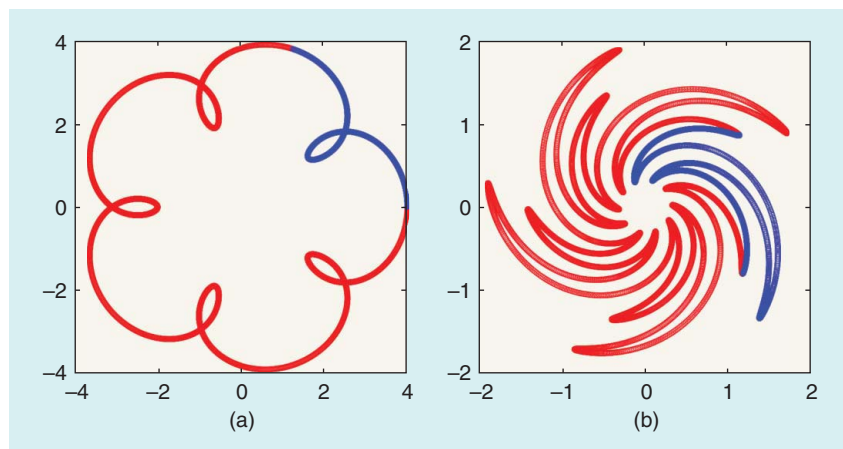


FIGURE 7. (a) A mystery curve made by polar form $r(t) = 3 + \cos(5t)$ and $\theta(t) = t + \sin(5t)/3$. (b) Another polar form mystery curve. $r(t) = 1 - (\cos(5t)/4) + (2\cos(10t)/3)$, $\theta(t) = t + (2\cos(5t)/5) - \cos(10t)$.

Figure 9, we show the spectrum of the complex signal $c(t)$ shown in Figure 8. There are some significant bins in the spectrum, and we can perform low-pass

filtering to approximate or smooth the signal. The filtered results are given in Figure 10(a) and (b). One can see that Figure 10(b) is already very similar

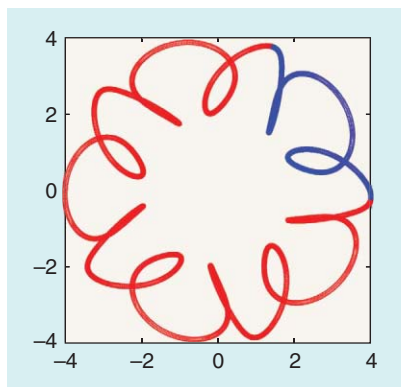


FIGURE 8. Another mystery curve made by polar form, as shown in *mysterycurve_made2.m* [3].

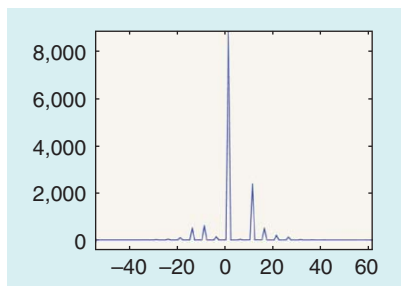


FIGURE 9. The spectrum of the complex signal $c(t)$ shown in Figure 8.

to Figure 8, but there are only five nonzero terms in the frequency response.

From the view of the polar form, we can also notice that the $x(t) = r(t)\cos(\theta(t))$ and $y(t) = r(t)\sin(\theta(t))$ are actually AM-FM signals, which are widely used in speech analysis [8]–[10] and other areas [11], [12]. In particular, the $r(t)$ is called the AM part and $d\theta/dt$ is the instantaneous frequency that corresponds to the FM part. In Figure 11(b),

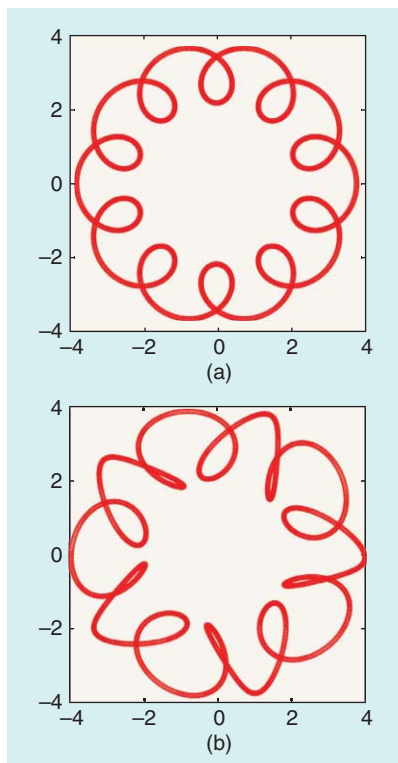


FIGURE 10. (a) Preserving the lowest two bins in Figure 9. (b) Preserving the lowest five bins.

an AM-FM signal is given. The envelope [AM part, $r(t)$] is given in Figure 11(a). The mystery curve corresponding to the AM-FM signal is given in Figure 11(c). This mystery curve, compared with the other curves mentioned in this article, is less sharp. The reason is that the AM signal has a slow changing envelope ($r(t)$) and a rather fast phase ($\theta(t)$) variation. In other words, in a small amount of time Δt , the difference of $r(t)$ and $r(t + \Delta t)$

is small while the phase shift $\theta(t)$ and $\theta(t + \Delta t)$ is large, which makes the curve round.

Mystery curve in 3-D

Like the case in 2-D, the mystery curve in 3-D can be defined by time shift and rotation operation. Recall that, by the Euler rotation theorem, any combination of three dimension rotations is equivalent to a single 2-D rotation [13] about an axis. So, without loss of generality, we can assume the axis to be the z -axis. Therefore, the $x(t)$ and $y(t)$ must satisfy the same constraints as the 2-D case, and $z(t)$ is a periodic signal with period $2\pi/N$. For example, Figure 12(a) is a 2-D mystery curve, and if we put it into 3-D by setting $z(t) = 1 + \cos(5t)$, the result is shown in Figure 12(b). The code is given in *mysterycurve_3D.m* [3].

Because the rotation axis is z , it is easy to generate a 3-D fivefold symmetric curve whose shape looks like a vase or pineapple, as we can observe in Figure 12(c), which contains only the color red, and Figure 12(d), which is distinguished by five colors.

Another example is given in Figure 12(e) and (f). In Figure 12(e), we illustrate a threefold 3-D mystery curve whose rotation axis is not z . To show the symmetry, we color the curve in red, green, blue (RGB) and provide another view angle in Figure 12(f). The code is given in *mysterycurve_3D2.m* [3].

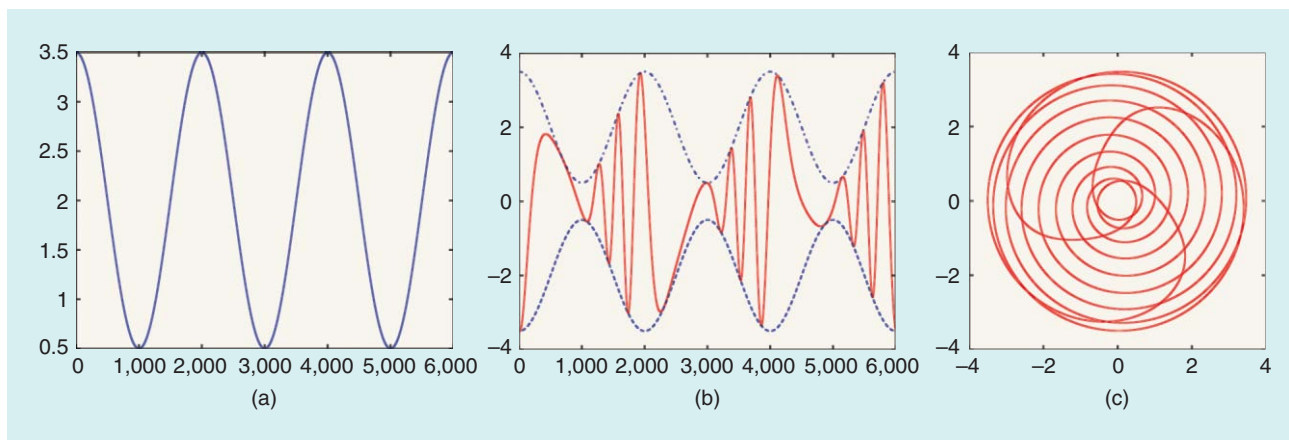


FIGURE 11. (a) The AM part $r(t) = 2 + 1.5\cos(3t)$. (b) The AM-FM signal $r(t)\cos(\theta(t))$ where $r(t)$, indicated as the blue dash line, is the same as (a) and $\theta(t) = 11t + 3\cos(3t)$. (c) The mystery curve corresponding to the same $r(t)$ and $\theta(t)$. The code is given in *myster_AMFM.m* [3].

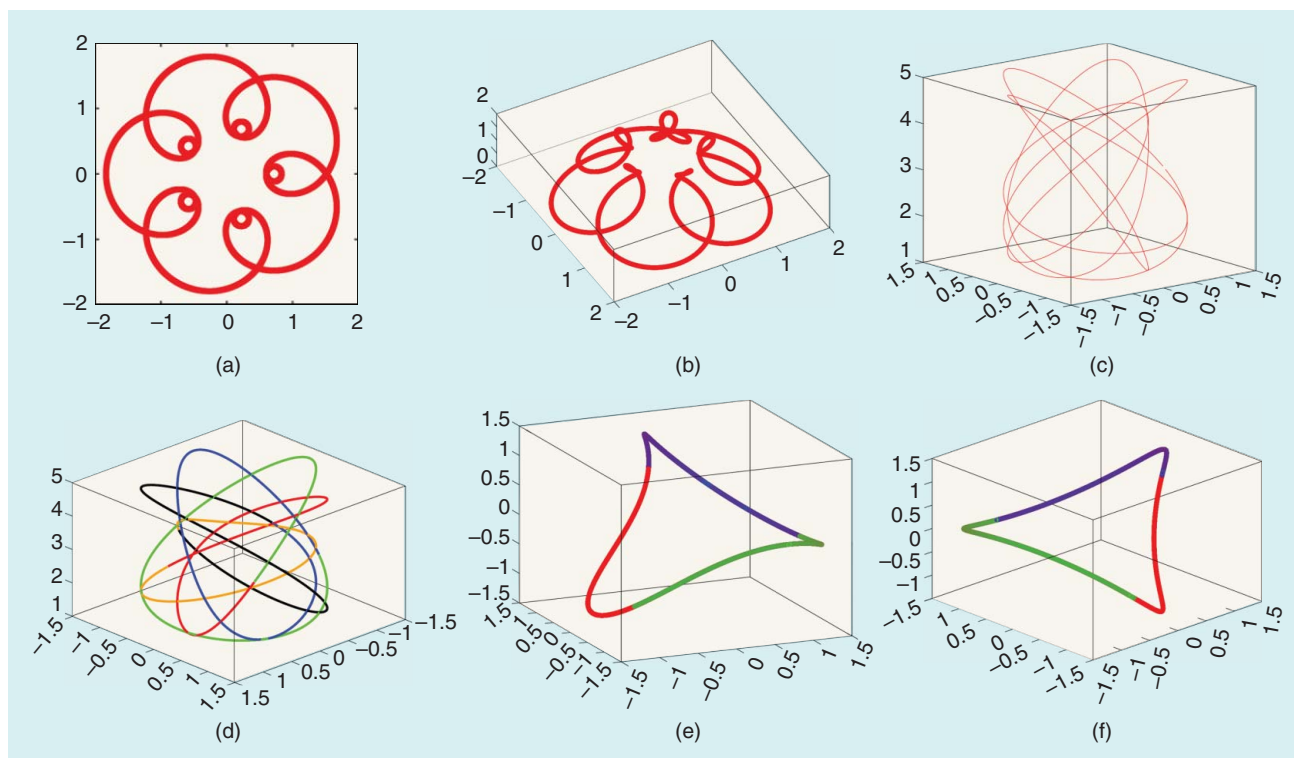


FIGURE 12. Mystery curves in 3-D. (a) The original 2-D mystery curve. (b) The 3-D mystery curve made from (a) by setting $z(t) = 1 + \cos(5t)$. (c) The 3-D curve looks like a pineapple. (d) The 3-D curve looks like a pineapple from another point of view. This curve is fivefold symmetric so that we use five different colors to indicate each fold. (e) Another 3-D mystery curve. This is a threefold symmetric curve, distinguished by the colors RGB, and the rotation axis is not z-axis. (f) The same curve as (e) with a different angle to show symmetry.

What we have learned

In this article, we have learned some properties and results of the mystery curve. By introducing the polar form of such a signal, which is related to a special case of AM-FM signal, we can now easily design these beautiful curves and the repeated patterns can be changed flexibly. We have also learned how to approximate the curve by taking finite terms after low-pass filter. By the Euler rotation theorem, the results we achieved from 2-D cases can extend to 3-D cases directly.

Authors

Soo-Chang Pei (peisc@ntu.edu.tw) received the B.S. degree from National Taiwan University, Taipei, in 1970 and the M.S. and Ph.D. degrees from the University of California, Santa Barbara in 1972 and 1975, respectively, all in electrical engineering. Currently, he is a professor in the Electrical Engineering Department at National Taiwan University. His research interests include digital signal processing, image

processing, optical information processing, and laser holography. He is an IEEE Life Fellow and a member of Eta Kappa Nu and the Optical Society of America.

Kuo-Wei Chang (muslimchang@gmail.com) received the B.S. degree from the Department of Electrical Engineering in 2004 and the M.S. degree from the Graduate Institute of Communication Engineering in 2006, both from National Taiwan University, Taipei, where he is a Ph.D. student in communication engineering. He is currently a researcher with Telecommunication Laboratories, Chunghwa Telecom Co., Ltd. Some of his current research areas include digital signal processing, music information retrieval, and pattern recognition.

References

- [1] F. Farris, *Creating Symmetry: The Artful Mathematics of Wallpaper Patterns*. Princeton, NJ: Princeton Univ. Press, 2015.
- [2] B. Yorgey. (2015). Mystery curve, animated. [Online]. Available: <https://mathlesstraveled.com/2015/06/04/random-cyclic-curves-5/>
- [3] K.-W. Chang. (2017). Mystery curve codes. [Online]. Available: <https://github.com/KuoWeiChang/MysteryCurve>

[4] B. Yorgey. (2015). Mystery curve, animated. [Online]. Available: <https://mathlesstraveled.com/2015/06/04/random-cyclic-curves-5/>

[5] A. V. Oppenheim, W. Schaffer, and R. B. John, *Discrete-Time Signal Processing*. Englewood Cliffs, NJ: Prentice Hall, 1989.

[6] I. Hafner. (2016). Symmetry of a mystery curve. [Online]. Available: <http://demonstrations.wolfram.com/SymmetryOfAMysteryCurve/>

[7] C. Hill. (2016). The mystery curve. [Online]. Available: <http://scipython.com/blog/the-mystery-curve/>

[8] D. Dimitriadis, P. Maragos, and A. Potamianos, "Robust AM-FM features for speech recognition," *IEEE Signal Process. Lett.*, vol. 12, no. 9, pp. 621–624, 2005.

[9] Y. Pantazis, O. Rosec, and Y. Stylianou, "Adaptive AM-FM signal decomposition with application to speech analysis," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 2, pp. 290–300, 2011.

[10] R. Sharma, L. Vignolo, G. Schlotthauer, M. Colominas, H. L. Rufiner, and S. Prasanna, "Empirical mode decomposition for adaptive AM-FM analysis of speech: A review," *Speech Commun.*, vol. 88, pp. 39–64, Jan. 2017.

[11] C. T. Nguyen and J. P. Havlicek, "On the amplitude and phase computation of the AM-FM image model," in *Proc. 2014 IEEE Int. Conf. Image Processing*, 2014, pp. 4318–4322.

[12] G. Biagetti, P. Crippa, A. Curzi, S. Orcioni, and C. Turchetti, "Analysis of the EMG signal during cyclic movements using multicomponent AM-FM decomposition," *IEEE J. Biomed. Health Inform.*, vol. 19, no. 5, pp. 1672–1681, 2015.

[13] B. Palais, R. Palais, and S. Rodi, "A disorienting look at Euler's theorem on the axis of a rotation," *Amer. Math. Monthly*, vol. 116 no. 10, pp. 892–909, 2009.

SP

TIPS & TRICKS

Vicente Torres, Javier Valls,
and Richard Lyons

Fast- and Low-Complexity atan2(a,b) Approximation

This article presents a new entry to the class of published algorithms for the fast computation of the arctangent of a complex number. Our method uses a look-up table (LUT) to reduce computational errors. We also show how to convert a large-sized LUT addressed by two variables to an equivalent-performance smaller-sized LUT addressed by only one variable. In addition, we demonstrate how and why the use of follow-on LUTs applied to other simple arctan algorithms produce unexpected and interesting results.

Introduction

The computation of the arctangent function atan2(a,b), i.e., obtaining the angle of a complex number $c = b + ja$, has been the subject of extensive study because this computation is needed in many applications, for example, in the frequency, phase, and time synchronization stages of digital communications, digital FM demodulation, target tracking in wireless sensor networks, and object recognition in the field of image processing. From a designer's point of view, it is useful to have several computation choices since the performance requirements (speed, accuracy, power consumption, etc.) may be different depending on the specific application, and one of those choices may be better suited than others for a given application.

A high-speed computation of atan2(a,b) can be achieved with LUTs,

where the bit-level concatenation of a and b are the values used to address the ROM that stores the output of the function. The LUT method is fast but much memory is required when a decent arctangent accuracy is needed. Another popular option is to use high-order algebraic polynomials, like Chebyshev polynomials or the Taylor series [1]. These methods give good precision, but since the arctangent is highly nonlinear, they lead to long polynomials and intensive computations. In other cases, approximations based on rational functions are used [2]–[4], as they may provide acceptable results with few computations. The coordinated rotation digital computer (CORDIC) algorithm, which requires only shift and add operations, is frequently used to compute the arctangent [1]. However, its sequential nature makes it less adequate when throughput speed is critical.

Instead of using a single complicated equation to achieve high accuracy, as proposed by other authors, our proposal is a two-stage process with a first stage that uses a low-complexity coarse approximation and a second stage that improves the accuracy by means of a small LUT that stores precomputed error values (as a function of the first stage output). Our proposal computes a full-quadrant arctangent faster than other popular options that achieve the same accuracy. We now describe the two processing stages of our proposed atan2(a,b) algorithm.

First stage

The idea behind this stage is to conceptually generate a continuous real-valued

sinusoid $p(t)$ that has the same initial phase angle as the phase of our complex number $c = b + ja$. If $c = |c|e^{j\theta}$, that sinusoid would be

$$p(t) = |c| \cdot \cos\left(\frac{2\pi t}{T} - \theta\right), \quad (1)$$

where t is time and T is the sinusoid's period, as shown in Figure 1.

The reason we care about this $p(t)$ sinusoid is that the time location of $p(t)$'s maximum value, t_m in Figure 1, is proportional to the desired phase angle of $c = b + ja = |c|e^{j\theta}$. The relationship between t_m and θ is found by setting the time derivative of $p(t)$ equal to zero and solving for t_m . Doing so gives us

$$t_m = \frac{T\theta}{2\pi}. \quad (2)$$

The time-domain dimensions of variables t_m and T must, of course, be identical. With no loss in generality, and out of convenience, we assume the time between the $p[n]$ samples is unity. Thus t_m is measured in units and $T = 4$ units. So when we use (2) to compute θ , the

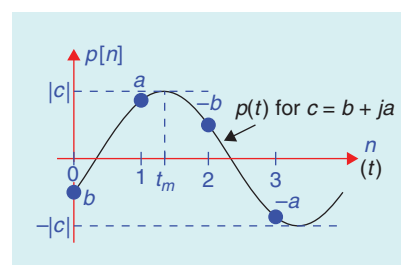


FIGURE 1. The real-valued sequence $p[n]$ and continuous sinusoid $p(t)$ associated with a given complex number $c = |c|e^{j\theta}$.

Digital Object Identifier 10.1109/MSP.2017.2730898
Date of publication: 13 November 2017

ratio t_m/T will be dimensionless and θ will be measured in radians. In the first-stage processing we will estimate t_m and show how this will yield an estimate of angle θ , for a given accuracy, more efficiently than performing other $\text{atan2}(a,b)$ algorithms.

It can easily be seen that the sequence $p[n] = \{b, a, -b, -a\}$ is a sampled version of $p(t)$, as defined by (1) as

$$\begin{aligned}
 p[n] &= \{b, a, -b, -a\} = |c| \cdot \{\cos(\theta), \sin(\theta), -\cos(\theta), -\sin(\theta)\} \\
 &= |c| \cdot \left\{ \cos(-\theta), \cos\left(\frac{\pi}{2} - \theta\right), \cos\left(\frac{2\pi}{2} - \theta\right), \cos\left(\frac{3\pi}{2} - \theta\right) \right\} \\
 &= |c| \cdot \cos\left(\frac{n\pi}{2} - \theta\right) = p(t) \Big|_{t=\frac{nT}{4}}
 \end{aligned}
 \tag{3}$$

where $n = 0, 1, 2, 3$.

Our goal is to compute t_m from the $p[n]$ samples.

To clarify our scenario here, Figure 2 shows the various $p(t)$ waveforms that result from various values of our complex number input c . The time location of the absolute maximum of the sinusoidal $p(t)$ waveform, t_m , is proportional to the angle of c .

The first step of the first-stage processing is to determine the time location of the largest sample of four-sample sequence $p[n]$ (determined from the signs of $a + b$ and $a - b$), a parameter that we call *offset*. The second step of the first-stage processing computes the time location of the maximum value of $p(t)$ relative to offset, a parameter that we call f .

Based on the previously given concepts and relationships, we conclude the first step of the first-stage processing by determining the value for offset, which will be 0, 1, 2, or 3. In the second step of the first-stage processing, we complete the estimation of t_m approximating the value of time variable f . Specifically, we approximate the $p(t)$ signal by a second-order Taylor series in the vicinity of the largest $p[n]$ sample as detailed in "Appendix," which gives us f_r , an approximation of the time location of the maximum value of $p(t)$ relative to that sample

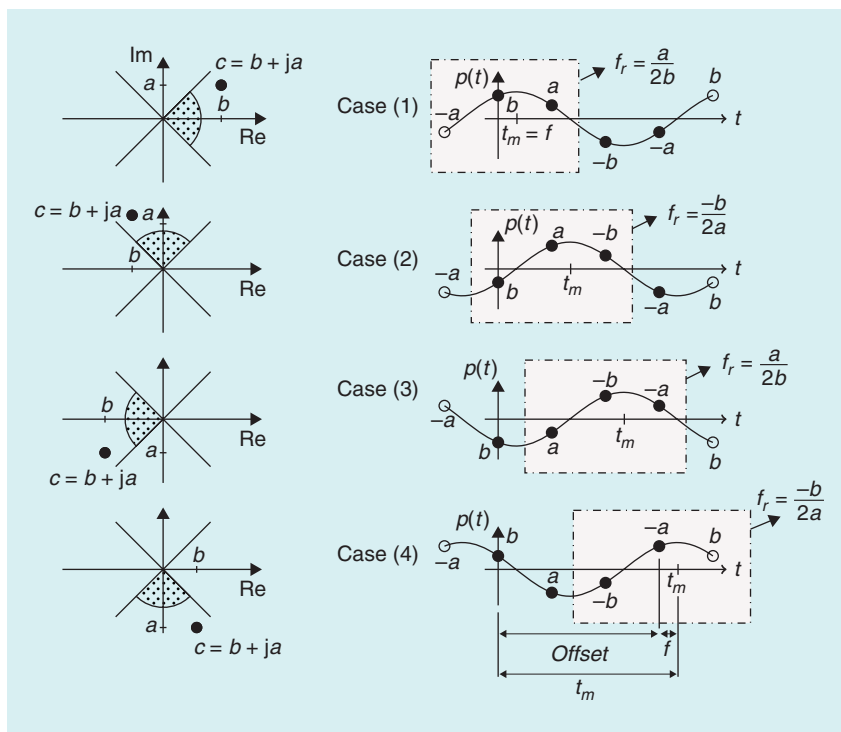


FIGURE 2. The proposed $\text{atan2}(a,b)$ algorithm illustrated for four different possible $c = b + ja$ values.

$$f \approx f_r \equiv \frac{-p'(0)}{p''(0)} = \frac{p(1)}{2p(0)}. \tag{4}$$

In a general case, this computation would require three samples: the biggest of the four samples of the waveform, and also the two samples adjacent to that sample, as depicted for case 1 in Figure 2. But since those two adjacent samples have the same absolute value and opposite sign, only two samples are required in (4): the largest sample $p(0)$ and its following sample $p(1)$. Using (4), we compile our desired processing parameters in Table 1. (Note that a negative value of f_r indicates that $p(t)$ maximum value

occurred prior to the largest sample in $p[n]$.)

Based on the values for offset and f_r from Table 1 and using (2), assuming $T = 4$, the result of the first-stage processing is an approximation of $\text{atan2}(a,b)$, normalized to the range $[0, 1)$, as follows:

$$\begin{aligned}
 \frac{\text{atan2}(a,b)}{2\pi} &= \frac{\theta}{2\pi} = \frac{t_m}{4} \pmod{1} \\
 &\approx \frac{\text{offset} + f_r}{4} \pmod{1},
 \end{aligned}
 \tag{5}$$

where the mod operator is needed to translate negative values to the desired

Table 1. The deduction of the expression for f , as a function of the signs of $a + b$ and $a - b$.

Case	$a + b > 0$	$a - b > 0$	$p(0)$	$p(1)$	Offset	$f_r = \frac{p(1)}{2p(0)}$
1	1	0	b	a	0	$\frac{a}{2b}$
2	1	1	a	-b	1	$\frac{-b}{2a}$
3	0	1	-b	-a	2	$\frac{a}{2b}$
4	0	0	-a	b	3	$\frac{-b}{2a}$

Appendix

The derivation of (4) proceeds in three steps: 1) derive a polynomial expression approximating Figure S1's continuous $p(t)$ sinusoid in terms of known $p[n]$ samples; 2) set that expression's time derivative equal to zero; and 3) replace t with f and solve for f .

Our derivation begins by assuming the largest of our known $p[n]$ samples is located at time $t=0$. Approximating Figure S1's $p(t)$ sinusoid with a second-order Taylor series expression in the vicinity of $t=0$, we begin by writing

$$p(t) \approx p(0) + p'(0)t + \frac{1}{2}p''(0)t^2, \quad (S1)$$

where $p'(0)$ and $p''(0)$ represent the first and second derivatives of $p(t)$ at time $p(t)=0$. Given the (S1) polynomial, we next approximate the unknown $p'(0)$ and $p''(0)$ coefficients by using the central difference formula. Doing so we write the first-order derivative $p'(0)$ as

$$p'(0) \approx \frac{p(h) - p(-h)}{2h}. \quad (S2)$$

Having an approximation of $p'(0)$, we next approximate the second-order derivative $p''(0)$ using the first-order derivatives centered at the hypothetical $p(-h/2)$ and $p(h/2)$ samples in Figure S1. Those first-order derivatives are

$$p'(-h/2) \approx \frac{p(0) - p(-h)}{h} \text{ and}$$

$$p'(h/2) \approx \frac{p(h) - p(0)}{h}.$$

Given $p'(-h/2)$ and $p'(h/2)$, we write our desired second-order derivative $p''(0)$ as

$$p''(0) \approx \frac{p'(h/2) - p'(-h/2)}{h}$$

$$\approx \frac{\frac{p(h) - p(0)}{h} - \frac{p(0) - p(-h)}{h}}{h}$$

$$= \frac{p(h) - 2p(0) + p(-h)}{h^2}. \quad (S3)$$

[0, 1) range. The computation of parameters offset and f_r are shown as the first-stage processing in Figure 3.

The neat trick of our proposed algorithm is that neither the $p[n]$ sequence nor the continuous $p(t)$ signal need to be computed. Our first-stage processing produces a rough estimate of the angle

of $c = b + ja$ based upon some simple logic and simple arithmetic using values a and b . The offset can be computed using the signs of $a + b$ and $a - b$, as shown in Table 1. The determination of the two samples needed for the computation of f_r can also be performed using the signs of $a + b$ and $a - b$. It should

be pointed out that the variable used in the denominator in both expressions for f_r is always the largest absolute value between a and b .

Note that, in [5], Shima used an approximation for the one-variable $\text{atan}(x)$ (derived from a first-order Lagrange polynomial interpolation of

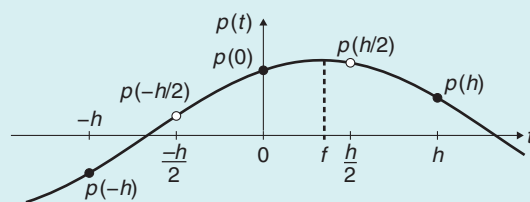


FIGURE S1. The sinusoidal $p(t)$ signal and the desired time value f .

Assuming the time between our known $p[n]$ samples is unity sets $h = 1$ and recalling that, for our $p[n]$ samples, $p[-1] = -p[1]$, we can rewrite (S2) and (S3) as

$$p'(0) \approx \frac{p(1) - p(-1)}{2} = p(1) \quad (S4)$$

$$p''(0) \approx \frac{p(1) - 2p(0) + p(-1)}{1^2}$$

$$= -2p(0). \quad (S5)$$

Substituting (S4) and (S5) as coefficients in (S1), our desired approximation of $p(t)$ is

$$p(t) \approx p(0) + p(1)t - p(0)t^2. \quad (S6)$$

That completes the first step of our derivation. As the second step of our derivation we take the derivative of $p(t)$ to produce

$$p'(t) \approx p(1) - 2p(0)t. \quad (S7)$$

Setting (S7)'s $p'(t) = 0$ gives us an approximation of the time location of the maximum value of Figure S1's $p(t)$ signal. Doing so and defining that estimated time value as f_r we write

$$0 = p(1) - 2p(0)f_r. \quad (S8)$$

Finally, solving (S8) for our desired expression for f in terms of known $p[n]$ sample values we arrive at the final form of (4) as

$$f \approx f_r \equiv \frac{-p'(0)}{p''(0)} = \frac{p(1)}{2p(0)}. \quad (S9)$$

the function in two octants) that would lead to the same expression as ours assuming our a/b or b/a were substituted for his x and his expression is extended to the four quadrants using trigonometric identities.

The coarse approximation of $\text{atan2}(a,b)$ in this first stage can be modeled using the following MATLAB-style code, which returns a normalized $\theta/(2\pi)$ value for the arctangent:

```
function angl=ap_atan2(a,b)
    b0=(a+b)>0;
    b1=(a-b)>0;
    offset=2*not(b0)+not(xor(b1,b0));
    if b0==b1
        fr=-0.5*b/a;
    else
        fr=0.5*a/b;
    end
    angl=mod((offset+fr)/4,1);
end
```

Second stage

The normalized error obtained using the first stage is shown in Figure 4(a), as a function of the actual angle of the complex number $c = b + ja$.

This error function should not be directly stored in an LUT, as it needs to be addressed by the concatenation of the inputs a and b , requiring a large amount of storage. However, we use the trick of transforming that error function to one that only depends on the coarse approximation calculated in the previous stage. Therefore, our proposed second stage improves the accuracy of the first-stage result using an error LUT addressed by $|f_r|$.

The MATLAB function “errorLUTcontents” indicates how this two-variable to one-variable addressing transformation is done where the absolute value of the error is as a function of the absolute value of variable f_r .

```
function f2=errorLUTcontents(wLUT);
    x=linspace(0,0.5,2^wLUT);
    t=linspace(0,pi/4,2^wLUT);
    c=exp(1j*t);
    f1=atan2(imag(c),real(c))/(2*pi);
```

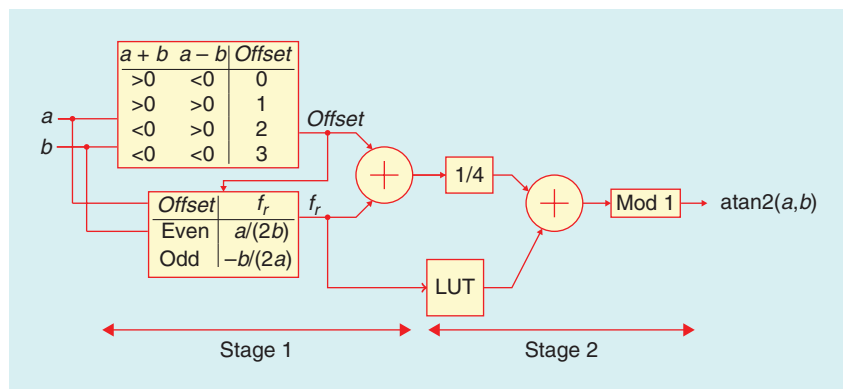


FIGURE 3. Building blocks for the proposed $\text{atan2}(a,b)$ algorithm.

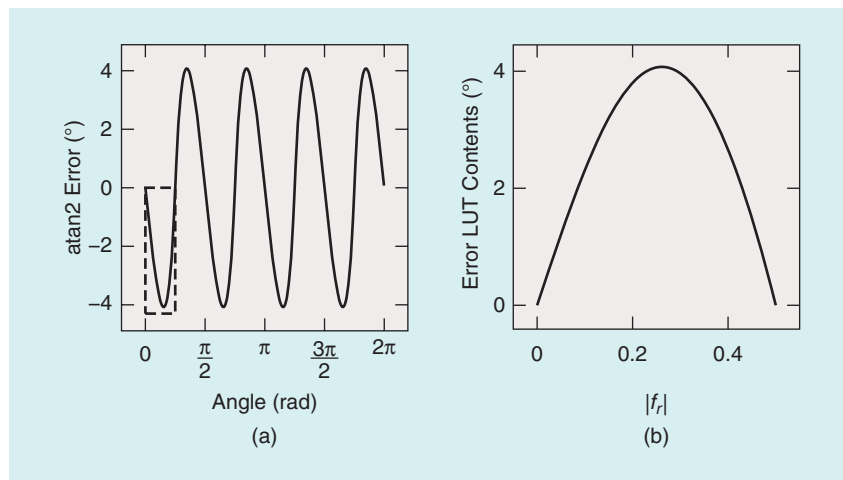


FIGURE 4. Error curves: (a) the first stage arctangent error in degrees and (b) the contents of the second-stage error LUT.

```
f2=0.125*abs(imag(c))./(abs(real(c)));
error=f1-f2;
f2=interp(4*f2,error,x,'pchip');
end
```

The contents of an LUT-named table that contains 2^{wLUT} words can be computed using

```
table=errorLUTcontents(wLUT);
```

In function “ap_atan2,” right after f_r is computed, the following two sentences would be used to include the second stage in the model

```
fix=sign(fr)*table(1+floor(abs(fr)*2^(wLUT+1)));
angl=mod(angl+fix,1);
```

Note that, if this method were implemented using finite precision arithmetic,

the least-significant bit of the table would be around three positions lower than the target accuracy desired for the whole operator.

In summary, our complete algorithm to approximate the angle of a complex number is to 1) identify the maximum of the four $p[n]$ samples in Figure 1 to determine the value of offset, 2) compute the time location f_r of the $p(t)$'s maximum and combine that f_r with offset value from step 1, and 3) improve the result of step 2 using a relatively small-error LUT.

Results and performance

In this section, we will compare our approach to several known low-complexity approximations of the atan2 function.

Table 2 summarizes the computational resources needed by our proposal and a few other atan2 approximations, grouped for akin accuracies. Whenever

Table 2. The computational cost comparison of our proposed algorithm with various previously published arctangent algorithms.

Method	/	*	+	LUT Size (Words)	Maximum Error (Degrees)
[5, eq. (4.13)]	1	0	3	–	± 4.075
[6, eq. (12)]	2	0	4	–	± 4.075
Ours	1	0	5	32	± 0.249
[2, eq. (2)]	1	3	5	–	± 0.276
Ratio+LUT	1	0	4	128	± 0.224
Ours	1	0	5	64	± 0.126
[4, eq. (18)]	1	4	4	–	± 0.162
[3, eq. (9)]	1	3	5	–	± 0.086
Ratio+LUT	1	0	4	256	± 0.112
Ours	1	0	5	1K	± 0.008
[4, eq. (16)]	1	7	6	–	± 0.008
Ratio+LUT	1	0	4	4K	± 0.007

is used, the approximation of the $\text{atan2}(a,b)$, has a maximum error of $\pm 4.07^\circ$, which corresponds to 6.5 exact bits [i.e., the number of most significant bits that are zero in the binary representation of the maximum absolute value of the error; this value can be obtained as $-\log_2(\max(\text{abs}(\text{error})))$, assuming a $[0, 1)$ normalization of the values]. Another coarse approximation was presented in [6, eq. (12)], achieving the same accuracy with higher computational cost.

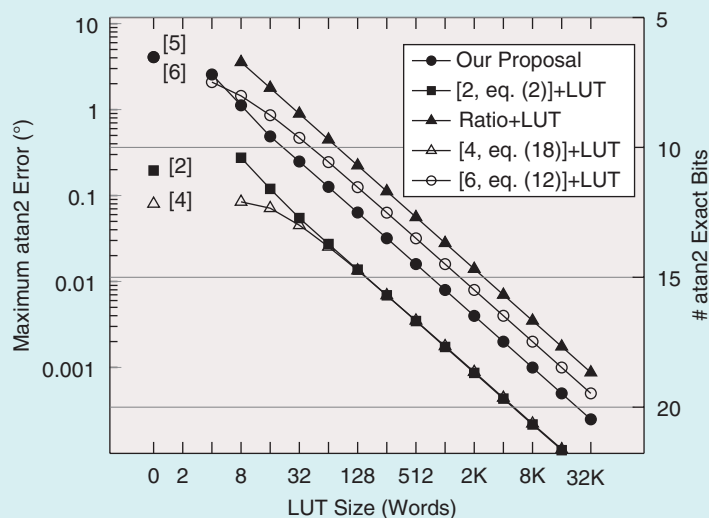
As shown in Table 2, by using a small error LUT of 32 values our proposal achieves similar accuracy to Lyons' [2, eq. (2)] and Rajan et al.'s [3, eq. (9)], but requiring three and two fewer multiplications, respectively. If larger LUT sizes are used, similar accuracy to [4]–(18) and (16) can be achieved with fewer arithmetic resources by our method.

Finally, another option considered for comparison purposes is the method described in [7], which we call the *Ratio+LUT method*: first, the ratio $z = a/b$ is calculated with a division operation (to avoid a large LUT storing a two-variable function), second, the one-variable function $\text{atan}(z)$ is computed using a LUT. Note that, to extend the computation to a full-quadrant atan2 , four additions would be needed. As seen in Table 2, when using the Ratio+LUT, the size of this LUT would be four times larger than in our proposal, for the same final accuracy. Moreover, in fixed-point implementations of the algorithm, the largest value stored in the atan2 LUT would be 3.5 bits larger than in our case, making the quantization noise worse for the same LUT size and word length.

Using a different first stage

As we have shown (see Table 2), the first stage of our algorithm requires the least arithmetic resources. Nevertheless, in this section, we explore the idea of adding a second stage based on an error LUT to other atan2 algorithms. Adding a second stage could be an easy way of improving the accuracy of an existing implementation with minimum design cost.

We have used [6, eq. (12)], [2, eq. (2)], and [4, eq. (18)] as first stages in a two-stage algorithm. Their error was

**FIGURE 5.** The maximum arctan error and number of exact bits versus LUT size.

an algorithm is proposed for a two-octant one-variable $\text{atan}(x)$, three extra addition/subtractions are included in this table as the cost of extending the approximation to all the quadrants. In Table 2, we only consider divisions, multiplications, additions/subtractions, and storage requirements to evaluate the computational cost of the algorithms. Other required operations have a computational cost that can be highly platform dependent. Operators like binary shifts, $\text{mod}()$, $\text{floor}()$, bit string concatenations, etc. have no cost at all in fixed-point application-specific integrated

circuit (ASIC) or field-programmable gate array (FPGA) implementations but may result in additional computational time in pure software implementations. For example, in a fixed-point FPGA implementation, the computation of $\text{offset} = 2 * \text{not}(b_0) + \text{not}(\text{xor}(b_1, b_0))$ doesn't involve multiplications nor additions, just bit concatenations and simple logic operators. In such a case, the hardware architecture could be implemented following the data flow illustrated in Figure 3.

When only the first stage of the algorithm, i.e., without the error LUT,

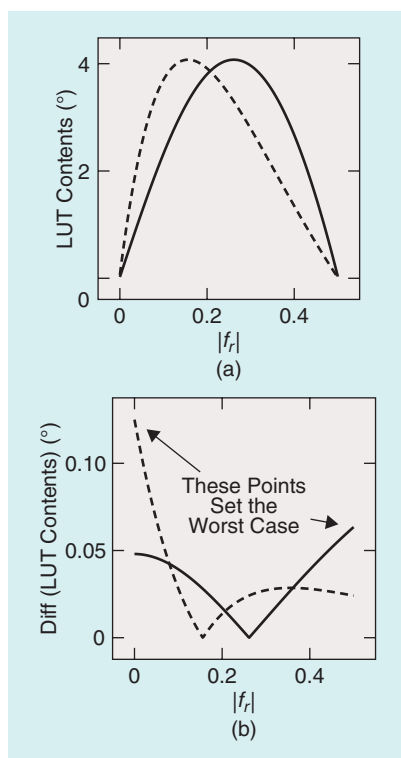


FIGURE 6. LUT characteristics: (a) LUT contents and (b) absolute value for the derivatives of the LUT contents for our proposal (solid line) and [6, eq. (12)] (dashed line).

computed, using a modified version of the “errorLUTcontents” function, and the maximum atan2 error was measured for several LUT sizes. The results are shown in Figure 5. When a second-stage LUT is used, the accuracy is improved significantly. As can be seen, the maximum error is halved (i.e., one exact bit more is achieved) when the size of the LUT is doubled.

Another interesting observation is that, when the second stage is added, the atan2 maximum error depends on the maximum first derivative of the error curve stored in the LUT. This explains why, for example, for the same atan2 accuracy, [6, (eq. 12)] would require a LUT twice as large as ours. Figure 6(a) shows the LUT contents for both approaches: ours with a solid line and [6, eq. (12)]’s with a dashed line. Figure 6(b) shows the absolute value of the difference between consecutive values in the LUT, for the specific case of an LUT with 256 words. As can be seen in those figures, our approach has a smaller maximum of the absolute value of the first derivative of the error curve.

On the contrary, [4, (eq. 18)] achieves better accuracy than [2, (eq. 2)], but when a second-stage LUT is added, their accuracy is similar. That’s because, in this case, they have similar maximum values of the first derivative of their error curves. An important lesson learned is that atan2 approximations with smaller first derivative error values are better suited for the addition of a second-stage LUT.

Conclusions

We propose a full-quadrant algorithm for the computation of the arctangent of a complex number $c = b + ja$, particularly suitable for implementations in hardware, e.g., FPGA, ASIC, etc., where there is no penalty incurred when accessing a LUT. The second stage of the method we propose could be applied to other low-complexity algorithms for the approximation of the atan2 function, but for a given accuracy there is a tradeoff between the complexity of the approximation used for the first stage and the required storage resources used in the second stage. As we have shown, algorithms with a smaller first derivative of their error curve are best suited for improving the accuracy by the addition of a second-stage LUT. Because our proposed method can be easily improved by increasing the size of a memory when higher accuracy is needed, it is an attractive arctan method in high-speed applications where moderate accuracy is required (e.g., in systems where the precision of the measured a and b variables is, say, 14 bits or fewer).

Acknowledgments

This work is funded by the Spanish Ministerio de Economía y Competitividad and FEDER under grant TEC2015-70858-C2-2-R.

Authors

Vicente Torres (vtorres@eln.upv.es) received a telecommunications engineering degree from the Universitat Politècnica de Valencia (UPV) in 1994 and received the Ph.D. degree in telecommunication engineering from the same university in 2001. He is an associate professor at UPV. He is currently a member of the Institute for Telecommunications and Multimedia

Applications. His research work is focused on digital signal processing in electronic devices. He has contributed to more than 50 papers in renowned technical journals and conferences and has been an active member of several financially aided research projects.

Javier Valls (jvalls@eln.upv.es) received the telecommunication engineering degree from the Universitat Politècnica de Catalunya, Spain, and the Ph.D. degree in telecommunication engineering from the Universitat Politècnica de Valencia, Spain, in 1993 and 1999, respectively. He has been with the Department of Electronics at the Universitat Politècnica de Valencia since 1993, where he is an associate professor. His current research interests include the design of field-programmable gate array-based systems, computer arithmetic, very-large-scale integration signal processing, and digital communications.

Richard Lyons (R.Lyons@ieee.org) is a consulting signal processing engineer. Winner of the IEEE 2012 Education Award, he is the author of *Understanding Digital Signal Processing 3/E* (Prentice Hall, 2010). He is the editor of, and contributor to, *Streamlining Digital Signal Processing, A Tricks of the Trade Guidebook* (IEEE Press/Wiley, 2007) and the coauthor of *The Essential Guide to Digital Signal Processing* (Prentice Hall, 2014).

References

- [1] J.-M. Muller, *Elementary Functions: Algorithms and Implementation*. Cambridge, MA: Birkhäuser, 1997.
- [2] R. G. Lyons, “Another contender in the arctangent race,” *IEEE Signal Process. Mag.*, vol. 21, no. 1, pp. 109–110, Jan. 2004.
- [3] S. Rajan, S. Wang, R. Inkol, and A. Joyal, “Efficient approximations for the arctangent function,” *IEEE Signal Process. Mag.*, vol. 23, no. 3, pp. 108–111, May 2006.
- [4] X. Girones, C. Julia, and D. Puig, “Full quadrant approximations for the arctangent function,” *IEEE Signal Process. Mag.*, vol. 30, no. 1, pp. 130–135, Jan. 2013.
- [5] J. M. Shima, “FM demodulation using a digital radio and digital signal processing,” M.S. thesis, Univ. Florida, Gainesville, 1995.
- [6] S. Winitzki, *Uniform Approximations for Transcendental Functions*. Berlin, Heidelberg: Springer, 2003, pp. 780–789.
- [7] R. Gutierrez and J. Valls, “Implementation on FPGA of a LUT-based atan(y/x) operator suitable for synchronization algorithms,” in *Proc. Int. Conf. Field Programmable Logic and Applications*, 2007, pp. 472–475.



DATES AHEAD

Please send calendar submissions to:
Dates Ahead, Attn: Jessica Welsh, E-mail: j.welsh@ieee.org

2017

NOVEMBER

Fifth IEEE Global Conference on Signal and Information Processing (GlobalSIP)

14–16 November, Montréal, Canada.
General Cochairs: Warren Gross and Kostas Plataniotis
URL: <http://2017.ieeeglobalsip.org>

DECEMBER

Ninth IEEE Workshop on Information Forensics and Security (WIFS)

4–7 December, Rennes, France.
General Cochairs: Teddy Furon and Carmela Troncoso
URL: <http://wifs2017.org/>

Seventh IEEE Conference of the Sensor Signal Processing for Defence (SSPD)

6–7 December, Edinburgh, Great Britain.
General Chairs: Mike Davies, Jonathon Chambers, and Paul Thomas
URL: www.sspd.eng.ed.ac.uk/

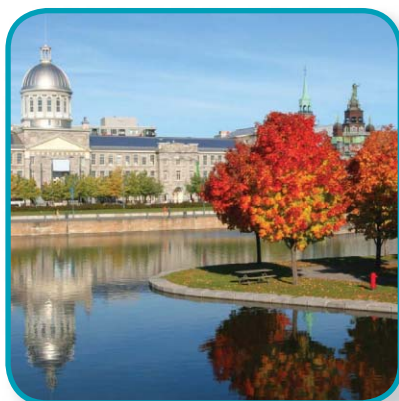
17th IEEE International Workshop on Computational Advances in Multisensor Adaptive Processing (CAMSAP)

10–13 December, Curacao, Dutch Antilles.
General Chairs: André L.F. de Almeida and Martin Haardt
URL: <http://www.cs.huji.ac.il/conferences/CAMSAP17/>

IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)

16–20 December, Okinawa, Japan.
General Chairs: John Hershey and Tomohiro Nakatani
URL: <https://asru2017.org>

Digital Object Identifier 10.1109/MSP.2017.2743878
Date of publication: 13 November 2017



GlobalSIP 2017 will be held in Montréal, Canada, from 14 to 16 November.

17th IEEE International Symposium on Signal Processing and Information Technology (ISSPIT)

18–20 December, Bilbao, Spain.
General Cochairs: Begona García-Zapirain and Adel Elmaghraby
URL: <http://www.isspit.org/isspit/2017/>

2018

MARCH

Data Compression Conference (DCC)

27–30 March, Snowbird, Utah, United States.
General Cochairs: Michael W. Marcellin and James A. Storer
URL: <http://www.cs.brandeis.edu/~dcc/>

APRIL

IEEE International Symposium on Biomedical Imaging (ISBI)

4–7 April, Washington, D.C., United States.
Conference Chairs: Amir Amini and Scott Acton
URL: <https://biomedicalimaging.org/2018/>

43rd IEEE International Conference on Acoustic, Speech, and Signal Processing (ICASSP)

15–20 April, Calgary, Canada.
General Chair: Monson Hayes
General Cochair: Hanseok Ko
URL: <http://2018.ieeeicassp.org/>

JUNE

20th IEEE Statistical Signal Processing Workshop (SSP)

10–13 June, Freiburg, Germany.
General Chair: Peter Schreier
URL: <https://ssp2018.org/>

19th IEEE International Workshop on Signal Processing Advances in Wireless Communications (SPAWC)

25–28 June, Kalamata, Greece.
General Cochairs: Athina Petropulu and Constantinos B. Papadias
URL: <http://spawc2018.org/>

JULY

IEEE International Conference on Multimedia and Expo (ICME)

23–27 July, San Diego, California, United States.
General Chairs: C.-C. Jay Kuo, Truong Nguyen, and Wenjun Zeng
URL: <http://www.icme2018.org/>

OCTOBER

25th IEEE International Conference on Image Processing (ICIP)

7–10 October, Athens, Greece.

NOVEMBER

Sixth IEEE Global Conference on Signal and Information Processing (GlobalSIP)

26–28 November, Anaheim, CA, United States.
General Chairs: Shuguang Cui and Hamid Jafarkhani
URL: <http://2018.ieeeglobalsip.org/>

PERSPECTIVES (continued from page 176)

but can they be a crystal ball to predict the stock market and earn a large excess return? Before we dream of the Alpha-Go for market prediction, let us review a few basic principles on asset pricing to better understand the role AI could play. I refer to the stock market in this discussion, but the same applies to other assets and securities such as currencies, commodities, and bonds.

Stock price reflects a market equilibrium when supply meets demand, a state in which sellers and buyers agree

In a free market, the price “clears the market,” which means the price at which the quantity people buy and the quantity they sell are the same. Why do prices move? The basic concept is represented in the supply-demand curves. When the quantity that sellers are willing to sell does not meet the quantity that buyers are willing to buy, prices change until a new equilibrium is reached. See Figure 1 for the supply-demand curves.

Now add this to the mix: assume that an AI tool can predict tomorrow’s stock price. Say the price is going to be US\$10 higher tomorrow, represented by the red curves and $P(t+1)$. What will happen?

If everyone in the market has a magic AI tool or at least knows the tool’s prediction, then today’s price will immediately rise to the predicted point. Why? Because no one would sell at a lower price. Indeed, as long as a few market participants have this knowledge, the price will change immediately. There are many examples of this in the real world such as an acquisition announcement. Company A is trading at US\$25. Company B announces it has entered into an agreement to purchase Company A in one month’s time for US\$35. Company’s A stock immediately jumps to US\$35 in anticipation of the deal close price. As a

result, no one can make profit from a known prediction.

An informal way to think about it is that today’s price is the average price of the many possible prices we think we might see tomorrow, the “expected” price. And furthermore, when determining today’s price, people discount tomorrow’s expected price because, in general, people have an aversion to future uncertainty. The dark shadow curves in Figure 1 represent another possibility tomorrow.

When no one else but one person, say, Alice, has this magic AI tool, then there are two kinds of people trading. Alice, who knows the future price, and ignorant people trading at the wrong price today (see the aforementioned first argument) but will gain this knowledge tomorrow. So if Alice buys the stock today, she would make a free profit when the price rises to where she knows it will be tomorrow. Note that as Alice buys the stock, the price would be moving up along the supply curve, and the demand curve would be moving slightly right toward its correct place. The more Alice is buying, the more the demand curve will move to the right. Another compounding effect would be that, though other people (suppliers) do not have Alice’s information, they may quickly infer from Alice’s order for a large quantity

of shares that there is something there and raise their expected price, moving the supply curve to the left. In such a case, the free profit Alice can make is apparently limited. Of course, if Alice is not trading in a large quantity and no one pays attention to her trading, she may make a one-time profit for herself by keeping her information secret. Real-world examples such as undetected insider trading, or, arguably, the opaque strategies employed by Jim Simons at Renaissance Technologies, can only profit if others do not know or infer what they know.

Risks or uncertain future outcomes are an inherent nature of the financial market and the underlying economic activities

Stock price is the result of a market equilibrium when market participants optimize risk-returns. The current stock price represents the best knowledge of sellers and buyers for the expected (average) future of the stocks/firms. The excess expected return in the stock market comes as the reward of holding market risk, also called *systematic risk*, which cannot be eliminated.

This is a more involved argument that can be best illustrated by the Capital Asset Pricing Model (CAPM). The argument also holds for other asset-pricing models. A CAPM (see Figure 2) is an equilibrium relationship between stock prices and the market portfolio. It is obtained when people optimize the expected return (mean of the stock return) for a given risk represented by volatility, i.e., the standard deviation of the stock return. This optimization leads to a market equilibrium, i.e., a mean-variance-efficient (MVE) tangency portfolio T , such that the prices of all stocks satisfy the CAPM relationship according to their correlation to the market portfolio, the optimal portfolio. A market index, such as the S&P 500 index,

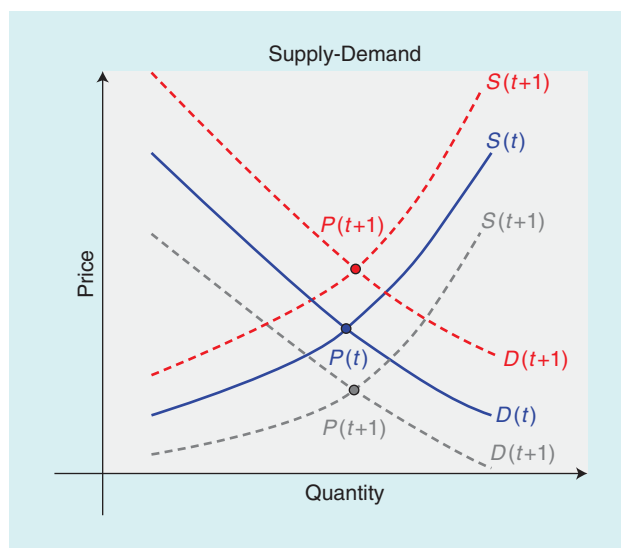


FIGURE 1. An example of the supply-demand and price equilibrium movement at today (time t) and tomorrow (time $t+1$).

is often used as a proxy of the market portfolio. Note that the MVE tangency portfolio is the optimal portfolio in that it gives maximum expected return for a given volatility and risk-free rate. A CAPM is an equilibrium relationship that stock prices must satisfy with regard to their risks and correlations. In a CAPM world, the market portfolio is the MVE tangency portfolio. No one can hold a better portfolio than the market portfolio—the MVE tangency portfolio T . Readers can refer to [2] for an introduction to the CAPM with a signal processing perspective as well as finance jargon I use here.

Of course, such market equilibrium and optimal market portfolio will only hold when major market participants agree on the knowledge, i.e., mean and covariance, of all stock returns. If an AI tool provides better estimate of the knowledge, as long as everyone in the market knows, no one can beat the market portfolio in a risk-return sense.

Now, what if only one group of investors, represented by Alice, has the better knowledge, and the other group of investors, represented by Bob, disagrees and stubbornly holds and acts on the wrong knowledge of the firms? In such a case [3], the informed Alice will hold the true MVE tangency portfolio T , while the misinformed Bob in aggregate holds a non-MVE portfolio B . The market portfolio M is now the value-weight portfolio of T and B . Let w denote the proportion of the total value of risk asset owned by Alice, and R_A and R_B denote the expected returns of portfolios A and B , respectively. Then the market portfolio return R_M is

$$R_M = wR_A + (1 - w)R_B.$$

The relationship among portfolios is illustrated in Figure 2.

Remarks

- Apparently, Alice’s portfolio T has the highest Sharpe ratio and outperforms the market portfolio M .

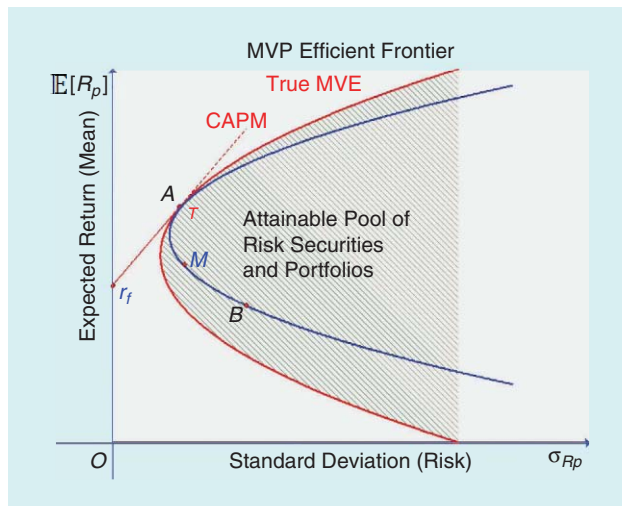


FIGURE 2. Portfolio risk-return optimization and market equilibrium: r_f represents the risk-free rate. The risk-free rate is the theoretical rate of return of an investment with zero risk. In practice, it is often represented by the interest rate on U.S. Treasury bills.

- The stubborn misinformed Bob’s portfolio underperforms the market.
- Note that, if the misinformed investors in aggregate hold the tangency portfolio, the CAPM still holds and no one will outperform the market.
- When the overall weight of the misinformed investors is large enough, it can drag down the market portfolio. This may happen in a market with many nonprofessional stubborn active investors. But in the U.S. stock market, the institutional investment composes the majority. Therefore, M is expected to be similar to the true tangency portfolio T .
- If the nonprofessional retail investors are equipped with AI, they would act more like informed investors that will bring the market portfolio closer to M , making the market more efficient and practically losing less money compared to the informed professional investors.
- Here’s what’s interesting: if Bob knows that he is not as informed as Alice but doesn’t know who Alice is or what her secret holdings are, his best strategy is to be a passive investor holding the market portfolio. If all misinformed investors know they are misinformed and therefore hold the market portfolio, the market portfolio again becomes the optimal MVE tangency portfolio T . However, if

Bob uses a wrong AI tool or data and becomes overconfident that he’s the informed investor in the market, guess what—he becomes the stubborn guy holding the wrong portfolio who’s worse off, dragging the market portfolio down and hurting all the passive investors. Warren Buffett put it this way: “If you’ve been playing poker for half an hour and you still don’t know who the patsy is, you’re the patsy.” The caveat is that it’s easier said than done: you have to have enough unbiased data samples to draw statistically significant conclusion to know who the patsy is.

My conclusion? AI may provide new tools for information. But are you a seasoned financial professional who works closely with the market, or are you a casual investor? If you are one of us, the majority, don’t expect AI to bring a quick buck. If you are the former, you have a tough job but there’s a chance to make a fortune.

Summary and Q&A

Q: Can AI (or any other technologies) help us better evaluate the risk?

A: Yes, it is possible. But from society’s perspective, it is not necessarily a good thing. We want smart people to create totally new knowledge, products, art, etc. that always involve high risk. Accurate risk evaluation may cause these high-risk start-ups to find themselves short of funders, just like an insurance company may not want to insure certain high-risk patients.

Q: Can AI help us better allocate capital in a financial market?

A: Yes.

Q: Can AI replace some financial analysts?

A: Possibly. That may happen in any industry and has happened before in the stock market. The famed trading pits in New York City and Chicago stuffed with floor traders shoving and shouting are no more. They’ve been replaced by electronic exchanges. Now automation

with AI is threatening financial managers who do work at their desktops.

Q: Can AI generate a few new successful hedge funds?

A: Maybe, there is always a chance to become the next Jim Simons who was a mathematician, or D.E. Shaw, who was a computer scientist. But the overall quantity and amount of successful hedge funds are unlikely to change. The top hedge funds today have no choice but to recruiting top AI talent. It's a Darwinian world where you retain your edge or the hedge fund dies.

Q: But can AI make every smart AI expert excessive risk-free money from stock market?

A: The answer is a solid "no!"

Author

Xiao-Ping (Steven) Zhang (xzhang@ryerson.ca) received his B.S. and Ph.D. degrees from Tsinghua University, Beijing, China, in 1992 and 1996, respectively, both in electronic engineering. He received an M.B.A. degree in finance, economics, and entrepreneurship with honors from the University of Chicago Booth School of Business in 2008. He is a professor of electrical and computer engineering and cross-appointed to the Finance Department at the Ted Rogers School of Management at Ryerson University, Ontario, Canada. His research interests include signal processing, machine learning,

big data, finance, and marketing. He is the cofounder and chief executive officer of EidoSearch, an Ontario-based company offering a content-based search and analysis engine for financial big data.

References

[1] (May 24, 2014). Sea of tranquility: The return of moderation. *Economist.*, vol. 411, no. 8888, pp. 70. [Online]. Available: <https://www.economist.com/news/finance-and-economics/21602734-volatility-has-disappeared-economy-and-markets-could-be>

[2] X.-P. Zhang and F. Wang, "Signal processing for finance, economics, and marketing," *IEEE Signal Processing Mag.*, vol. 34, no. 3, pp. 14–35, May 2017.

[3] E. F. Fama and K. R. French, "Disagreement, tastes, and asset prices," *J. Financial Econ.*, vol. 83, no. 3, pp. 667–689, Mar. 2007.

SP

Sign Up or Renew Your 2018 SPS Memberships

A Membership in the IEEE Signal Processing Society (SPS), the IEEE's first society, can help you lay the foundation for many years of success ahead:

- **CONNECT** with more than 19,000 signal processing professionals through SPS conferences, and local events hosted by more than 170 SPS Chapters worldwide.
- **SAVE** with member discounts on conferences and publications, and access to travel grants, SigPort repository, and SPS Resource Center.
- **ADVANCE** with world-class educational resources, awards and recognitions, and society-wide volunteer opportunities in publications, conferences, membership, and more.



Learn more about membership options (including choices of electronic access and print option of the *IEEE Signal Processing Magazine*, SPS Digital Library, and more):

<http://signalprocessingsociety.org/get-involved/membership>

Already a SPS Member? Refer a friend, a student or a colleague to join IEEE & SPS.

IEEE "Member-Get-a-Member" reward is available up to \$90 per year.

Electronic membership options and special rate are available in certain countries.



New benefit from the IEEE Signal Processing Society
SPS Resource Center

The SPS Resource Center is the new home for the IEEE Signal Processing Society's online library of tutorials, lectures, presentations, and more. Unrestricted access to our fast-growing archive is now included with your SPS membership.

<http://rc.signalprocessingsociety.org>

We accept submissions, too!
Interested in submitting your educational materials?
sps-resourcecenter@ieee.org

Digital Object Identifier 10.1109/MSP.2017.2760819



© GRAPHIC STOCK

ADVERTISING & SALES

The Advertisers Index contained in this issue is compiled as a service to our readers and advertisers: the publisher is not liable for errors or omissions although every effort is made to ensure its accuracy. Be sure to let our advertisers know you found them through *IEEE Signal Processing Magazine*.

IEEE SIGNAL PROCESSING MAGAZINE REPRESENTATIVE

Mark David, Director, Business Development — Media & Advertising, Phone: +1 732 465 6473, Fax: +1 732 981 1855, m.david@ieee.org

COMPANY	PAGE NUMBER	WEBSITE	PHONE
Clemson University	9	http://apply.interfolio.com/39731	
Mathworks	CVR 4	www.mathworks.com/wireless	
Southern New Hampshire University	3	snhu.edu	+1 603 645 9611
Southern University of Science and Technology	11	sustc.edu.cn	+86 755 88018558

Digital Object Identifier 10.1109/MSP.2017.2701061

IEEE SIGNAL PROCESSING CUP 2018

GLOBAL UNDERGRADUATE COMPETITION IN SIGNAL PROCESSING

FORENSIC CAMERA MODEL IDENTIFICATION CHALLENGE: The 2018 Signal Processing Cup competition will be on forensic identification of camera model from images. Teams will be tasked with designing a system to determine which camera model captured a digital image without relying upon information in the image's metadata.

WHO CAN PARTICIPATE? Teams formed of 3 to 10 undergraduate students, at most one graduate student, and one faculty member.

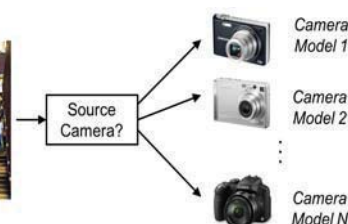
PRIZES: GRAND PRIZE VALUED UP TO \$10K TOTAL

Monetary prizes (up to \$5000), plus travel grants for the top three teams to showcase their work at ICASSP 2018.

IMPORTANT DATES:

- December 1, 2017: Team registration deadline
- January 21, 2018: Project submission deadline
- February 10, 2018: Announcement of finalists
- April 22-27, 2018: Final competition at ICASSP

URL: <http://signalprocessingsociety.org/get-involved/signal-processing-cup>



IEEE
Signal Processing Society

Digital Object Identifier 10.1109/MSP.2017.2768459

To the Victor Go the Spoils: AI in Financial Markets

Artificial intelligence (AI) for finance is a hot topic these days. Top hedge funds are waging a talent war over AI experts the way sports teams compete for pro athletes. Recently, a friend of mine, an AI expert working at a tech firm, got an offer from a top hedge fund. From a financial perspective, it is “an offer he can’t refuse,” as Don Corleone says. In the sci-fi film *Transcendence*, Johnny Depp applied his newly digitized brain to the markets, pocketing hundreds of millions overnight. We’ve witnessed the invincibility of machines—AlphaGo, Watson, DeepBlue—and can’t help but wonder what’s next. In January of this year, a new AI named Libratus joined the club by defeating four of the world’s best professional poker players. How close are we to AI’s conquest of the game of finance? Will AI experts become the new titans of Wall Street?

Note that AI and signal processing (SP) have a natural connection. Indeed, AI for SP recently has achieved pioneering success in speech and image processing through deep neural network (DNN) technologies. There exists the unbridled expectation people and companies might have on this connection (SP + AI) to make money.

To properly align our expectation, we need to understand the nature of financial markets and, within the laws of that universe, what AI can do for us as a

community and for you as an individual investor. From this perspective, I will show that AI cannot change the fundamental structure and dynamics of the financial market. Like all other technologies and innovations introduced to finance, AI may advance our markets and economy, automate away jobs, and even bring a select few vast wealth. But the vast majority of people—no matter their level of AI expertise—will not achieve large excess returns. So don’t plan for an early retirement just yet.

A fundamental function of the financial market is resource allocation, i.e., for businesses or people to secure resources for promising investment projects. Everyone knows, or quickly learns, that investment returns are not guaranteed. Investors assume risk when allocating resources because they have imperfect knowledge of the future. A good investment has expected rewards that outweigh these potential losses. Investors of high-risk projects such as start-ups demand high expected returns. To minimize the risk or achieve the optimum risk-return investment, people seek information that helps them see the future more clearly. The financial market is driven by information.

Financial market innovations often help resources flow to the projects that we think are the most promising and needed. An example related to our daily lives is the invention of the credit card, which allows individuals easy access to capital (with high interest/risk) when

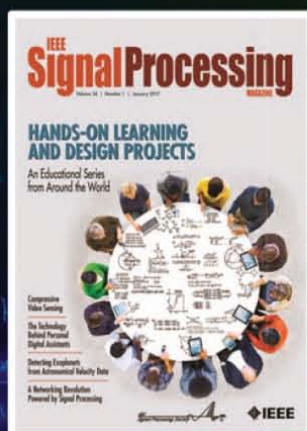
people are in need. Many economists [1] believe that consumers’ continuous spending using the debt accessible through credit cards in economic down times is an important cause for the Great Moderation in the United States in the 1990s. It is easy to take for granted how much technology and the capability to evaluate personal credit through data are a driving force for the popularity of credit cards.

The stock market processes an incalculable amount of data, converting it to information about the future, and boiling everything down into a single number: price. Due to the fluctuations of security prices, many people liken the stock and bond market to one giant casino and attempt to profit by speculating on what the asset prices should be. The best of quantitative hedge funds, such as Renaissance Technologies, D.E. Shaw, Two Sigma, and Citadel, have minted some of the world’s wealthiest people, encouraging those of us who are working in SP and AI to fantasize about our shot at a big win. If we can use our skills to predict tomorrow’s stock price, or even the direction of the market, we could make big money. It can feel as if we are just one DNN away from riches. Indeed, many open-source DNN tools are available for everyone to use in analysis.

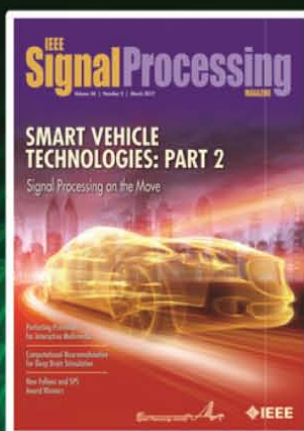
DNN and future AI tools undoubtedly may help us to gain new information,

(continued on page 171)

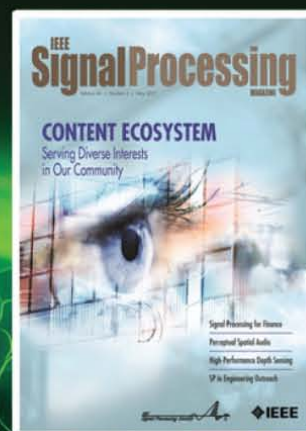
January 2017
Hands-on Learning/Design



March 2017
Smart Vehicle – Part 2



May 2017
Feature Article Collection

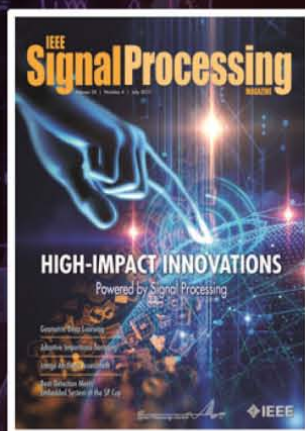


Publish with IEEE Signal Processing Magazine

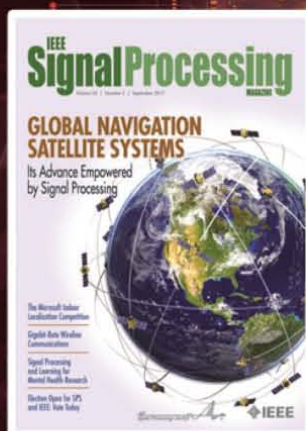
HIGH IMPACT among all electrical engineering publications

REACH a broad signal processing audience worldwide

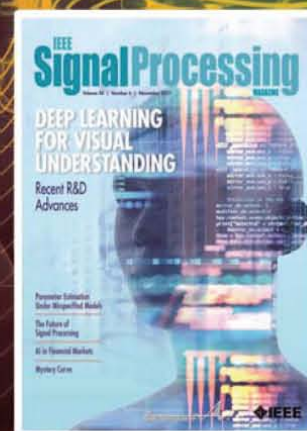
WELCOME proposals for Special Issues and Feature Articles, and contributions to Columns



July 2017
Feature Article Collection



September 2017
Global Navigation Satellite Systems



December 2017
Deep Learning

MATLAB SPEAKS WIRELESS DESIGN

You can simulate, prototype, and verify wireless systems right in MATLAB. Learn how today's MATLAB supports RF, LTE, WLAN and 5G development and SDR hardware.

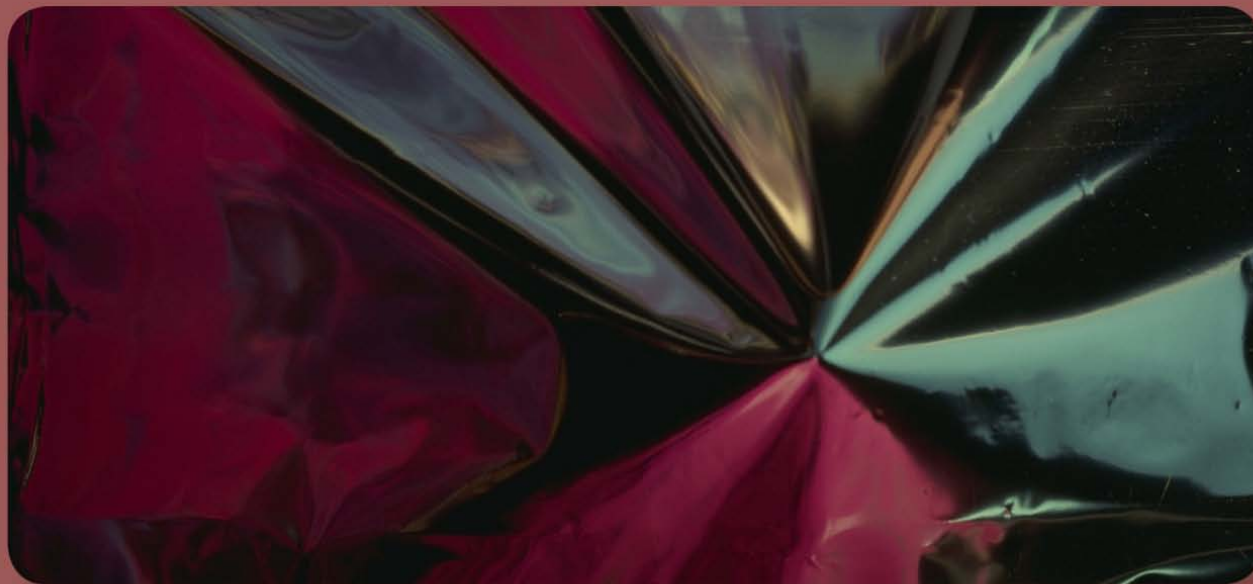
mathworks.com/wireless

IEEE Signal Processing society

Content Gazette

November 2017

ISSN 2167-5023



T-SP November 1 2017 Vol. 65 #20

<http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8071105>

T-ASLP October 2017 Vol. 25 #10

<http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8048690>

T-IP October 2017 Vol. 26 #10

<http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8038880>

IFS November 2017 Vol. 12 #11

<http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8048691>

T-MM November 2017 Vol. 19 #11

<http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8067622>

T-JSTSP September 2017 Vol. 11 #6

<http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8010944>

T-SPL -October 2017 Vol. 24 #10

<http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8049584>

T-SPL November 2017 Vol. 24 #11

<http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8071068>

T-CI September 2017 Vol. 3 #3

<http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8003530>

T-SIPN September 2017 Vol. 3 #3

<http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8010498>

http://www.ieee.org/publications_standards/publications/authors/author_ethics.html



Organizing Committee

General Chairs

Christophoros Nikou, University of Ioannina, Greece

Kostas Plataniotis, University of Toronto, Canada

Technical Program Chairs

Nikolaos Boulgouris, Brunel University London, UK

Lisimachos P. Kondi, University of Ioannina, Greece

Finance Chair

Aggelos Pikrakis, University of Piraeus, Greece

Plenary Chairs

John Apostolopoulos, Cisco Systems, USA

Athanassios Skodras, University of Patras, Greece

Tutorial Chairs

Christine Guillemot, INRIA, Rennes, France

Rafael Molina, University of Granada, Spain

Special Sessions Chairs

Guy Côté, Apple Inc., USA

Adriana Dumitras, USA

Awards Chair

Jean-Philippe Thiran, EPFL, Switzerland

Exhibits/Demo Chair

Adrian Bors, University of York, UK

Publication Chair

Jean-Luc Dugelay, EURECOM, France

Industry Liaison Chairs

Panos Nasiopoulos, University of British Columbia, Canada

Amir Said, Qualcomm, USA

Local Arrangement Chair

Stefanos Kollias, National Technical University of Athens, Greece

Publicity Chairs

Nikos Nikolaidis, Aristotle University of Thessaloniki, Greece

Konstantinos Papadias, Athens Information Technology, Greece

Students/Young Professionals Activities and Doctoral Consortium Chairs

Patrizio Campisi, Università degli Studi Roma Tre, Italy

Sotiris Tsaftaris, University of Edinburgh, UK

Nikolaos Thomos, University of Essex, UK

Registration Chair

Kostas Berberidis, University of Patras, Greece

IEEE Student Activities Chair

Kostas Karpouzis, National Technical University of Athens, Greece

Innovation Program Chairs

Jill Boyce, Intel Corporation

Haohong Wang, TCL Research America

International Liaisons

Panos Papamichalis, Southern Methodist University, USA

Xiao Ping Zhang, Ryerson University, Canada

Yong Man Ro, KAIST, Republic of Korea

Advisory Board

Aggelos Katsaggelos

Petros Maragos

Ioannis Pitas

Sergios Theodoridis

ICIP 2018

<http://2018.ieeeicip.org>

IEEE INTERNATIONAL CONFERENCE ON IMAGE PROCESSING
OCTOBER 7 - 10, 2018 * ATHENS, GREECE



The 25th IEEE International Conference on Image Processing (ICIP) will be held in the Megaron Athens International Conference Centre, Athens, Greece, on October 7-10, 2018. ICIP is the world's largest and most comprehensive technical conference focused on image and video processing and computer vision. The conference will feature world-class speakers, tutorials, exhibits, and a vision technology showcase.

Topics of interest include, but are not limited to:

- Filtering, Transforms, Multi-Resolution Processing
- Restoration, Enhancement, Super-Resolution
- Computer Vision Algorithms and Technologies
- Compression, Transmission, Storage, Retrieval
- Multi-View, Stereoscopic, and 3D Processing
- Multi-Temporal and Spatio-Temporal Processing
- Biometrics, Forensics, and Content Protection
- Biological and Perceptual-based Processing
- Medical Image and Video Analysis
- Document and Synthetic Visual Processing
- Color and Multispectral Processing
- Scanning, Display, and Printing
- Applications to various fields
- Computational Imaging
- Video Processing and Analytics
- Visual Quality Assessment
- Deep learning for Images and Video
- Image and Video Analysis for the Web
- Image Processing for VR Systems
- Image Processing for Autonomous Vehicles

Paper Submission

Authors are invited to submit papers of not more than four pages for technical content including figures and references, with one optional page containing only references. Submission instructions, templates for the required paper format, and information on "no show" policy are available at 2018.ieeeicip.org.

Journal Paper Presentations

Authors of papers published in all IEEE Signal Processing Society fully owned journals as well as in IEEE Wireless Communication Letters will be given the opportunity to present their work at ICIP 2018, subject to space availability and approval by the Technical Program Chairs of IEEE ICIP 2018.

Innovation Program

Following the tradition that started in 2016, the ICIP 2018 Innovation Program Chairs will arrange an outstanding event with prominent speakers from the Industry.

Tutorials, Special Sessions, and Challenge Sessions Proposals

Tutorials will be held on October 7, 2018. Tutorial proposals must include title, outline, contact information, biography and selected publications for the presenter(s), and a description of the tutorial and material to be distributed to participants. For detailed submission guidelines, please refer to the tutorial proposals page. Special Sessions and Challenge Session Proposals must include a topical title, rationale, session outline, contact information, and a list of invited papers/participants. For detailed submission guidelines, please refer the ICIP 2018 website at 2018.ieeeicip.org.

Important Dates

Special Session Proposals:	November 15, 2017
Notification of Special Session Acceptance:	December 15, 2017
Tutorial Proposals:	December 15, 2017
Notification of Tutorial Acceptance:	January 15, 2018
Paper Submission:	February 7, 2018
Notification of Acceptance:	April 30, 2018
Camera-Ready Papers:	May 31, 2018

IEEE TRANSACTIONS ON SIGNAL PROCESSING

A PUBLICATION OF THE IEEE SIGNAL PROCESSING SOCIETY



www.signalprocessingsociety.org

Indexed in PubMed® and MEDLINE®, products of the United States National Library of Medicine



NOVEMBER 1, 2017

VOLUME 65

NUMBER 21

ITPRED

(ISSN 1053-587X)

REGULAR PAPERS

Augmented Nested Arrays With Enhanced DOF and Reduced Mutual Coupling http://dx.doi.org/10.1109/TSP.2017.2736493	5549
On Phase Response Function Based Decentralized Phase Desynchronization http://dx.doi.org/10.1109/TSP.2017.2733452	5564
A Full Mean Square Analysis of CLMS for Second-Order Noncircular Inputs http://dx.doi.org/10.1109/TSP.2017.2739098	5578
Sparsity-Based Two-Dimensional DOA Estimation for Coprime Array: From Sum-Difference Coarray Viewpoint http://dx.doi.org/10.1109/TSP.2017.2739105	5591
Optimal Joint Remote Radio Head Selection and Beamforming Design for Limited Fronthaul C-RAN http://dx.doi.org/10.1109/TSP.2017.2739102	5605
Multuser Detection Based on MAP Estimation With Sum-of-Absolute-Values Relaxation http://dx.doi.org/10.1109/TSP.2017.2740164	5621
Similarity Function Tracking Using Pairwise Comparisons http://dx.doi.org/10.1109/TSP.2017.2739100	5635
Robust Calibration of Radio Interferometers in Non-Gaussian Environment http://dx.doi.org/10.1109/TSP.2017.2733496	5649
Adaptive Diffusion Schemes for Heterogeneous Networks http://dx.doi.org/10.1109/TSP.2017.2740199	5661
Phase Noise Compensation for OFDM Systems http://dx.doi.org/10.1109/TSP.2017.2740165	5675
Working Locally Thinking Globally: Theoretical Guarantees for Convolutional Sparse Coding http://dx.doi.org/10.1109/TSP.2017.2733447	5687



On the Optimal Power Allocation for Two-Way Full-Duplex AF Relay Networks http://dx.doi.org/10.1109/TSP.2017.2736507	5702
..... <i>J.-W. Li and C. Lin</i>	
On Performance of Sparse Fast Fourier Transform and Enhancement Algorithm http://dx.doi.org/10.1109/TSP.2017.2740198	5716
..... <i>G.-L. Chen, S.-H. Tsai, and K.-J. Yang</i>	
Misspecified and Asymptotically Minimax Robust Quickest Change Detection http://dx.doi.org/10.1109/TSP.2017.2740202	5730
..... <i>T. L. Molloy and J. J. Ford</i>	
Component-Based Modeling and Processing of Medical Ultrasound Signals http://dx.doi.org/10.1109/TSP.2017.2731303	5743
..... <i>Y. Yankelevsky, Z. Friedman, and A. Feuer</i>	
Fast and Flexible Successive-Cancellation List Decoders for Polar Codes http://dx.doi.org/10.1109/TSP.2017.2740204	5756
..... <i>S. A. Hashemi, C. Condo, and W. J. Gross</i>	
Tensor-Based Large-Scale Blind System Identification Using Segmentation http://dx.doi.org/10.1109/TSP.2017.2736505	5770
..... <i>M. Boussé, O. Debals, and L. De Lathauwer</i>	
Nonparametric Detection of Anomalous Data Streams http://dx.doi.org/10.1109/TSP.2017.2733472	5785
..... <i>S. Zou, Y. Liang, H. V. Poor, and X. Shi</i>	
Distributed Cophasing With Autonomous Constellation Selection http://dx.doi.org/10.1109/TSP.2017.2739106	5798
..... <i>R. Chopra, R. Annavajjala, and C. R. Murthy</i>	
Large-Scale Multiantenna Multisine Wireless Power Transfer http://dx.doi.org/10.1109/TSP.2017.2739112	5812
..... <i>Y. Huang and B. Clerckx</i>	



19th IEEE INTERNATIONAL WORKSHOP ON SIGNAL PROCESSING ADVANCES IN WIRELESS COMMUNICATIONS
25-28 JUNE, KALAMATA, GREECE

Organizing Committee

General Chairs

Athina Petropulu

Rutgers University, USA

Constantinos Papadias

Athens Information Technology, Greece

Technical Program Chairs

Waheed Bajwa

Rutgers University, USA

Urbashi Mitra

University of Southern California, USA

Special Sessions Chair

Rick Brown

Worcester Polytechnic Institute, USA

Plenaries Chair

Qing Zhao

Cornell University, USA

Tutorials Chair

Ioannis Krikidis

University of Cyprus, Cyprus

Industrial Liaison Chairs

Marios Kountouris

Huawei, France

Yupeng Liu

Nokia, USA

Publications Chair

Angeliki Alexiou

University of Piraeus, Greece

Publicity Chair

Eleftherios Kofidis

University of Piraeus, Greece

SPAWC 2018 will be held in Kalamata, Greece on June 25-28, 2018. The workshop is devoted to advances in signal processing for wireless communications, networking, and information theory. The technical program features tutorials, plenary talks, thematic oral talks, as well as invited and contributed papers presented in poster format.

Call for Papers

Thematically, SPAWC 2018 will in particular focus on the areas of:

- Machine learning and data analytics
- Physical-layer security and privacy
- Biological communications and signal processing
- 5G and beyond

In addition, special session proposals and regular papers are also being solicited in the general areas of:

- Smart antennas, MIMO systems, massive MIMO, and space-time processing
- Reliable wireless communications for autonomous vehicles
- Signal processing for ad-hoc, multi-hop, and sensor networks
- Cooperative communication, coordinated multipoint transmission and reception
- Distributed resource allocation and scheduling
- Convex and non-convex optimization; game theory for communications
- Heterogeneous networks, small cells
- Millimeter wave, 60 GHz communications
- Full duplex systems
- Cognitive radio and networks
- Cooperative sensing, compressed sensing, sparse signal processing
- Machine-to-machine, device-to-device communications
- Modeling, estimation and equalization of wireless channels
- Acquisition, synchronization, localization and tracking
- Low latency & delay-limited communications
- Signal processing for optical, satellite, and underwater communications
- Energy efficiency and energy harvesting

Three best papers, with students as primary authors, will be recognized at the workshop through a student-paper competition. All invited and regular papers, with up to five pages in length, will be published through IEEE Xplore. Papers will be submitted via EDAS.

Important Dates

- Special Session Proposals: Dec-4, 2017
- Decisions Due for Special Sessions: Dec-25, 2017
- Initial Invited and Regular Paper Submission: Feb-19, 2018
- Paper Decisions Due: Apr-30, 2018
- Camera-Ready Papers Due: May-14, 2018
- Registration Deadline for Authors: May-14, 2018
- Workshop Dates: 25-28 June, 2018



IEEE
Signal Processing Society



IEEE TRANSACTIONS ON SIGNAL AND INFORMATION PROCESSING OVER NETWORKS

Now accepting paper submissions

The new *IEEE Transactions on Signal and Information Processing over Networks* publishes high-quality papers that extend the classical notions of processing of signals defined over vector spaces (e.g. time and space) to processing of signals and information (data) defined over networks, potentially dynamically varying. In signal processing over networks, the topology of the network may define structural relationships in the data, or may constrain processing of the data. Topics of interest include, but are not limited to the following:

Adaptation, Detection, Estimation, and Learning

- Distributed detection and estimation
- Distributed adaptation over networks
- Distributed learning over networks
- Distributed target tracking
- Bayesian learning; Bayesian signal processing
- Sequential learning over networks
- Decision making over networks
- Distributed dictionary learning
- Distributed game theoretic strategies
- Distributed information processing
- Graphical and kernel methods
- Consensus over network systems
- Optimization over network systems

Communications, Networking, and Sensing

- Distributed monitoring and sensing
- Signal processing for distributed communications and networking
- Signal processing for cooperative networking
- Signal processing for network security
- Optimal network signal processing and resource allocation

Modeling and Analysis

- Performance and bounds of methods
- Robustness and vulnerability
- Network modeling and identification

Modeling and Analysis (cont.)

- Simulations of networked information processing systems
- Social learning
- Bio-inspired network signal processing
- Epidemics and diffusion in populations

Imaging and Media Applications

- Image and video processing over networks
- Media cloud computing and communication
- Multimedia streaming and transport
- Social media computing and networking
- Signal processing for cyber-physical systems
- Wireless/mobile multimedia

Data Analysis

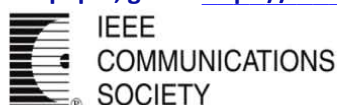
- Processing, analysis, and visualization of big data
- Signal and information processing for crowd computing
- Signal and information processing for the Internet of Things
- Emergence of behavior

Emerging topics and applications

- Emerging topics
- Applications in life sciences, ecology, energy, social networks, economic networks, finance, social sciences, smart grids, wireless health, robotics, transportation, and other areas of science and engineering

Editor-in-Chief: Petar M. Djurić, Stony Brook University (USA)

To submit a paper, go to: <https://mc.manuscriptcentral.com/tsipn-ieee>



IEEE/ACM TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING

A PUBLICATION OF THE IEEE SIGNAL PROCESSING SOCIETY



www.signalprocessingsociety.org

Indexed in PubMed® and MEDLINE®, products of the United States National Library of Medicine



OCTOBER 2017

VOLUME 25

NUMBER 10

ITASFA

(ISSN 2329-9290)

REGULAR PAPERS

An Exemplar-Based Approach to Frequency Warping for Voice Conversion http://dx.doi.org/10.1109/TASLP.2017.2723721	1863
Identifying Missing and Extra Notes in Piano Recordings Using Score-Informed Dictionary Learning http://dx.doi.org/10.1109/TASLP.2017.2724203	1877
Joint Estimation of PLDA and Nonlinear Transformations of Speaker Vectors http://dx.doi.org/10.1109/TASLP.2017.2724198	1890
Multitalker Speech Separation With Utterance-Level Permutation Invariant Training of Deep Recurrent Neural Networks http://dx.doi.org/10.1109/TASLP.2017.2726762	1901
Unsupervised Iterative Deep Learning of Speech Features and Acoustic Tokens with Applications to Spoken Term Detection http://dx.doi.org/10.1109/TASLP.2017.2729024	1914
Room Impulse Response Interpolation Using a Sparse Spatio-Temporal Representation of the Sound Field http://dx.doi.org/10.1109/TASLP.2017.2730284	1929
Deep Feature Engineering for Noise Robust Spoofing Detection http://dx.doi.org/10.1109/TASLP.2017.2732162	1942
Augmented Intensity Vectors for Direction of Arrival Estimation in the Spherical Harmonic Domain http://dx.doi.org/10.1109/TASLP.2017.2736067	1956
Spherical Harmonic Smoothing for Localizing Coherent Sound Sources http://dx.doi.org/10.1109/TASLP.2017.2738698	1969
Intelligibility Enhancement of Telephone Speech Using Gaussian Process Regression for Normal-to-Lombard Spectral Tilt Conversion http://dx.doi.org/10.1109/TASLP.2017.2740004	1985



Multiple-Speaker Localization Based on Direct-Path Features and Likelihood Maximization With Spatial Sparsity Regularization http://dx.doi.org/10.1109/TASLP.2017.2740001	<i>X. Li, L. Girin, R. Horaud, and S. Gannot</i>	1997
Finite Element Synthesis of Diphthongs Using Tuned Two-Dimensional Vocal Tracts http://dx.doi.org/10.1109/TASLP.2017.2735179	<i>M. Arnela and O. Guasch</i>	2013
Joint Denoising and Dereverberation Using Exemplar-Based Sparse Representations and Decaying Norm Constraint http://dx.doi.org/10.1109/TASLP.2017.2744261	<i>D. Baby and H. Van hamme</i>	2024

EDICS—Editor’s Information Classification Scheme http://dx.doi.org/10.1109/TASLP.2017.2754207		2036
Information for Authors http://dx.doi.org/10.1109/TASLP.2017.2725104		2038

IEEE TRANSACTIONS ON IMAGE PROCESSING

A PUBLICATION OF THE IEEE SIGNAL PROCESSING SOCIETY



www.signalprocessingsociety.org

Indexed in PubMed® and MEDLINE®, products of the United States National Library of Medicine



OCTOBER 2017

VOLUME 26

NUMBER 10

IIPRE4

(ISSN 1057-7149)

PAPERS

Derivative Kernels: Numerics and Applications http://dx.doi.org/10.1109/TIP.2017.2713950	<i>M. S. Hosseini and K. N. Plataniotis</i>	4596
Multi-Label Classification by Semi-Supervised Singular Value Decomposition http://dx.doi.org/10.1109/TIP.2017.2719939	<i>L. Jing, C. Shen, L. Yang, J. Yu, and M. K. Ng</i>	4612
Joint Chroma Subsampling and Distortion-Minimization-Based Luma Modification for RGB Color Images With Application http://dx.doi.org/10.1109/TIP.2017.2719945	<i>K.-L. Chung, T.-C. Hsu, and C.-C. Huang</i>	4626
High-Quality Parallel-Ray X-Ray CT Back Projection Using Optimized Interpolation http://dx.doi.org/10.1109/TIP.2017.2706521	<i>M. T. McCann and M. Unser</i>	4639
Action Recognition Using 3D Histograms of Texture and A Multi-Class Boosting Classifier http://dx.doi.org/10.1109/TIP.2017.2718189	<i>B. Zhang, Y. Yang, C. Chen, L. Yang, J. Han, and L. Shao</i>	4648
Robust Face Recognition With Kernelized Locality-Sensitive Group Sparsity Representation http://dx.doi.org/10.1109/TIP.2017.2716180	<i>S. Tan, X. Sun, W. Chan, L. Qu, and L. Shao</i>	4661
Track Everything: Limiting Prior Knowledge in Online Multi-Object Recognition http://dx.doi.org/10.1109/TIP.2017.2696744	<i>S. C. Wong, V. Stamatescu, A. Gatt, D. Kearney, I. Lee, and M. D. McDonnell</i>	4669
Visual Attention Modeling for Stereoscopic Video: A Benchmark and Computational Model http://dx.doi.org/10.1109/TIP.2017.2721112	<i>Y. Fang, C. Zhang, J. Li, J. Lei, M. Perreira Da Silva, and P. Le Callet</i>	4684
Gaussian Process Domain Experts for Modeling of Facial Affect http://dx.doi.org/10.1109/TIP.2017.2721114	<i>S. Eleftheriadis, O. Rudovic, M. P. Deisenroth, and M. Pantic</i>	4697
Diversity-Aware Multi-Video Summarization http://dx.doi.org/10.1109/TIP.2017.2708902	<i>R. Panda, N. C. Mithun, and A. K. Roy-Chowdhury</i>	4712
Large-Scale Crowdsourced Study for Tone-Mapped HDR Pictures http://dx.doi.org/10.1109/TIP.2017.2713945	<i>D. Kundu, D. Ghadiyaram, A. C. Bovik, and B. L. Evans</i>	4725
Low-Rank and Joint Sparse Representations for Multi-Modal Recognition http://dx.doi.org/10.1109/TIP.2017.2721838	<i>H. Zhang, V. M. Patel, and R. Chellappa</i>	4741
Detecting Anatomical Landmarks From Limited Medical Imaging Data Using Two-Stage Task-Oriented Deep Neural Networks http://dx.doi.org/10.1109/TIP.2017.2721106	<i>J. Zhang, M. Liu, and D. Shen</i>	4753
Greedy Batch-Based Minimum-Cost Flows for Tracking Multiple Objects http://dx.doi.org/10.1109/TIP.2017.2723239	<i>X. Wang, B. Fan, S. Chang, Z. Wang, X. Liu, D. Tao, and T. S. Huang</i>	4765
Visual Attention Saccadic Models Learn to Emulate Gaze Patterns From Childhood to Adulthood http://dx.doi.org/10.1109/TIP.2017.2722238	<i>O. Le Meur, A. Coutrot, Z. Liu, P. Rämä, A. Le Roch, and A. Helo</i>	4777

QoE-Guided Warping for Stereoscopic Image Retargeting http://dx.doi.org/10.1109/TIP.2017.2721546	4790
..... F. Shao, W. Lin, W. Lin, Q. Jiang, and G. Jiang	
Part-Based Deep Hashing for Large-Scale Person Re-Identification http://dx.doi.org/10.1109/TIP.2017.2695101	4806
..... F. Zhu, X. Kong, L. Zheng, H. Fu, and Q. Tian	
ESIM: Edge Similarity for Screen Content Image Quality Assessment http://dx.doi.org/10.1109/TIP.2017.2718185	4818
..... Z. Ni, L. Ma, H. Zeng, J. Chen, C. Cai, and K.-K. Ma	
Correlation-Based Tracker-Level Fusion for Robust Visual Tracking http://dx.doi.org/10.1109/TIP.2017.2699791	4832
..... M. K. Rapuru, S. Kakanuru, P. M. Venugopal, D. Mishra, and G. R. K. S. Subrahmanyam	
Going Deeper With Contextual CNN for Hyperspectral Image Classification http://dx.doi.org/10.1109/TIP.2017.2725580	4843
..... H. Lee and H. Kwon	
Fast Segmentation From Blurred Data in 3D Fluorescence Microscopy http://dx.doi.org/10.1109/TIP.2017.2716843	4856
..... M. Storath, D. Rickert, M. Unser, and A. Weinmann	
Robust Web Image Annotation via Exploring Multi-Facet and Structural Knowledge http://dx.doi.org/10.1109/TIP.2017.2717185	4871
..... M. Hu, Y. Yang, F. Shen, L. Zhang, H. T. Shen, and X. Li	
Quality Assessment of Perceptual Crosstalk on Two-View Auto-Stereoscopic Displays http://dx.doi.org/10.1109/TIP.2017.2717180	4885
..... J. Kim, T. Kim, S. Lee, and A. C. Bovik	
Volumetric Image Registration From Invariant Keypoints http://dx.doi.org/10.1109/TIP.2017.2722689	4900
..... B. Rister, M. A. Horowitz, and D. L. Rubin	
Higher Order Energies for Image Segmentation http://dx.doi.org/10.1109/TIP.2017.2722691	4911
..... J. Shen, J. Peng, X. Dong, L. Shao, and F. Porikli	
Blind Deep S3D Image Quality Evaluation via Local to Global Feature Aggregation http://dx.doi.org/10.1109/TIP.2017.2725584	4923
..... H. Oh, S. Ahn, J. Kim, and S. Lee	
Distance Metric Learning via Iterated Support Vector Machines http://dx.doi.org/10.1109/TIP.2017.2725578	4937
..... W. Zuo, F. Wang, D. Zhang, L. Lin, Y. Huang, D. Meng, and L. Zhang	
Inverse Rendering and Relighting From Multiple Color Plus Depth Images http://dx.doi.org/10.1109/TIP.2017.2728184	4951
..... S. Liu and M. N. Do	
Anti-Impulse-Noise Edge Detection via Anisotropic Morphological Directional Derivatives http://dx.doi.org/10.1109/TIP.2017.2726190	4962
..... P.-L. Shui and F.-P. Wang	
Deep Learning on Sparse Manifolds for Faster Object Segmentation http://dx.doi.org/10.1109/TIP.2017.2725582	4978
..... J. C. Nascimento and G. Carneiro	
Temporal Coherence-Based Deblurring Using Non-Uniform Motion Optimization http://dx.doi.org/10.1109/TIP.2017.2731206	4991
..... C. Qiao, R. W. H. Lau, B. Sheng, B. Zhang, and E. Wu	
Texture Characterization Using Shape Co-Occurrence Patterns http://dx.doi.org/10.1109/TIP.2017.2726182	5005
..... G.-S. Xia, G. Liu, X. Bai, and L. Zhang	
Fast and Orthogonal Locality Preserving Projections for Dimensionality Reduction http://dx.doi.org/10.1109/TIP.2017.2726188	5019
..... R. Wang, F. Nie, R. Hong, X. Chang, X. Yang, and W. Yu	
Comments and Corrections to “An Efficient Adaptive Binary Arithmetic Coder Based on Logarithmic Domain” http://dx.doi.org/10.1109/TIP.2017.2729880	5031
..... E. S. Jang and J.-W. Chong	
Learning the Image Processing Pipeline http://dx.doi.org/10.1109/TIP.2017.2713942	5032
..... H. Jiang, Q. Tian, J. Farrell, and B. A. Wandell	
ISAR Imaging of High-Speed Maneuvering Target Using Gapped Stepped-Frequency Waveform and Compressive Sensing http://dx.doi.org/10.1109/TIP.2017.2728182	5043
..... M.-S. Kang, S.-J. Lee, S.-H. Lee, and K.-T. Kim	
Bilinear Optimized Product Quantization for Scalable Visual Content Analysis http://dx.doi.org/10.1109/TIP.2017.2722224	5057
..... L. Yu, Z. Huang, F. Shen, J. Song, H. T. Shen, and X. Zhou	
Ocular Recognition for Blinking Eyes http://dx.doi.org/10.1109/TIP.2017.2713041	5070
..... P. Liu, J.-M. Guo, S.-H. Tseng, K. Wong, J.-D. Lee, C.-C. Yao, and D. Zhu	
EDICS—Editor’s Information Classification Scheme http://dx.doi.org/10.1109/TIP.2017.2740786	5082
Information for Authors http://dx.doi.org/10.1109/TIP.2017.2740787	5083

IEEE Statistical Signal Processing Workshop 2018

10 – 13 June 2018, Freiburg, Germany

www.ssp2018.org



The 2018 IEEE Workshop on Statistical Signal Processing (SSP) will be held from 10-13 June 2018 in Freiburg, Germany. The SSP Workshop is a unique meeting that brings members of the IEEE Signal Processing Society together with researchers from allied fields such as bioinformatics, communications, machine learning, and statistics. One of its key features is having all contributed and special sessions as poster sessions allowing extensive interaction and networking. The scientific program of SSP 2018 will include invited plenary talks, and regular and special sessions with contributed research papers. All submitted papers are reviewed by experts, and all accepted papers will be published on IEEExplore.

General Chair
Peter Schreier
Univ. Paderborn

Technical Co-Chairs
Javier Via
Univ. Cantabria

Arie Yeredor
Tel-Aviv Univ.

Special Sessions
Wing-Kin (Ken) Ma
Chinese Univ. HK

Alle-Jan van der Veen
TU Delft

Finance
Raviv Raich
Oregon State Univ.

Florian Römer
TU Ilmenau

Publications
David Ramirez
Univ. Carlos III Madrid

Local Arrangements
Arno Blau
Stryker Corp

Christian Debes
AGT International

Webmaster
Tim Marrinan
Univ. Paderborn

**Publicity/
International Liaison**
Tülay Adalı
Univ. Maryland BC

Abdelhak Zoubir
TU Darmstadt

We invite submitting original research papers on topics including, but not limited to, the following areas:

Foundations, methods, and algorithms

- Detection and estimation theory
- Machine learning and pattern recognition
- Signal separation methods
- Data driven methods
- Bayesian techniques
- Sampling and reconstruction
- Signal and system modeling
- Adaptive signal processing
- Distributed signal processing
- Signal processing over graphs and networks
- Optimization
- Sparsity-aware processing
- Matrix and tensor methods

Application areas

- Bioinformatics and genomics
- Big data
- Signal processing for the internet of things
- Array processing, radar, and sonar
- Communication systems and networks
- Sensor networks
- Information forensics and security
- Medical and biomedical imaging
- Social networks
- Smart grids and industrial applications
- Geoscience
- Astrophysics
- Financial signal processing

Submission of papers: Prospective authors are invited to submit full papers, with up to four pages of technical content (references may be listed on a fifth page), using the template and formatting guidelines posted at www.ssp2018.org. All accepted papers must be presented at the workshop in order to be included in the proceedings. There will be best student paper awards.

Submission of proposals for special sessions: Special session proposals must include a title, rationale, session outline, list of invited papers, and contact information. Please refer to www.ssp2018.org for further information regarding the submission of proposals.

Important dates:

- | | |
|--|------------------|
| • Submission of proposals for special sessions | 24 November 2017 |
| • Notification of acceptance of special sessions | 8 December 2017 |
| • Submission of full papers | 26 January 2018 |
| • Notification of paper acceptance | 23 March 2018 |
| • Registration and camera-ready papers | 20 April 2018 |

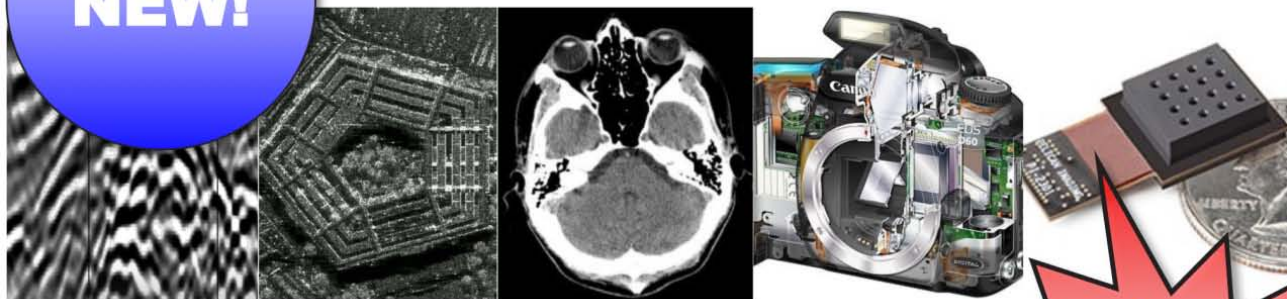
Venue: The conference will be held at the spectacular Historical Merchants' Hall, one of the most outstanding historical buildings in Freiburg dating back to the 14th century. It is situated in the historical center of the city right next to Freiburg cathedral and its main square, which features al fresco dining and boutique shopping.

Freiburg is a famous old university town, known for its high standard of living, its beautiful natural setting, and for being the sunniest and warmest city in Germany. It is located in the heart of the Baden wine-growing region close to the Swiss and French borders and serves as the main entry point to the breathtaking beauty of the Black Forest. Freiburg has a convenient high-speed train connection to Frankfurt International Airport.





IEEE TRANSACTIONS ON COMPUTATIONAL IMAGING



Editor-in-Chief

W. Clem Karl
Boston University

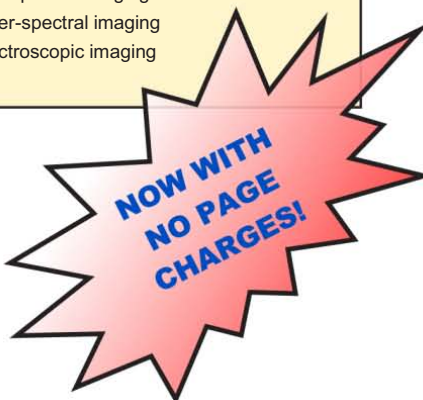
Technical Committee

- Charles Bouman
- Eric Miller
- Peter Corcoran
- Jong Chul Ye
- Dave Brady
- William Freeman

The IEEE Transactions on Computational Imaging publishes research results where computation plays an integral role in the image formation process. All areas of computational imaging are appropriate, ranging from the principles and theory of computational imaging, to modeling paradigms for computational imaging, to image formation methods, to the latest innovative computational imaging system designs. Topics of interest include, but are not limited to the following:

<p>Computational Imaging Methods and Models</p> <ul style="list-style-type: none"> • Coded image sensing • Compressed sensing • Sparse and low-rank models • Learning-based models, dictionary methods • Graphical image models • Perceptual models <p>Computational Image Formation</p> <ul style="list-style-type: none"> • Sparsity-based reconstruction • Statistically-based inversion methods • Multi-image and sensor fusion • Optimization-based methods; proximal iterative methods, ADMM <p>Computational Photography</p> <ul style="list-style-type: none"> • Non-classical image capture • Generalized illumination • Time-of-flight imaging • High dynamic range imaging • Plenoptic imaging 	<p>Computational Consumer Imaging</p> <ul style="list-style-type: none"> • Mobile imaging, cell phone imaging • Camera-array systems • Depth cameras, multi-focus imaging • Pervasive imaging, camera networks <p>Computational Acoustic Imaging</p> <ul style="list-style-type: none"> • Multi-static ultrasound imaging • Photo-acoustic imaging • Acoustic tomography <p>Computational Microscopy</p> <ul style="list-style-type: none"> • Holographic microscopy • Quantitative phase imaging • Multi-illumination microscopy • Lensless microscopy • Light field microscopy <p>Imaging Hardware and Software</p> <ul style="list-style-type: none"> • Embedded computing systems • Big data computational imaging • Integrated hardware/digital design 	<p>Tomographic Imaging</p> <ul style="list-style-type: none"> • X-ray CT • PET • SPECT <p>Magnetic Resonance Imaging</p> <ul style="list-style-type: none"> • Diffusion tensor imaging • Fast acquisition <p>Radar Imaging</p> <ul style="list-style-type: none"> • Synthetic aperture imaging • Inverse synthetic aperture imaging <p>Geophysical Imaging</p> <ul style="list-style-type: none"> • Multi-spectral imaging • Ground penetrating radar • Seismic tomography <p>Multi-spectral Imaging</p> <ul style="list-style-type: none"> • Multi-spectral imaging • Hyper-spectral imaging • Spectroscopic imaging
---	--	---

For more information on the IEEE Transactions on Computational Imaging see <http://www.signalprocessingsociety.org/publications/periodicals/tci/>



IEEE TRANSACTIONS ON INFORMATION FORENSICS AND SECURITY

A PUBLICATION OF THE IEEE SIGNAL PROCESSING SOCIETY



www.signalprocessingsociety.org

NOVEMBER 2017

VOLUME 12

NUMBER 11

ITIFA6

(ISSN 1556-6013)

REGULAR PAPERS

Directional Age-Primitive Pattern (DAPP) for Human Age Group Recognition and Age Estimation http://dx.doi.org/10.1109/TIFS.2017.2695456	<i>M. T. B. Iqbal, M. Shoyaib, B. Ryu, M. Abdullah-Al-Wadud, and O. Chae</i>	2505
ExpSOS: Secure and Verifiable Outsourcing of Exponentiation Operations for Mobile Cloud Computing http://dx.doi.org/10.1109/TIFS.2017.2710941	<i>K. Zhou, M. H. Affi, and J. Ren</i>	2518
A Probabilistic Logic of Cyber Deception http://dx.doi.org/10.1109/TIFS.2017.2710945	<i>S. Jajodia, N. Park, F. Pierazzi, A. Pugliese, E. Serra, G. I. Simari, and V. S. Subrahmanian</i>	2532
Deep Learning Hierarchical Representations for Image Steganalysis http://dx.doi.org/10.1109/TIFS.2017.2710946	<i>J. Ye, J. Ni, and Y. Yi</i>	2545
The Spatial–Temporal Perspective: The Study of the Propagation of Modern Social Worms http://dx.doi.org/10.1109/TIFS.2017.2711424	<i>T. Wang, C. Xia, Z. Li, X. Liu, and Y. Xiang</i>	2558
Someone in Your Contact List: Cued Recall-Based Textual Passwords http://dx.doi.org/10.1109/TIFS.2017.2712126	<i>N. Alomar, M. Alsaleh, and A. Alarifi</i>	2574
Further Improving Efficiency of Higher Order Masking Schemes by Decreasing Randomness Complexity http://dx.doi.org/10.1109/TIFS.2017.2713323	<i>R. Zhang, S. Qiu, and Y. Zhou</i>	2590
Quality-Specific Hand Vein Recognition System http://dx.doi.org/10.1109/TIFS.2017.2713340	<i>J. Wang and G. Wang</i>	2599
Protecting Secret Key Generation Systems Against Jamming: Energy Harvesting and Channel Hopping Approaches http://dx.doi.org/10.1109/TIFS.2017.2713342	<i>E. V. Belmega and A. Chorti</i>	2611
Real-Time Digital Signatures for Time-Critical Networks http://dx.doi.org/10.1109/TIFS.2017.2716911	<i>A. A. Yavuz, A. Mudgerikar, A. Singla, I. Papapanagiotou, and E. Bertino</i>	2627
No Bot Expects the DeepCAPTCHA! Introducing Immutable Adversarial Examples, With Applications to CAPTCHA Generation http://dx.doi.org/10.1109/TIFS.2017.2718479	<i>M. Osadchy, J. Hernandez-Castro, S. Gibson, O. Dunkelman, and D. Pérez-Cabo</i>	2640
A New Rule for Cost Reassignment in Adaptive Steganography http://dx.doi.org/10.1109/TIFS.2017.2718480	<i>W. Zhou, W. Zhang, and N. Yu</i>	2654
Testing the Trustworthiness of IC Testing: An Oracle-Less Attack on IC Camouflaging http://dx.doi.org/10.1109/TIFS.2017.2710954	<i>M. Yasin, O. Sinanoglu, and J. Rajendran</i>	2668
Achieving Perfect Location Privacy in Wireless Devices Using Anonymization http://dx.doi.org/10.1109/TIFS.2017.2713341	<i>Z. Montazeri, A. Houmansadr, and H. Pishro-Nik</i>	2683
Physical Layer Security for Channel-Aware Random Access With Opportunistic Jamming http://dx.doi.org/10.1109/TIFS.2017.2714842 ..	<i>J. Choi</i>	2699
Blind 3D Mesh Watermarking for 3D Printed Model by Analyzing Layering Artifact http://dx.doi.org/10.1109/TIFS.2017.2718482	<i>J.-U. Hou, D.-G. Kim, and H.-K. Lee</i>	2712
Optimal Differentially Private Mechanisms for Randomised Response http://dx.doi.org/10.1109/TIFS.2017.2718487	<i>N. Holohan, D. J. Leith, and O. Mason</i>	2726

Security as a Service for Cloud-Enabled Internet of Controlled Things Under Advanced Persistent Threats: A Contract Design Approach http://dx.doi.org/10.1109/TIFS.2017.2718489	<i>J. Chen and Q. Zhu</i>	2736
Efficient Tensor-Based 2D+3D Face Verification http://dx.doi.org/10.1109/TIFS.2017.2718490	<i>A. Ouamane, A. Chouchane, E. Boutellaa, M. Belahcene, S. Bourennane, and A. Hadid</i>	2751
Privacy-Preserving Similarity Joins Over Encrypted Data http://dx.doi.org/10.1109/TIFS.2017.2721221	<i>X. Yuan, X. Wang, C. Wang, C. Yu, and S. Nutanong</i>	2763
Zipf's Law in Passwords http://dx.doi.org/10.1109/TIFS.2017.2721359	<i>D. Wang, H. Cheng, P. Wang, X. Huang, and G. Jian</i>	2776
Approximating Private Set Union/Intersection Cardinality With Logarithmic Complexity http://dx.doi.org/10.1109/TIFS.2017.2721360	<i>C. Dong and G. Loukides</i>	2792
Near-Optimal and Practical Jamming-Resistant Energy-Efficient Cognitive Radio Communications http://dx.doi.org/10.1109/TIFS.2017.2721931	<i>P. Zhou, Q. Wang, W. Wang, Y. Hu, and D. Wu</i>	2807

IEEE TRANSACTIONS ON **MULTIMEDIA**

A PUBLICATION OF
THE IEEE CIRCUITS AND SYSTEMS SOCIETY
THE IEEE SIGNAL PROCESSING SOCIETY
THE IEEE COMMUNICATIONS SOCIETY
THE IEEE COMPUTER SOCIETY



<http://www.signalprocessingsociety.org/tmm/>

NOVEMBER 2017

VOLUME 19

NUMBER 11

ITREAE

(ISSN 1520-9210)

PAPERS

3-D Video Signal Processing

Texture Plus Depth Video Coding Using Camera Global Motion Information *F. Cheng, T. Tillo, J. Xiao, and B. Jeon* 2361

Compression and Coding

Fast Algorithm and VLSI Architecture of Rate Distortion Optimization in H.265/HEVC *H. Sun, D. Zhou, L. Hu, S. Kimura, and S. Goto* 2375

Nuclear Norm-Based 2DLPP for Image Classification *Y. Lu, C. Yuan, Z. Lai, X. Li, W. K. Wong, and D. Zhang* 2391

Signal Dependent Transform Based on SVD for HEVC Intra-coding *T. Zhang, H. Chen, M.-T. Sun, D. Zhao, and W. Gao* 2404

Image/Video/Graphics Analysis and Synthesis

A Saliency Prior Context Model for Real-Time Object Tracking *C. Ma, Z. Miao, X.-P. Zhang, and M. Li* 2415

An Imbalance Compensation Framework for Background Subtraction *X. Zhang, C. Zhu, H. Wu, Z. Liu, and Y. Xu* 2425

Personalized Social Image Recommendation Method Based on User-Image-Tag Model *J. Zhang, Y. Yang, Q. Tian, L. Zhuo, and X. Liu* 2439

System Design and Optimization

A Pipeline-Based Ray-Tracing Runtime System for HSA-Compliant Frameworks *C.-C. Kao, Y.-T. Miao, and W.-C. Hsu* 2450

3-D Processing and Presentation

Deep Multimetric Learning for Shape-Based 3D Model Retrieval *J. Xie, G. Dai, and Y. Fang* 2463

(Contents Continued on Back Cover)



(Contents Continued from Front Cover)

Subjective and Objective Quality Assessment and User Experience

- Blind Stereo Quality Assessment Based on Learned Features From Binocular Combined Images
 *M. Karimi, M. Nejati, S. M. R. Soroushmehr, S. Samavi, N. Karimi, and K. Najarian* 2475
- Blind Image Quality Assessment Based on Rank-Order Regularized Regression
 *Q. Wu, H. Li, Z. Wang, F. Meng, B. Luo, W. Li, and K. N. Ngan* 2490
- Visual Importance and Distortion Guided Deep Image Quality Assessment Framework
 *J. Guan, S. Yi, X. Zeng, W.-K. Cham, and X. Wang* 2505

Big Data Support for Multimedia

- Compact Hash Codes for Efficient Visual Descriptors Retrieval in Large Scale Databases
 *S. Ercoli, M. Bertini, and A. Del Bimbo* 2521

Multimedia Search and Retrieval

- Trip Outfits Advisor: Location-Oriented Clothing Recommendation
 *X. Zhang, J. Jia, K. Gao, Y. Zhang, D. Zhang, J. Li, and Q. Tian* 2533
- Learning Efficient Binary Codes From High-Level Feature Representations for Multilabel Image Retrieval
 *L. Ma, H. Li, F. Meng, Q. Wu, and K. N. Ngan* 2545

Social and Web Multimedia

- Predicting Popularity of Online Videos Using Support Vector Regression
 *T. Trzciński and P. Rokita* 2561

Media Cloud Computing and Communications

- Social-Aware Rate Based Content Sharing Mode Selection for D2D Content Sharing Scenarios
 *D. Wu, L. Zhou, and Y. Cai* 2571

Multimedia Content Delivery Networks

- Tradeoffs Between Cost and Performance for CDN Provisioning Based on Coordinate Transformation
 *H. Yin, X. Zhang, S. Zhao, Y. Luo, C. Tian, and V. Sekar* 2583
- QoS Provisionings for Device-to-Device Content Delivery in Cellular Networks
 *Y. Xu and F. Liu* 2597

Multimedia Storytelling and Cross-Modal Translations Between Multimedia Contents

- A Probabilistic Approach to People-Centric Photo Selection and Sequencing
 *V. Vonikakis, R. Subramanian, J. Arnfred, and S. Winkler* 2609

CORRESPONDENCE

Error Resilience and Concealment

- Improved Depth-Assisted Error Concealment Algorithm for 3D Video Transmission
 *P.-C. Huang, J.-R. Lin, G.-L. Li, K.-H. Tai, and M.-J. Chen* 2625

ANNOUNCEMENTS

- Introducing IEEE Collabratec 2633
- MGM Program 2634

-
- Information for Authors 2635
-



General chairs

Yiannis Kompatsiaris,
CERTH-ITI (GR)

Thrasyvoulos N. Pappas,
Northwestern University (US)

Technical program co-chairs

Vasileios Mezaris, CERTH-ITI (GR)
Alessandro Foi, Tampere University
of Technology (FIN)
Brendt Wohlberg, Los Alamos
National Laboratory (US)

Financial chair

Anastasios Karakostas,
CERTH-ITI (GR)

Publication co-chairs

Ioannis Patras, Queen Mary
University of London (UK)
Giulia Boato, University of Trento
(IT)

Publicity

Jenny Benois Pineau, University of
Bordeaux (FR)

US Liaison

Charlie Bouman, Purdue University
(US)

Asia Liaison

Alex Kot (SING)

Local arrangement co-chairs

Anastasios Karakostas,
CERTH-ITI (GR)
Sofia Tsekeridou, INTRASOFT
International R&D (GR)

The 2018 IEEE Image, Video, and Multidimensional Signal Processing (IVMSP) Workshop is the 13th of a series of unique meetings that bring together researchers in academia and industry to share the most recent and exciting advances in image, video, and multidimensional signal processing and analysis. The main themes of the 2018 IVMSP Workshop are **Big Data, Social Media and Computational Imaging**.

The scientific program of IVMSP 2018 will include plenary talks, regular and special sessions. We welcome contributions within the three main themes, as well as at their intersection, e.g. use of imagery from social networks for learning models for image reconstruction tasks.

Big Data

- Deep learning methods for large scale multimedia
- Multimedia big data processing, analysis and retrieval
- Distributed solutions and architectures for big multimedia
- Convergence between Internet of Things, wearables and social media
- Big Data applications (summarization, surveillance, mobile, etc)

Social Media

- Social media data collection, filtering, and indexing
- Social media data representation and understanding
- User profiling, collective behavior and privacy aspects of social media
- Monitoring, sensing and prediction in social signals
- Detection, analysis and verification of emergent events

Computational Imaging

- Image models, sparse, low rank, and statistical models
- Image formation, model based inversion, image fusion, and optimization-based methods
- Computational imaging systems, computational photography and microscopy, medical and radar imaging
- Hardware and software for computational imaging, embedded systems, big data, and non-traditional sensors

PAPER SUBMISSION

Papers cannot be longer than 5 pages (double-column IEEE conference format), including all text, figures, tables, references, etc. The 5th page is available for references only. See the website for additional information regarding the submission process: www.ivmsp2018.org.

BEST STUDENT PAPER AWARDS

The IVMSP Best Student Paper Awards will be granted to the first, second and third best overall papers for which a student is the principal author and presenter. The selection will be based on the technical quality, originality, and clarity of the submission.

IMPORTANT DATES

- | | |
|---------------------------------|------------------|
| • Submission of full papers | January 30, 2018 |
| • Notification of acceptance | March 28, 2018 |
| • Author advance registration | April 26, 2018 |
| • Camera-ready paper submission | April 26, 2018 |

IEEE JOURNAL OF SELECTED TOPICS IN SIGNAL PROCESSING


www.ieee.org/sp/index.html

SEPTEMBER 2017

VOLUME 11

NUMBER 6

IJSTGY

(ISSN 1932-4553)

 EDITORIAL

Introduction to the Cooperative Special Issue on Graph Signal Processing in the IEEE Transactions on Signal and Information Processing over Networks and IEEE Journal of Selected Topics in Signal Processing http://dx.doi.org/10.1109/IJSTSP.2017.2733938	<i>P. Frossard, P. L. Dragotti, A. Ortega, M. Rabbat, and A. Ribeiro</i>	771
--	--	-----

 PAPERS

Graph Filters and the Z-Laplacian http://dx.doi.org/10.1109/IJSTSP.2017.2730040	<i>X. Yan, B. M. Sadler, R. J. Drost, P. L. Yu, and K. Lerman</i>	774
Spectral Projector-Based Graph Fourier Transforms http://dx.doi.org/10.1109/IJSTSP.2017.2731599	<i>J. A. Deri and J. M. F. Moura</i>	785
On the Graph Fourier Transform for Directed Graphs http://dx.doi.org/10.1109/IJSTSP.2017.2726979	<i>S. Sardellitti, S. Barbarossa, and P. Di Lorenzo</i>	796
Almost Tight Spectral Graph Wavelets With Polynomial Filters http://dx.doi.org/10.1109/IJSTSP.2017.2726972	<i>D. B. H. Tay, Y. Tanaka, and A. Sakiyama</i>	812
Graph Learning From Data Under Laplacian and Structural Constraints http://dx.doi.org/10.1109/IJSTSP.2017.2726975	<i>H. E. Egilmez, E. Pavez, and A. Ortega</i>	825
Graph Signal Recovery via Primal-Dual Algorithms for Total Variation Minimization http://dx.doi.org/10.1109/IJSTSP.2017.2726978	<i>P. Berger, G. Hannak, and G. Matz</i>	842
Kernel-Based Reconstruction of Space-Time Functions on Dynamic Graphs http://dx.doi.org/10.1109/IJSTSP.2017.2726976	<i>D. Romero, V. N. Ioannidis, and G. B. Giannakis</i>	856
Time-Varying Graph Signal Reconstruction http://dx.doi.org/10.1109/IJSTSP.2017.2726969	<i>K. Qiu, X. Mao, X. Shen, X. Wang, T. Li, and Y. Gu</i>	870
Robust Spatial Filtering With Graph Convolutional Neural Networks http://dx.doi.org/10.1109/IJSTSP.2017.2726981	<i>F. P. Such, S. Sah, M. A. Dominguez, S. Pillai, C. Zhang, A. Michael, N. D. Cahill, and R. Ptucha</i>	884
Nonmonotonic Front Propagation on Weighted Graphs With Applications in Image Processing and High-Dimensional Data Classification http://dx.doi.org/10.1109/IJSTSP.2017.2731520	<i>X. Desquesnes and A. Elmoataz</i>	897
Query Adaptive Fusion for Graph-Based Visual Reranking http://dx.doi.org/10.1109/IJSTSP.2017.2726977	<i>M. Fang and Y.-J. Zhang</i>	908
Information for Authors http://dx.doi.org/10.1109/IJSTSP.2017.2736321		918



IEEE Journal on Selected Topics in Signal Processing

Call for Papers

Special Issue on

“Information-Theoretic Methods in Data Acquisition, Analysis, and Processing”

The field of information theory addresses fundamental questions in various areas including statistical decision theory, data communications, data compression, security, and networking. In particular, information-theoretic methods can be used to illuminate fundamental limits and gauge the effectiveness of algorithms for various problems associated with these fields.

Recent years have witnessed a renaissance in the use of information-theoretic methods to address various problems in the general field of information processing beyond communications and networking, including signal acquisition, signal analysis and processing, compressive sensing, dictionary learning, supervised and unsupervised learning, reinforcement learning, graph mining, and more.

With a world-wide drive in both academia and industry for new approaches to data science, it is generally believed that information-theoretic methods have the potential to illuminate theory and algorithms that will underpin this emerging field.

This special issue covers emerging topics at the interface of information theory and data acquisition, analysis, and processing, with applications to the general area of data science. Its overarching aim is to map out this emerging research landscape as well as current and future research directions.

Topics of interest include (but are not limited to):

- New information measures to capture limits in modern data acquisition, analysis, and processing problems
- Information-theoretic limits on and algorithms for data acquisition and processing
- Limits on and algorithms for feature extraction, data sketching, and information embedding
- Limits on and algorithms for community detection, graph selection, and ranking
- Limits in active learning, supervised and unsupervised learning, reinforcement learning, and deep learning
- Limits on and algorithms for data acquisition, analysis, and processing problems in the presence of communication and / or computation constraints
- New approaches from the fields of approximation theory and harmonic analysis to unveil limits on and algorithms for data acquisition, analysis, and processing
- Application of new techniques to problems in signal processing, imaging, decision theory, machine learning, data analysis, security, and privacy.

Prospective authors should follow the instructions given on the IEEE JSTSP webpage: <https://signalprocessingsociety.org/publications-resources/ieee-journal-selected-topics-signal-processing>, and submit their manuscript through the web submission system at: <https://mc.manuscriptcentral.com/jstsp-ieee>.

Important Dates:

- Manuscript submission: December 1, 2017
- 1st review completed: February 1, 2018
- Revised manuscript due: April 1, 2018
- 2nd review completed: June 1, 2018
- Final manuscript due: July 1, 2018
- Publication: October 2018

Guest Editors:

- Helmut Bölcskei, ETH Zurich
- Stark Draper, University of Toronto
- Yonina Eldar, Technion – Israel Institute of Technology
- Miguel Rodrigues, University College London
- Vincent Tan, National University of Singapore

IEEE

SIGNAL PROCESSING LETTERS

A PUBLICATION OF THE IEEE SIGNAL PROCESSING SOCIETY


www.ieee.org/sp/index.html

OCTOBER 2017

VOLUME 24

NUMBER 10

ISPLEM

(ISSN 1070-9908)

LETTERS

Max–Min Multicell-Aware Precoding and Power Allocation for Downlink Massive MIMO Systems http://dx.doi.org/10.1109/LSP.2017.2715501	<i>S. Zarei, J. Aulin, W. Gerstacker, and R. Schober</i>	1433
Reflection Symmetry Axes Detection Using Multiple Model Fitting http://dx.doi.org/10.1109/LSP.2017.2735630	<i>R. Nagar and S. Raman</i>	1438
A Novel Iterative Shrinkage Algorithm for CS-MRI via Adaptive Regularization http://dx.doi.org/10.1109/LSP.2017.2736159	<i>Z. Chen, Y. Fu, Y. Xiang, and R. Rong</i>	1443
Real-Time Perceptual Model for Distraction in Interfering Audio-on-Audio Scenarios http://dx.doi.org/10.1109/LSP.2017.2733084	<i>J. Rämö, S. Bech, and S. H. Jensen</i>	1448
A Scalable ADMM Algorithm for Rigid Registration http://dx.doi.org/10.1109/LSP.2017.2737518	<i>R. Sanyal, S. M. Ahmed, M. Jaiswal, and K. N. Chaudhury</i>	1453
Extended Locality-Constrained Linear Self-Coding for Saliency Detection http://dx.doi.org/10.1109/LSP.2017.2737650	<i>C. Yang, J. Pu, G.-S. Xie, Y. Dong, and Z. Liu</i>	1458
Detecting the Presence of ENF Signal in Digital Videos: A Superpixel-Based Approach http://dx.doi.org/10.1109/LSP.2017.2741440	<i>S. Vatansever, A. E. Dirik, and N. Memon</i>	1463
Sparse Overcomplete Denoising: Aggregation Versus Global Optimization http://dx.doi.org/10.1109/LSP.2017.2734119	<i>D. Carrera, G. Boracchi, A. Foi, and B. Wohlberg</i>	1468
Weak RIC Analysis of Finite Gaussian Matrices for Joint Sparse Recovery http://dx.doi.org/10.1109/LSP.2017.2729022	<i>A. Elzanaty, A. Giorgetti, and M. Chiani</i>	1473
Two-Stream Deep Correlation Network for Frontal Face Recovery http://dx.doi.org/10.1109/LSP.2017.2736542	<i>T. Zhang, Q. Dong, M. Tang, and Z. Hu</i>	1478
Order-Based Disparity Refinement Including Occlusion Handling for Stereo Matching http://dx.doi.org/10.1109/LSP.2017.2739150	<i>X. Ye, Y. Gu, L. Chen, J. Li, H. Wang, and X. Zhang</i>	1483
Chirp Spread Spectrum Toward the Nyquist Signaling Rate—Orthogonality Condition and Applications http://dx.doi.org/10.1109/LSP.2017.2737596	<i>X. Ouyang, O. A. Dobre, Y. L. Guan, and J. Zhao</i>	1488
Gaze-Based Object Segmentation http://dx.doi.org/10.1109/LSP.2017.2739200	<i>R. Shi, N. K. Ngan, and H. Li</i>	1493
Flexible Multiple Base Station Association and Activation for Downlink Heterogeneous Networks http://dx.doi.org/10.1109/LSP.2017.2738027	<i>K. Shen, Y.-F. Liu, D. Y. Ding, and W. Yu</i>	1498
An Analytical Study of Circularly Pulse-Shaped FBMC-OQAM Waveforms http://dx.doi.org/10.1109/LSP.2017.2738620	<i>A. RezaazadehReyhani and B. Farhang-Boroujeni</i>	1503
A Parallel Approach to HRTF Approximation and Interpolation Based on a Parametric Filter Model http://dx.doi.org/10.1109/LSP.2017.2741724	<i>G. Ramos, M. Cobos, B. Bank, and J. A. Belloch</i>	1507
Second-Order Statistics of One-Sided CPM Signals http://dx.doi.org/10.1109/LSP.2017.2740964	<i>D. Darsena, G. Gelli, I. Iudice, and F. Verde</i>	1512

Shape Projectors for Landmark-Based Spline Curves http://dx.doi.org/10.1109/LSP.2017.2743692	<i>D. Schmitter and M. Unser</i>	1517
Direct Derivation of the Stochastic CRB of DOA Estimation for Rectilinear Sources http://dx.doi.org/10.1109/LSP.2017.2744673	<i>H. Abeida and J. P. Delmas</i>	1522
Distributed Alamouti Relay Beamforming Scheme in Multiuser Relay Networks http://dx.doi.org/10.1109/LSP.2017.2735806	<i>W. Li and M. Dong</i>	1527
Distributed Estimation Recovery Under Sensor Failure http://dx.doi.org/10.1109/LSP.2017.2749265	<i>M. Doostmohammadian, H. R. Rabiee, H. Zarrabi, and U. A. Khan</i>	1532
Multi-View Nonparametric Discriminant Analysis for Image Retrieval and Recognition http://dx.doi.org/10.1109/LSP.2017.2748392	<i>G. Cao, A. Iosifidis, and M. Gabbouj</i>	1537
An Efficient Manifold Algorithm for Constructive Interference Based Constant Envelope Precoding http://dx.doi.org/10.1109/LSP.2017.2748230	<i>F. Liu, C. Masouros, P. V. Amadori, and H. Sun</i>	1542
Automatic Steganographic Distortion Learning Using a Generative Adversarial Network http://dx.doi.org/10.1109/LSP.2017.2745572	<i>W. Tang, S. Tan, B. Li, and J. Huang</i>	1547
Making Likelihood Ratios Digestible for Cross-Application Performance Assessment http://dx.doi.org/10.1109/LSP.2017.2748899	<i>A. Nausch, D. Meuwly, D. Ramos, J. Lindh, and C. Busch</i>	1552
Joint Transceiver Optimization of MIMO SWIPT Systems for Harvested Power Maximization http://dx.doi.org/10.1109/LSP.2017.2749405	<i>Z. Chen, Q. Shi, Q. Wu, and W. Xu</i>	1557
Deep Ensemble Tracking http://dx.doi.org/10.1109/LSP.2017.2749458	<i>J. Guo and T. Xu</i>	1562
Relationships Between Nonlinear and Space-Variant Linear Models in Hyperspectral Image Unmixing http://dx.doi.org/10.1109/LSP.2017.2747478	<i>L. Drumetz, B. Ehsandoust, J. Chanussot, B. Rivet, M. Babaie-Zadeh, and C. Jutten</i>	1567
Contravariant Adaptation on the Manifold of Causal, FIR, Invertible Multivariable Matrix Systems http://dx.doi.org/10.1109/LSP.2017.2749483	<i>T. K. Moon and J. H. Gunther</i>	1572

IEEE

SIGNAL PROCESSING LETTERS

A PUBLICATION OF THE IEEE SIGNAL PROCESSING SOCIETY


www.ieee.org/sp/index.html

NOVEMBER 2017

VOLUME 24

NUMBER 11

ISPLEM

(ISSN 1070-9908)

LETTERS

Stacking <i>PCANet+</i> : An Overly Simplified ConvNets Baseline for Face Recognition http://dx.doi.org/10.1109/LSP.2017.2749763	1581
..... <i>C.-Y. Low, A. B.-J. Teoh, and K.-A. Toh</i>	
Forensic Face Photo-Sketch Recognition Using a Deep Learning-Based Architecture http://dx.doi.org/10.1109/LSP.2017.2749266	1586
..... <i>C. Galea and R. A. Farrugia</i>	
Generalized Rational Sampling Rate Conversion Polyphase FIR Filter http://dx.doi.org/10.1109/LSP.2017.2750904	1591
..... <i>A. Kumar, S. Yadav, and N. Purohit</i>	
An Adaptive Computation of Contour Representations for Mode Decomposition http://dx.doi.org/10.1109/LSP.2017.2750802	1596
..... <i>D.-H. Pham and S. Meignen</i>	
An Individualized Super-Gaussian Single Microphone Speech Enhancement for Hearing Aid Users With Smartphone as an Assistive Device http://dx.doi.org/10.1109/LSP.2017.2750979	1601
..... <i>C. K. A. Reddy, N. Shankar, G. S. Bhat, R. Charan, and I. Panahi</i>	
Determining the Dimension of the Improper Signal Subspace in Complex-Valued Data http://dx.doi.org/10.1109/LSP.2017.2751959	1606
..... <i>T. Hasija, C. Lameiro, and P. J. Schreiber</i>	
Optimal Parameter Encoding Based on Worst Case Fisher Information Under a Secrecy Constraint http://dx.doi.org/10.1109/LSP.2017.2749517	1611
..... <i>Ç. Göken and S. Gezici</i>	
On the Spark of Binary LDPC Measurement Matrices From Complete Protographs http://dx.doi.org/10.1109/LSP.2017.2749043	1616
..... <i>H. Liu, H. Zhang, and L. Ma</i>	
Non-orthogonal Simultaneous Diagonalization of K-Order Complex Tensors for Source Separation http://dx.doi.org/10.1109/LSP.2017.2751038	1621
..... <i>V. Maurandi and E. Moreau</i>	
Automatic Modulation Classification Using Deep Learning Based on Sparse Autoencoders With Nonnegativity Constraints http://dx.doi.org/10.1109/LSP.2017.2752459	1626
..... <i>A. Ali and F. Yangyu</i>	
A Consensus Nonlinear Filter With Measurement Uncertainty in Distributed Sensor Networks http://dx.doi.org/10.1109/LSP.2017.2751611	1631
..... <i>K. Shen, Z. Jing, and P. Dong</i>	
Distributed Learning With Time Correlated Information http://dx.doi.org/10.1109/LSP.2017.2751086	1636
..... <i>P. Guerreiro and J. Xavier</i>	
Simultaneous Sparse Bayesian Learning With Partially Shared Supports http://dx.doi.org/10.1109/LSP.2017.2753770	1641
..... <i>W. Chen</i>	
On Probability of Support Recovery for Orthogonal Matching Pursuit Using Mutual Coherence http://dx.doi.org/10.1109/LSP.2017.2753939	1646
..... <i>E. Miandji, M. Emadi, J. Unger, and E. Afshari</i>	
CorrC2G: Color to Gray Conversion by Correlation http://dx.doi.org/10.1109/LSP.2017.2755077	1651
..... <i>H. Z. Nafchi, A. Shahkolaei, R. Hedjam, and M. Cheriet</i>	
No-Reference Image Quality Assessment Using Image Statistics and Robust Feature Descriptors http://dx.doi.org/10.1109/LSP.2017.2754539	1656
..... <i>M. Oszust</i>	
Multimodal Image Registration Through Simultaneous Segmentation http://dx.doi.org/10.1109/LSP.2017.2754263	1661
..... <i>I. Aganj and B. Fischl</i>	
Joint Human Detection and Head Pose Estimation via Multistream Networks for RGB-D Videos http://dx.doi.org/10.1109/LSP.2017.2731952	1666
..... <i>G. Zhang, J. Liu, H. Li, Y. Q. Chen, and L. S. Davis</i>	

Evaluating Multiexposure Fusion Using Image Information http://dx.doi.org/10.1109/LSP.2017.2752233	1671
Does Vector Gaussian Approximation After LMMSE Filtering Improve the LLR Quality? http://dx.doi.org/10.1109/LSP.2017.2751570	1676
Generalized Least Squares for ESPRIT-Type Direction of Arrival Estimation http://dx.doi.org/10.1109/LSP.2017.2751303	1681
Nonconvex Weighted ℓ_p Minimization Based Group Sparse Representation Framework for Image Denoising http://dx.doi.org/10.1109/LSP.2017.2731791	1686
A Fiber Bundle Geometry Approach for Edge Detection of Chromaticity Distributions http://dx.doi.org/10.1109/LSP.2017.2723426	1691
Optimal Bandwidth for Multitaper Spectrum Estimation http://dx.doi.org/10.1109/LSP.2017.2719943	1696
A Multiple Image-Based Noise Level Estimation Algorithm http://dx.doi.org/10.1109/LSP.2017.2755687	1701
A New Optimality Property of the Capon Estimator http://dx.doi.org/10.1109/LSP.2017.2729658	1706
Iterative Target Localization in Distributed MIMO Radar From Bistatic Range Measurements http://dx.doi.org/10.1109/LSP.2017.2747479	1709
Proactive Monitoring via Jamming in Amplify-and-Forward Relay Networks http://dx.doi.org/10.1109/LSP.2017.2727045	1714
Mixture Reduction on Matrix Lie Groups http://dx.doi.org/10.1109/LSP.2017.2723765	1719
Chirp Rate and Instantaneous Frequency Estimation: Application to Recursive Vertical Synchrosqueezing http://dx.doi.org/10.1109/LSP.2017.2714578	1724
Artificial Noise Aided Hybrid Precoding Design for Secure mmWave MISO Systems With Partial Channel Knowledge http://dx.doi.org/10.1109/LSP.2017.2751459	1729
A Close to Optimal Adaptive Filter for Sudden System Changes http://dx.doi.org/10.1109/LSP.2017.2757147	1734
Frequency-Domain Intergroup Interference Coordination for V2V Communications http://dx.doi.org/10.1109/LSP.2017.2757529	1739
Adaptive Detection and Range Estimation of Point-Like Targets With Symmetric Spectrum http://dx.doi.org/10.1109/LSP.2017.2756076	1744
Effect of Prosody Modification on Children's ASR http://dx.doi.org/10.1109/LSP.2017.2756347	1749

IEEE TRANSACTIONS ON COMPUTATIONAL IMAGING

A PUBLICATION OF
IEEE SIGNAL PROCESSING SOCIETY
IEEE ENGINEERING IN MEDICINE AND BIOLOGY SOCIETY
IEEE CONSUMER ELECTRONICS SOCIETY



TECHNICALLY CO-SPONSORED BY
IEEE GEOSCIENCE AND REMOTE SENSING SOCIETY



SEPTEMBER 2017

VOLUME 3

NUMBER 3

ITCIAJ

(ISSN 2333-9403)

GUEST EDITORIAL

Guest Editorial Special Issue on Extreme Imaging http://dx.doi.org/10.1109/TCL.2017.2726838	382
. <i>W. T. Freeman, A. Savakis, Y. Schechner, N. Snavely, and W. Heidrich</i>	

SPECIAL ISSUE ON EXTREME IMAGING

FlatCam: Thin, Lensless Cameras Using Coded Aperture and Computation http://dx.doi.org/10.1109/TCL.2016.2593662	384
. <i>M. S. Asif, A. Ayremlou, A. Sankaranarayanan, A. Veeraraghavan, and R. G. Baraniuk</i>	
Lensless Imaging With Compressive Ultrafast Sensing http://dx.doi.org/10.1109/TCL.2017.2684624	398
. <i>G. Satat, M. Tancik, and R. Raskar</i>	
Compressive Spectral Imaging via Polar Coded Aperture http://dx.doi.org/10.1109/TCL.2016.2617740	408
. <i>C. Fu, M. L. Don, and G. R. Arce</i>	
A Century of Portraits: A Visual Historical Record of American High School Yearbooks http://dx.doi.org/10.1109/TCL.2017.2699865	421
. <i>S. Ginosar, K. Rakelly, S. M. Sachs, B. Yin, C. Lee, P. Krähenbühl, and A. A. Efros</i>	
Spatial-Spectral Representation for X-Ray Fluorescence Image Super-Resolution http://dx.doi.org/10.1109/TCL.2017.2703987	432
. <i>Q. Dai, E. Pouyet, O. Cossairt, M. Walton, and A. K. Katsaggelos</i>	
A Few Photons Among Many: Unmixing Signal and Noise for Photon-Efficient Active Imaging http://dx.doi.org/10.1109/TCL.2017.2706028	445
. <i>J. Rapp and V. K. Goyal</i>	



A Bayesian Approach to Denoising of Single-Photon Binary Images http://dx.doi.org/10.1109/TCL.2017.2703900	460
..... <i>Y. Altmann, R. Aspden, M. Padgett, and S. McLaughlin</i>	
Object Depth Profile and Reflectivity Restoration From Sparse Single-Photon Data Acquired in Underwater Environments http://dx.doi.org/10.1109/TCL.2017.2669867	472
..... <i>A. Halimi, A. Maccarone, A. McCarthy, S. McLaughlin, and G. S. Buller</i>	
Acoustic Imaging of In-Duct Aeroengine Noise Sources Using Rotating Beamforming and Phased Arrays http://dx.doi.org/10.1109/TCL.2017.2721744	485
..... <i>L. C. Caldas, P. C. Greco, C. C. Pagani, and L. A. Baccalá</i>	
Model-Based Multiscale Gigapixel Image Formation Pipeline on GPU http://dx.doi.org/10.1109/TCL.2016.2612942	493
..... <i>Q. Gong, E. Vera, D. R. Golish, S. D. Feller, D. J. Brady, and M. E. Gehm</i>	

EDICS—Editor’s Classification Information Scheme http://dx.doi.org/10.1109/TCL.2017.2732639	503
Information for Authors http://dx.doi.org/10.1109/TCL.2017.2726864	504

IEEE TRANSACTIONS ON SIGNAL AND INFORMATION PROCESSING OVER NETWORKS

A PUBLICATION OF
THE IEEE SIGNAL PROCESSING SOCIETY
THE IEEE COMMUNICATIONS SOCIETY
THE IEEE COMPUTER SOCIETY



SEPTEMBER 2017

VOLUME 3

NUMBER 3

ITSIBW

(ISSN 2373-776X)

PREPARED COOPERATIVELY WITH THE IEEE JOURNAL OF SELECTED TOPICS IN SIGNAL PROCESSING
SPECIAL ISSUE ON GRAPH SIGNAL PROCESSING

EDITORIAL

Introduction to the Cooperative Special Issue on Graph Signal Processing in the IEEE Journal of Selected Topics in Signal Processing and the IEEE Transactions on Signal and Information Processing Over Networks
<http://dx.doi.org/10.1109/TSIPN.2017.2734178> P. Frossard, P. L. Dragotti, A. Ortega, M. Rabbat, and A. Ribeiro 448

SPECIAL ISSUE PAPERS

Graph Sampling for Covariance Estimation <http://dx.doi.org/10.1109/TSIPN.2017.2731161> S. P. Chepuri and G. Leus 451

Network Topology Inference from Spectral Templates <http://dx.doi.org/10.1109/TSIPN.2017.2731051> S. Segarra, A. G. Marques, G. Mateos, and A. Ribeiro 467

Learning Heat Diffusion Graphs <http://dx.doi.org/10.1109/TSIPN.2017.2731164> D. Thanou, X. Dong, D. Kressner, and P. Frossard 484

Online Bayesian Inference of Diffusion Networks <http://dx.doi.org/10.1109/TSIPN.2017.2731160> S. Shaghaghian and M. Coates 500

Inferring Structural Characteristics of Networks With Strong and Weak Ties From Fixed-Choice Surveys
<http://dx.doi.org/10.1109/TSIPN.2017.2731053> N. Momeni and M. G. Rabbat 513

Quickest Search for Local Structures in Random Graphs <http://dx.doi.org/10.1109/TSIPN.2017.2731125> J. Heydari and A. Tajer 526

Node Embedding via Word Embedding for Network Community Discovery <http://dx.doi.org/10.1109/TSIPN.2017.2731163> W. Ding, C. Lin, and P. Ishwar 539

Multilayer Spectral Graph Clustering via Convex Layer Aggregation: Theory and Algorithms
<http://dx.doi.org/10.1109/TSIPN.2017.2731123> P.-Y. Chen and A. O. Hero 553

REGULAR PAPERS

Emerging Topics and Applications

Decentralized Dynamic Optimization for Power Network Voltage Control <http://dx.doi.org/10.1109/TSIPN.2016.2631886> H. J. Liu, W. Shi, and H. Zhu 568



Adaptation, Detection, Estimation, and Learning

Convergence of Distributed Flooding and Its Application for Distributed Bayesian Filtering <http://dx.doi.org/10.1109/TSIPN.2016.2631944> *T. Li, J. M. Corchado, and J. Prieto* 580

Bayesian Learning Without Recall <http://dx.doi.org/10.1109/TSIPN.2016.2631943> *M. A. Rahimian and A. Jadbabaie* 592

Efficient Approximation and Denoising of Graph Signals Using the Multiscale Basis Dictionaries <http://dx.doi.org/10.1109/TSIPN.2016.2632039> *J. Irion and N. Saito* 607

Communication-Efficient Decentralized Sparse Bayesian Learning of Joint Sparse Signals <http://dx.doi.org/10.1109/TSIPN.2016.2632041> *S. Khanna and C. R. Murthy* 617

Modeling and Analysis

Learning the Interference Graph of a Wireless Network <http://dx.doi.org/10.1109/TSIPN.2016.2632040> *J. Yang, S. C. Draper, and R. Nowak* 631

EDICS—Editor’s Information Classification Scheme <http://dx.doi.org/10.1109/TSIPN.2017.2736242> 647

Information for Authors <http://dx.doi.org/10.1109/TSIPN.2017.2736244> 648




The Tenth IEEE Sensor Array and Multichannel
Signal Processing Workshop (www.sam2018.org)

8th-11th July 2018, Sheffield, United Kingdom



General Chairs

Wei Liu

University of Sheffield, UK

Peter Willett

University of Connecticut, US

Technical Chairs

Sergiy Vorobyov

Aalto University, Finland

Yimin D. Zhang

Temple University, US

IEEE SAM TC Representative

Mónica Bugallo

Stony Brook University, US

Special Session Chair

Hing Cheung So

City University of Hong Kong, HK

Finance Chair

Patrick Naylor

Imperial College London, UK

Publicity Chair

Hongbin Li

Stevens Institute of Technology, US

Local Arrangement Chair

Lyudmila Mihaylova

University of Sheffield, UK

Important Dates

Tutorial Proposals

22nd January, 2018

Special Session Proposals

5th February, 2018

Submission of Papers

24th February, 2018

Notification of Acceptance

30th April, 2018

Final Manuscript Submission

13th May, 2018

Advance Registration

20th May, 2018

Call for Papers

Technical Program

The SAM Workshop is an important IEEE Signal Processing Society event dedicated to sensor array and multichannel signal processing. The organizing committee invites the international community to contribute with state-of-the-art developments in the field. SAM 2018 will feature plenary talks by leading researchers in the field as well as poster and oral sessions with presentations by the participants.

Welcome to Sheffield!

The workshop will be held at Sheffield, the "Steel City". It is the third largest English district by population, and built on seven hills, like Rome. An estimated 2 million trees in the exuberant city, giving Sheffield the highest ratio of trees to people of any city in Europe. In particular, it is at the doorstep of the first UK national park -- the Peak District, offering breath-taking views and fantastic opportunities for pastimes such as cycling, walking and wildlife watching.

Research Areas

Authors are invited to submit contributions in the following areas:

- Adaptive beamforming
- Array processing for biomedical applications
- Array processing for communications
- Blind source separation and channel identification
- Computational and optimization techniques
- Compressive sensing and sparsity-based signal processing
- Detection and estimation
- Direction-of-arrival estimation
- Distributed and adaptive signal processing
- Intelligent systems and knowledge-based signal processing
- Microphone and loudspeaker array applications
- MIMO radar
- Multi-antenna systems: multiuser MIMO, massive MIMO and space-time coding
- Multi-channel imaging and hyperspectral processing
- Multi-sensor processing for smart grid and energy
- Non-Gaussian, nonlinear, and non-stationary models
- Performance evaluations with experimental data
- Radar and sonar array processing
- Sensor networks
- Source localization, classification and tracking
- Synthetic aperture techniques
- Space-time adaptive processing
- Statistical modelling for sensor arrays
- Waveform diverse sensors and systems

Submission of papers – Full-length five-page papers (last page with references only) will be accepted electronically at www.edas.info.

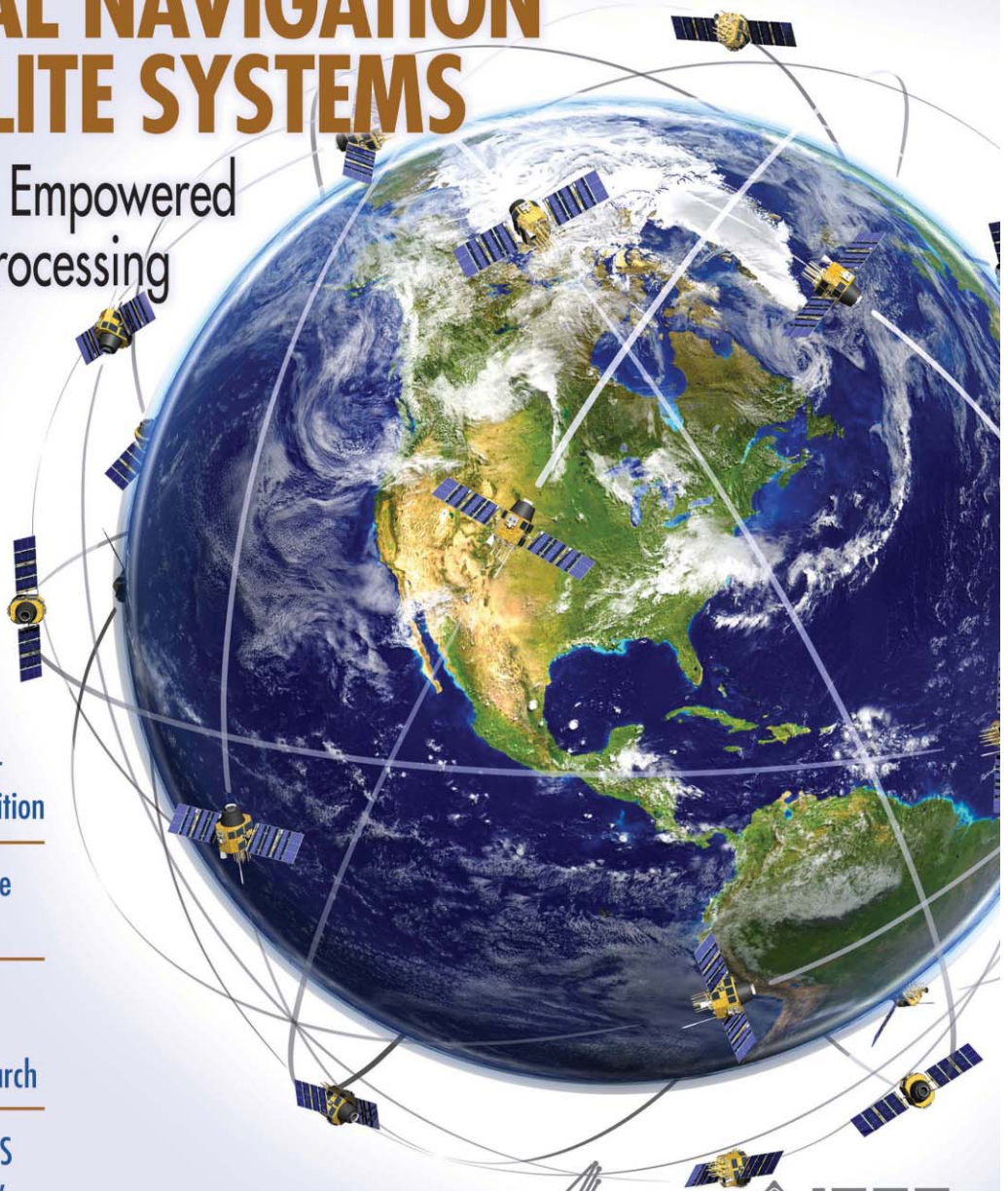
Submission of special session and tutorial proposals – details can be found at the workshop website.

IEEE Signal Processing MAGAZINE

Volume 34 | Number 5 | September 2017

GLOBAL NAVIGATION SATELLITE SYSTEMS

Its Advance Empowered
by Signal Processing



The Microsoft Indoor
Localization Competition

Gigabit-Rate Wireline
Communications

Signal Processing
and Learning for
Mental Health Research

Election Open for SPS
and IEEE: Vote Today

IEEE
Signal Processing Society

IEEE

Contents

Volume 34 | Number 5 | September 2017

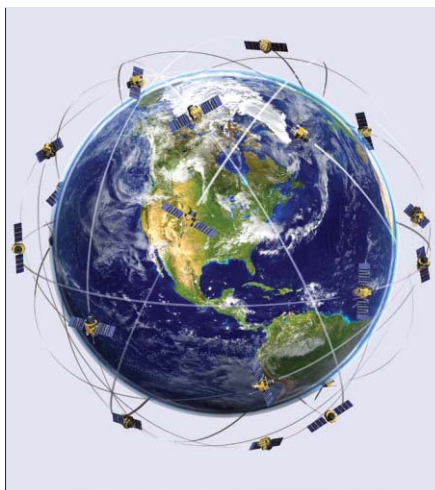
SPECIAL SECTION

ADVANCES IN SIGNAL PROCESSING FOR GLOBAL NAVIGATION SATELLITE SYSTEMS

- 12 FROM THE GUEST EDITORS**
Pau Closas, Marco Luise, José-Ángel Ávila-Rodríguez, Christopher Hegarty, and Jiyun Lee
- 16 SIGNAL MULTIPLEXING TECHNIQUES FOR GNSSs**
Zheng Yao and Mingquan Lu
- 27 SIGNAL STRUCTURE-BASED AUTHENTICATION FOR CIVIL GNSSs**
Davide Margaria, Beatrice Motella, Marco Anghileri, Jean-Jacques Floch, Ignacio Fernández-Hernández, and Matteo Paonni
- 38 UNAMBIGUOUS TECHNIQUES IN MODERNIZED GNSS SIGNALS**
Elena Simona Lohan, Diego Alonso de Diego, José A. López-Salcedo, Gonzalo Seco-Granados, Pedro Boto, and Pedro Fernandes



PG. 125



ON THE COVER

This special issue of *IEEE Signal Processing Magazine* begins by addressing the design of special global navigation satellite system signals and continues with the discussion of effective techniques for receiver performance enhancement. It finishes with the analysis of some vulnerabilities.

COVER IMAGE: ©ISTOCKPHOTO.COM/BLACKJACK3D

- 53 PROCESSING COST OF DOPPLER SEARCH IN GNSS SIGNAL ACQUISITION**
Sana U. Qaisar and Craig R. Benson
- 59 HIGH SENSITIVITY AND FAST ACQUISITION SIGNAL PROCESSING TECHNIQUES FOR GNSS RECEIVERS**
Seung-Hyun Kong
- 72 DIRECT POSITION ESTIMATION OF GNSS RECEIVERS**
Pau Closas and Adrià Gusi-Amigó
- 85 TIME-FREQUENCY ANALYSIS FOR GNSSs**
Moeness G. Amin, Daniele Borio, Yimin D. Zhang, and Lorenzo Galleani

96 MONITORING AND MITIGATION OF IONOSPHERIC ANOMALIES FOR GNSS-BASED SAFETY CRITICAL SYSTEMS

Jiyun Lee, Y.T. Jade Morton, Jinsil Lee, Hee-Seung Moon, and Jiwon Seo

111 I HEAR, THEREFORE I KNOW WHERE I AM

Zaher (Zak) M. Kassas, Joe Khalife, Kimia Shamaei, and Joshua Morales

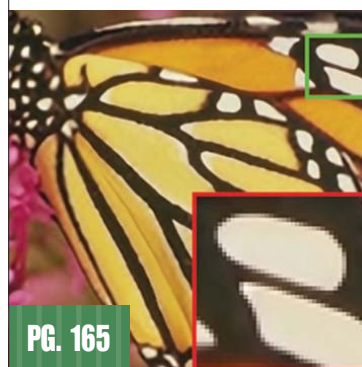
FEATURES

125 THE MICROSOFT INDOOR LOCALIZATION COMPETITION

Dimitrios Lymberopoulos and Jie Liu

141 SIGNAL PROCESSING FOR GIGABIT-RATE WIRELINE COMMUNICATIONS

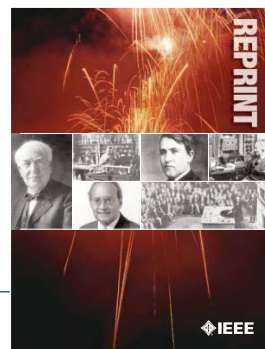
S.M. Zafaruddin, Itsik Bergel, and Amir Leshem



PG. 165

IEEE SIGNAL PROCESSING MAGAZINE (ISSN 1053-5888) (ISPREG) is published bimonthly by the Institute of Electrical and Electronics Engineers, Inc., 3 Park Avenue, 17th Floor, New York, NY 10016-5997 USA (+1 212 419 7900). Responsibility for the contents rests upon the authors and not the IEEE, the Society, or its members. Annual member subscriptions included in Society fee. Nonmember subscriptions available upon request. **Individual copies:** IEEE Members US\$20.00 (first copy only), nonmembers US\$241.00 per copy. Copyright and Reprint Permissions: Abstracting is permitted with credit to the source. Libraries are permitted to photocopy beyond the limits of U.S. Copyright Law for private use of patrons: 1) those post-1977 articles that carry a code at the bottom of the first page, provided the per-copy fee indicated in the code is paid through the Copyright Clearance Center, 222 Rosewood Drive, Danvers, MA 01923 USA; 2) pre-1978 articles without fee. Instructors are permitted to photocopy isolated articles for noncommercial classroom use without fee. **For all other copying, reprint, or republication permission,** write to IEEE Service Center, 445 Hoes Lane, Piscataway, NJ 08854 USA. Copyright © 2017 by the Institute of Electrical and Electronics Engineers, Inc. All rights reserved. Periodicals postage paid at New York, NY, and at additional mailing offices. **Postmaster:** Send address changes to IEEE Signal Processing Magazine, IEEE, 445 Hoes Lane, Piscataway, NJ 08854 USA. Canadian GST #125634188 **Printed in the U.S.A.**

Digital Object Identifier 10.1109/MSP.2017.2713740



IEEE ORDER FORM FOR REPRINTS

Purchasing IEEE Papers in Print is easy, cost-effective and quick.

Complete this form, send via our secure fax (24 hours a day) to 732-981-8062 or mail it back to us.

PLEASE FILL OUT THE FOLLOWING

Author: _____

Publication Title: _____

Paper Title: _____

RETURN THIS FORM TO:
 IEEE Publishing Services
 445 Hoes Lane
 Piscataway, NJ 08855-1331

Email the Reprint Department at reprints@ieee.org for questions regarding this form

PLEASE SEND ME

- 50 100 200 300 400 500 or _____ (in multiples of 50) reprints.
- YES NO Self-covering/title page required. COVER PRICE: \$74 per 100, \$39 per 50.
- \$58.00 Air Freight must be added for all orders being shipped outside the U.S.
- \$21.50 must be added for all USA shipments to cover the cost of UPS shipping and handling.

PAYMENT

- Check enclosed. Payable on a bank in the USA.
- Charge my: Visa Mastercard Amex Diners Club

Account # _____ Exp. date _____

Cardholder's Name (please print): _____

Bill me (you must attach a purchase order) Purchase Order Number _____

Send Reprints to: _____ Bill to address, if different: _____

Because information and papers are gathered from various sources, there may be a delay in receiving your reprint request. This is especially true with postconference publications. Please provide us with contact information if you would like notification of a delay of more than 12 weeks.

Telephone: _____ Fax: _____ Email Address: _____

2012 REPRINT PRICES (without covers)

Number of Text Pages

	1-4	5-8	9-12	13-16	17-20	21-24	25-28	29-32	33-36	37-40	41-44	45-48
50	\$129	\$213	\$245	\$248	\$288	\$340	\$371	\$408	\$440	\$477	\$510	\$543
100	\$245	\$425	\$479	\$495	\$573	\$680	\$742	\$817	\$885	\$953	\$1021	\$1088

Larger quantities can be ordered. Email reprints@ieee.org with specific details.

Tax Applies on shipments of regular reprints to CA, DC, FL, MI, NJ, NY, OH and Canada (GST Registration no. 12534188).
 Prices are based on black & white printing. Please call us for full color price quote, if applicable.

Authorized Signature: _____ Date: _____

2018 IEEE MEMBERSHIP APPLICATION

(students and graduate students must apply online)

Start your membership immediately: Join online www.ieee.org/join

Please complete both sides of this form, typing or **printing in capital letters**. Use only English characters and abbreviate only if more than 40 characters and spaces per line. We regret that incomplete applications cannot be processed.

1 Name & Contact Information

Please PRINT your name as you want it to appear on your membership card and IEEE correspondence. As a key identifier for the IEEE database, circle your last/surname.

Male Female Date of Birth (Day/Month/Year) ____/____/____

Title First/Given Name Middle Last/Family/Surname

▼ Primary Address Home Business (All IEEE mail sent here)

Street Address

City State/Province

Postal Code Country

Primary Phone

Primary E-mail

▼ Secondary Address Home Business

Company Name Department/Division

Street Address City State/Province

Postal Code Country

Secondary Phone

Secondary E-mail

To better serve our members and supplement member dues, your postal mailing address is made available to carefully selected organizations to provide you with information on technical services, continuing education, and conferences. Your e-mail address is not rented by IEEE. Please check box only if you do not want to receive these postal mailings to the selected address.

2 Attestation

I have graduated from a three- to five-year academic program with a university-level degree.
 Yes No

This program is in one of the following fields of study:

- Engineering
- Computer Sciences and Information Technologies
- Physical Sciences
- Biological and Medical Sciences
- Mathematics
- Technical Communications, Education, Management, Law and Policy
- Other (please specify): _____

This academic institution or program is accredited in the country where the institution is located.
 Yes No Do not know

I have ____ years of professional experience in teaching, creating, developing, practicing, or managing within the following field:

- Engineering
- Computer Sciences and Information Technologies
- Physical Sciences
- Biological and Medical Sciences
- Mathematics
- Technical Communications, Education, Management, Law and Policy
- Other (please specify): _____



3 Please Tell Us About Yourself

Select the numbered option that best describes yourself. This information is used by IEEE magazines to verify their annual circulation. Please enter numbered selections in the boxes provided.

A. Primary Line of Business

1. Computers
2. Computer peripheral equipment
3. Software
4. Office and business machines
5. Test, measurement, and instrumentation equipment
6. Communications systems and equipment
7. Navigation and guidance systems and equipment
8. Consumer electronics/appliances
9. Industrial equipment, controls, and systems
10. ICs and microprocessors
11. Semiconductors, components, sub-assemblies, materials, and supplies
12. Aircraft, missiles, space, and ground support equipment
13. Oceanography and support equipment
14. Medical electronic equipment
15. OEM incorporating electronics in their end product (not elsewhere classified)
16. Independent and university research, test and design laboratories, and consultants (not connected with a mfg. co.)
17. Government agencies and armed forces
18. Companies using and/or incorporating any electronic products in their manufacturing, processing, research, or development activities
19. Telecommunications services, telephone (including cellular)
20. Broadcast services (TV, cable, radio)
21. Transportation services (airline, railroad, etc.)
22. Computer and communications and data processing services
23. Power production, generation, transmission, and distribution
24. Other commercial users of electrical, electronic equipment, and services (not elsewhere classified)
25. Distributor (reseller, wholesaler, retailer)
26. University, college/other educational institutions, libraries
27. Retired
28. Other _____

B. Principal Job Function

1. General and corporate management
2. Engineering management
3. Project engineering management
4. Research and development management
5. Design engineering management —analog
6. Design engineering management —digital
7. Research and development engineering
8. Design/development engineering —analog
9. Design/development engineering—digital
10. Hardware engineering
11. Software design/development
12. Computer science
13. Science/physics/mathematics
14. Engineering (not elsewhere specified)
15. Marketing/sales/purchasing
16. Consulting
17. Education/teaching
18. Retired
19. Other _____

C. Principal Responsibility

1. Engineering and scientific management
2. Management other than engineering
3. Engineering design
4. Engineering
5. Software: science/mngmnt/engineering
6. Education/teaching
7. Consulting
8. Retired
9. Other _____

D. Title

1. Chairman of the Board/President/CEO
2. Owner/Partner
3. General Manager
4. VP Operations
5. VP Engineering/Dir. Engineering
6. Chief Engineer/Chief Scientist
7. Engineering Management
8. Scientific Management
9. Member of Technical Staff
10. Design Engineering Manager
11. Design Engineer
12. Hardware Engineer
13. Software Engineer
14. Computer Scientist
15. Dean/Professor/Instructor
16. Consultant
17. Retired
18. Other _____

Are you now or were you ever a member of IEEE?

Yes No If yes, provide, if known:

Membership Number Grade Year Expired

4 Please Sign Your Application

I hereby apply for IEEE Membership and agree to be governed by the IEEE Constitution, Bylaws, and Code of Ethics. I understand that IEEE will communicate with me regarding my individual membership and all related benefits. **Application must be signed.**

Signature _____ Date _____ Next Page

(continued on next page)

5 Add IEEE Society Memberships (Optional)

The 39 IEEE Societies support your technical and professional interests. Many society memberships include a personal subscription to the core journal, magazine, or newsletter of that society. **For a complete list of everything included with your IEEE Society membership, visit www.ieee.org/join.** All prices are quoted in US dollars.

Please check the appropriate box.

		BETWEEN 16 AUG 2017- 28 FEB 2018 PAY	BETWEEN 1 MAR 2018- 15 AUG 2018 PAY
IEEE Aerospace and Electronic Systems <input checked="" type="checkbox"/> <input checked="" type="checkbox"/>	AES010	25.00 <input type="checkbox"/>	12.50 <input type="checkbox"/>
IEEE Antennas and Propagation <input checked="" type="checkbox"/> <input checked="" type="checkbox"/>	AP003	15.00 <input type="checkbox"/>	7.50 <input type="checkbox"/>
IEEE Broadcast Technology <input checked="" type="checkbox"/> <input checked="" type="checkbox"/>	BT002	15.00 <input type="checkbox"/>	7.50 <input type="checkbox"/>
IEEE Circuits and Systems <input checked="" type="checkbox"/> <input checked="" type="checkbox"/>	CAS004	22.00 <input type="checkbox"/>	11.00 <input type="checkbox"/>
IEEE Communications <input checked="" type="checkbox"/> <input checked="" type="checkbox"/>	COM019	33.00 <input type="checkbox"/>	16.50 <input type="checkbox"/>
IEEE Computational Intelligence <input checked="" type="checkbox"/> <input checked="" type="checkbox"/>	CIS011	29.00 <input type="checkbox"/>	14.50 <input type="checkbox"/>
IEEE Computer <input checked="" type="checkbox"/> <input checked="" type="checkbox"/>	C016	60.00 <input type="checkbox"/>	30.00 <input type="checkbox"/>
IEEE Consumer Electronics <input checked="" type="checkbox"/> <input checked="" type="checkbox"/>	CE008	20.00 <input type="checkbox"/>	10.00 <input type="checkbox"/>
IEEE Control Systems <input checked="" type="checkbox"/> <input checked="" type="checkbox"/>	CS023	25.00 <input type="checkbox"/>	12.50 <input type="checkbox"/>
IEEE Dielectrics and Electrical Insulation <input checked="" type="checkbox"/> <input checked="" type="checkbox"/>	DEI032	26.00 <input type="checkbox"/>	13.00 <input type="checkbox"/>
IEEE Education <input checked="" type="checkbox"/> <input checked="" type="checkbox"/>	E025	20.00 <input type="checkbox"/>	10.00 <input type="checkbox"/>
IEEE Electromagnetic Compatibility <input checked="" type="checkbox"/> <input checked="" type="checkbox"/>	EMC027	31.00 <input type="checkbox"/>	15.50 <input type="checkbox"/>
IEEE Electron Devices <input checked="" type="checkbox"/> <input checked="" type="checkbox"/>	ED015	18.00 <input type="checkbox"/>	9.00 <input type="checkbox"/>
IEEE Electronics Packaging Society <input checked="" type="checkbox"/> <input checked="" type="checkbox"/>	EPO21	15.00 <input type="checkbox"/>	7.50 <input type="checkbox"/>
IEEE Engineering in Medicine and Biology <input checked="" type="checkbox"/> <input checked="" type="checkbox"/>	EMB018	45.00 <input type="checkbox"/>	22.50 <input type="checkbox"/>
IEEE Geoscience and Remote Sensing <input checked="" type="checkbox"/> <input checked="" type="checkbox"/>	GRS029	19.00 <input type="checkbox"/>	9.50 <input type="checkbox"/>
IEEE Industrial Electronics <input checked="" type="checkbox"/> <input checked="" type="checkbox"/>	IE013	9.00 <input type="checkbox"/>	4.50 <input type="checkbox"/>
IEEE Industry Applications <input checked="" type="checkbox"/> <input checked="" type="checkbox"/>	IA034	20.00 <input type="checkbox"/>	10.00 <input type="checkbox"/>
IEEE Information Theory <input checked="" type="checkbox"/> <input checked="" type="checkbox"/>	IT012	30.00 <input type="checkbox"/>	15.00 <input type="checkbox"/>
IEEE Instrumentation and Measurement <input checked="" type="checkbox"/> <input checked="" type="checkbox"/>	IM009	29.00 <input type="checkbox"/>	14.50 <input type="checkbox"/>
IEEE Intelligent Transportation Systems <input checked="" type="checkbox"/> <input checked="" type="checkbox"/>	ITSS038	35.00 <input type="checkbox"/>	17.50 <input type="checkbox"/>
IEEE Magnetics <input checked="" type="checkbox"/> <input checked="" type="checkbox"/>	MAG033	26.00 <input type="checkbox"/>	13.00 <input type="checkbox"/>
IEEE Microwave Theory and Techniques <input checked="" type="checkbox"/> <input checked="" type="checkbox"/>	MTT017	24.00 <input type="checkbox"/>	12.00 <input type="checkbox"/>
IEEE Nuclear and Plasma Sciences <input checked="" type="checkbox"/> <input checked="" type="checkbox"/>	NPS005	35.00 <input type="checkbox"/>	17.50 <input type="checkbox"/>
IEEE Oceanic Engineering <input checked="" type="checkbox"/> <input checked="" type="checkbox"/>	OE022	19.00 <input type="checkbox"/>	9.50 <input type="checkbox"/>
IEEE Photonics <input checked="" type="checkbox"/> <input checked="" type="checkbox"/>	PHO036	34.00 <input type="checkbox"/>	17.00 <input type="checkbox"/>
IEEE Power Electronics <input checked="" type="checkbox"/> <input checked="" type="checkbox"/>	PEL035	25.00 <input type="checkbox"/>	12.50 <input type="checkbox"/>
IEEE Power & Energy <input checked="" type="checkbox"/> <input checked="" type="checkbox"/>	PE031	30.00 <input type="checkbox"/>	15.00 <input type="checkbox"/>
IEEE Product Safety Engineering <input checked="" type="checkbox"/> <input checked="" type="checkbox"/>	PSE043	35.00 <input type="checkbox"/>	17.50 <input type="checkbox"/>
IEEE Professional Communication <input checked="" type="checkbox"/> <input checked="" type="checkbox"/>	PC026	31.00 <input type="checkbox"/>	15.50 <input type="checkbox"/>
IEEE Reliability <input checked="" type="checkbox"/> <input checked="" type="checkbox"/>	RL007	35.00 <input type="checkbox"/>	17.50 <input type="checkbox"/>
IEEE Robotics and Automation <input checked="" type="checkbox"/> <input checked="" type="checkbox"/>	RA024	9.00 <input type="checkbox"/>	4.50 <input type="checkbox"/>
IEEE Signal Processing <input checked="" type="checkbox"/> <input checked="" type="checkbox"/>	SP001	22.00 <input type="checkbox"/>	11.00 <input type="checkbox"/>
IEEE Social Implications of Technology <input checked="" type="checkbox"/> <input checked="" type="checkbox"/>	SIT030	33.00 <input type="checkbox"/>	16.50 <input type="checkbox"/>
IEEE Solid-State Circuits <input checked="" type="checkbox"/> <input checked="" type="checkbox"/>	SSC037	22.00 <input type="checkbox"/>	11.00 <input type="checkbox"/>
IEEE Systems, Man, & Cybernetics <input checked="" type="checkbox"/> <input checked="" type="checkbox"/>	SMC028	12.00 <input type="checkbox"/>	6.00 <input type="checkbox"/>
IEEE Technology & Engineering Management <input checked="" type="checkbox"/> <input checked="" type="checkbox"/>	TEM014	35.00 <input type="checkbox"/>	17.50 <input type="checkbox"/>
IEEE Ultrasonics, Ferroelectrics, & Frequency Control <input checked="" type="checkbox"/> <input checked="" type="checkbox"/>	UFFC020	20.00 <input type="checkbox"/>	10.00 <input type="checkbox"/>
IEEE Vehicular Technology <input checked="" type="checkbox"/> <input checked="" type="checkbox"/>	VT006	18.00 <input type="checkbox"/>	9.00 <input type="checkbox"/>

Legend—Society membership includes:

- One or more Society publications
- Society newsletter
- Online access to publication
- CD-ROM of selected society publications

Complete both sides of this form, sign, and return to:
 IEEE MEMBERSHIP APPLICATION PROCESSING
 445 HOES LN, PISCATAWAY, NJ 08854-4141 USA
 or fax to +1 732 981 0225
or join online at www.ieee.org/join

Please reprint your full name here _____

6 2018 IEEE Membership Rates (student rates available online)

IEEE member dues and regional assessments are based on where you live and when you apply. Membership is based on the calendar year from 1 January through 31 December. All prices are quoted in US dollars.

Please check the appropriate box.

RESIDENCE

United States.....	\$201.00 <input type="checkbox"/>	\$100.50 <input type="checkbox"/>
Canada (NB, NF, NS, and PEI HST)*.....	\$191.50 <input type="checkbox"/>	\$95.75 <input type="checkbox"/>
Canada (ON HST).....	\$188.50 <input type="checkbox"/>	\$94.25 <input type="checkbox"/>
Canada (GST)*.....	\$176.50 <input type="checkbox"/>	\$88.25 <input type="checkbox"/>
Canada (GST and QST Quebec).....	\$191.46 <input type="checkbox"/>	\$95.73 <input type="checkbox"/>
European Union.....	\$168.00 <input type="checkbox"/>	\$84.00 <input type="checkbox"/>
Africa, Europe,** Middle East.....	\$163.00 <input type="checkbox"/>	\$81.50 <input type="checkbox"/>
Latin America.....	\$154.00 <input type="checkbox"/>	\$77.00 <input type="checkbox"/>
Asia, Pacific.....	\$155.00 <input type="checkbox"/>	\$77.50 <input type="checkbox"/>

*IEEE Canada Business No. 125634188
 **Excludes European Union countries

Minimum Income or Unemployed Provision

Applicants who certify that their prior year income did not exceed US\$15,000 (or equivalent) or were not employed are granted 50% reduction in: full-year dues, regional assessment, and fees for one IEEE Membership plus one Society Membership. If applicable, please check appropriate box and adjust payment accordingly. Student members are not eligible.

- I certify I earned less than US\$15,000 in 2017
- I certify that I was unemployed in 2017

7 More Recommended Options

Proceedings of the IEEE.....	print \$51.00 <input type="checkbox"/> or online \$45.00 <input type="checkbox"/>
Proceedings of the IEEE (print/online combination).....	\$61.00 <input type="checkbox"/>
IEEE Standards Association (IEEE-SA).....	\$55.00 <input type="checkbox"/>
IEEE Women in Engineering (WIE).....	\$25.00 <input type="checkbox"/>

8 Payment Amount

Please total the membership dues, society dues, and other amounts from this page:

IEEE Membership dues ⑥.....	\$_____
IEEE Society dues (optional) ⑤.....	\$_____
IEEE-SA/WIE dues (optional) ⑦.....	\$_____
Proceedings of the IEEE (optional) ⑧.....	\$_____
Canadian residents pay 5% GST or appropriate HST (BC-12%; ON-13%; NB, NF, NS, PEI-15%) on Society payments & publications only.....	TAX \$_____
AMOUNT PAID	TOTAL \$_____

Payment Method

All prices are quoted in US dollars. You may pay for IEEE Membership by credit card (see below), check, or money order payable to IEEE, drawn on a US bank.

Check

Credit Card Number _____
 MONTH YEAR CARDHOLDER'S 5-DIGIT ZIP CODE
 EXPIRATION DATE (BILLING STATEMENT ADDRESS) US ONLY

Name as it appears on card _____

Signature _____

Auto Renew my memberships and subscriptions (available when paying by credit card).
 I agree to the Terms and Conditions located at www.ieee.org/autorenew

9 Were You Referred to IEEE?

Yes No If yes, provide the following:
 Member Recruiter Name _____
 IEEE Recruiter's Member Number (Required) _____

CAMPAIGN CODE _____ PROMO CODE _____

Information for Authors

(Updated/Effective July 2017)

For Transactions and Journals:

Authors are encouraged to submit manuscripts of Regular papers (papers which provide a complete disclosure of a technical premise), or Comment Correspondences (brief items that provide comment on a paper previously published in these TRANSACTIONS).

Submissions/resubmissions must be previously unpublished and may not be under consideration elsewhere.

Every manuscript must:

- i. provide a clear statement of the problem and what the contribution of the work is to the relevant research community;
- ii. state why this contribution is significant (what impact it will have);
- iii. provide citation of the published literature most closely related to the manuscript; and
- iv. state what is distinctive and new about the current manuscript relative to these previously published works.

By submission of your manuscript to these TRANSACTIONS, all listed authors have agreed to the authorship list and all the contents and confirm that the work is original and that figures, tables and other reported results accurately reflect the experimental work. In addition, the authors all acknowledge that they accept the rules established for publication of manuscripts, including agreement to pay all overlength page charges, color charges, and any other charges and fees associated with publication of the manuscript. Such charges are not negotiable and cannot be suspended. The corresponding author is responsible for obtaining consent from all co-authors and, if needed, from sponsors before submission.

In order to be considered for review, a paper must be within the scope of the journal and represent a novel contribution. A paper is a candidate for an Immediate Rejection if it is of limited novelty, e.g. a straightforward combination of theories and algorithms that are well established and are repeated on a known scenario. Experimental contributions will be rejected without review if there is insufficient experimental data. These TRANSACTIONS are published in English. Papers that have a large number of typographical and/or grammatical errors will also be rejected without review.

In addition to presenting a novel contribution, acceptable manuscripts must describe and cite related work in the field to put the contribution in context. Do not give theoretical derivations or algorithm descriptions that are easily found in the literature; merely cite the reference.

New and revised manuscripts should be prepared following the "Manuscript Submission" guidelines below, and submitted to the online manuscript system, ScholarOne Manuscripts. Do not send original submissions or revisions directly to the Editor-in-Chief or Associate Editors; they will access your manuscript electronically via the ScholarOne Manuscript system.

Manuscript Submission. Please follow the next steps.

1. *Account in ScholarOne Manuscripts.* If necessary, create an account in the on-line submission system ScholarOne Manuscripts. Please check first if you already have an existing account which is based on your e-mail address and may have been created for you when you reviewed or authored a previous paper.
 - All IEEE journals require an Open Researcher and Contributor ID (ORCID) for all authors. ORCID is a persistent unique identifier for researchers and functions similarly to an article's Digital Object Identifier (DOI). The author will need a registered ORCID in order to submit a manuscript or review a proof in this journal.
2. *Electronic Manuscript.* Prepare a PDF file containing your manuscript in double-column, single-spaced format using a font size of 10 points or larger, having a margin of at least 1 inch on all sides. Upload this version of the manuscript as a PDF file "double.pdf" to the ScholarOne- Manuscripts site. Since many reviewers prefer a larger font, you are strongly encouraged to also submit a single-column, double-spaced version (11 point font or larger), which is easy to create with the templates provided [IEEE Author Digital Toolbox \(http://www.ieee.org/publications_standards/publications/authors/authors_journals.html\)](http://www.ieee.org/publications_standards/publications/authors/authors_journals.html). Page length restrictions will be determined by the double-column version. Proofread your submission, confirming that all figures and equations are visible in your document before you "SUBMIT"

your manuscript. Proofreading is critical; once you submit your manuscript, the manuscript cannot be changed in any way. You may also submit your manuscript as a .PDF or MS Word file. The system has the capability of converting your files to PDF, however it is your responsibility to confirm that the conversion is correct and there are no font or graphics issues prior to completing the submission process.

3. *EDICS (Not applicable to Journal of Selected Topics in Signal Processing).* All submissions must be classified by the author with an EDICS (Editors' Information Classification Scheme) selected from the list of EDICS published online at the publication's EDICS webpage (*please see the list below). Upon submission of a new manuscript, please choose the EDICS categories that best suit your manuscript. Failure to do so will likely result in a delay of the peer review process.
4. *Additional Documents for Review.* Please upload pdf versions of all items in the reference list that are not publicly available, such as unpublished (submitted) papers. Graphical abstracts and supplemental materials intended to appear with the final paper (see below) must also be uploaded for review at the time of the initial submission for consideration in the review process. Use short filenames without spaces or special characters. When the upload of each file is completed, you will be asked to provide a description of that file.
5. *Supplemental Materials.* IEEE Xplore can publish multimedia files (audio, images, video, pseudocode and detailed algebraic manipulations of proofs), datasets, and software (e.g. Matlab code) along with your paper. Alternatively, you can provide the links to such files in a README file that appears on Xplore along with your paper. For details, please see IEEE Author Digital Toolbox (http://www.ieee.org/publications_standards/publications/authors/authors_journals.html) under "Multimedia." To make your work reproducible by others, the TRANSACTIONS encourages you to submit all files that can recreate the figures in your paper.
6. *Submission.* After uploading all files and proofreading them, submit your manuscript by clicking "Submit." A confirmation of the successful submission will open on screen containing the manuscript tracking number and will be followed with an e-mail confirmation to the corresponding and all contributing authors. Once you click "Submit," your manuscript cannot be changed in any way.
7. *Copyright Form and Consent Form.* By policy, IEEE owns the copyright to the technical contributions it publishes on behalf of the interests of the IEEE, its authors, and their employers; and to facilitate the appropriate reuse of this material by others. To comply with the IEEE copyright policies, authors are required to sign and submit a completed "IEEE Copyright and Consent Form" prior to publication by the IEEE. The IEEE recommends authors to use an effective electronic copyright form (eCF) tool within the ScholarOne Manuscripts system. You will be redirected to the "IEEE Electronic Copyright Form" wizard at the end of your original submission; please simply sign the eCF by typing your name at the proper location and click on the "Submit" button.

Comment Correspondence. Comment Correspondences provide brief comments on material previously published in these TRANSACTIONS. These items may not exceed 2 pages in double-column, single spaced format, using 9 point type, with margins of 1 inch minimum on all sides, and including: title, names and contact information for authors, abstract, text, references, and an appropriate number of illustrations and/or tables. Correspondence items are submitted in the same way as regular manuscripts (see "Manuscript Submission" above for instructions).

Authors may also submit manuscripts of overview articles, but note that these include an additional white paper approval process <http://www.signalprocessingsociety.org/publications/overview-articles/>. [This does not apply to the *Journal of Selected Topics in Signal Processing*. Please contact the Editor-in-Chief.]

Manuscript Length. For the initial submission of a regular paper, the manuscript may not exceed 13 double-column pages (10 point font), including title; names of authors and their complete contact information; abstract; text; all images, figures and tables, appendices and proofs; and all references. Supplemental materials and graphical

Digital Object Identifier 10.1109/TIP.2017.2740787

abstracts are not included in the page count. For regular papers, the revised manuscript may not exceed 16 double-column pages (10 point font), including title; names of authors and their complete contact information; abstract; text; all images, figures and tables, appendices and proofs; and all references. For Overview Papers, the maximum length is double that for regular submissions at each stage (please reference <http://www.signalprocessingsociety.org/publications/overview-articles/> for more information).

Note that any paper in excess of 10 pages will be subject to mandatory overlength page charges. Since changes recommended as a result of peer review may require additions to the manuscript, it is strongly recommended that you practice economy in preparing original submissions. Note: Papers submitted to the TRANSACTIONS ON MULTIMEDIA in excess of 8 pages will be subject to mandatory overlength page charges.

Exceptions to manuscript length requirements may, under extraordinary circumstances, be granted by the Editor-in-Chief. However, such exception does not obviate your requirement to pay any and all overlength or additional charges that attach to the manuscript.

Resubmission of Previously Rejected Manuscripts. Authors of manuscripts rejected from any journal are allowed to resubmit their manuscripts only once. At the time of submission, you will be asked whether your manuscript is a new submission or a resubmission of an earlier rejected manuscript. If it is a resubmission of a manuscript previously rejected by any journal, you are expected to submit supporting documents identifying the previous submission and detailing how your new version addresses all of the reviewers' comments. Papers that do not disclose connection to a previously rejected paper or that do not provide documentation as to changes made may be immediately rejected.

Author Misconduct. Author misconduct includes plagiarism, self-plagiarism, and research misconduct, including falsification or misrepresentation of results. All forms of misconduct are unacceptable and may result in sanctions and/or other corrective actions. Plagiarism includes copying someone else's work without appropriate credit, using someone else's work without clear delineation of citation, and the uncited reuse of an author's previously published work that also involves other authors. Self-plagiarism involves the verbatim copying or reuse of an authors own prior work without appropriate citation, including duplicate submission of a single journal manuscript to two different journals, and submission of two different journal manuscripts which overlap substantially in language or technical contribution. For more information on the definitions, investigation process, and corrective actions related to author misconduct, see the Signal Processing Society Policies and Procedures Manual, Section 6.1. <http://www.signalprocessingsociety.org/about-sps/governance/policy-procedure/part-2>. Author misconduct may also be actionable by the IEEE under the rules of Member Conduct.

Extensions of the Author's Prior Work. It is acceptable for conference papers to be used as the basis for a more fully developed journal submission. Still, authors are required to cite their related prior work; the papers cannot be identical; and the journal publication must include substantively novel aspects such as new experimental results and analysis or added theoretical work. The journal publication should clearly specify how the journal paper offers novel contributions when citing the prior work. Limited overlap with prior journal publications with a common author is allowed only if it is necessary for the readability of the paper, and the prior work must be cited as the primary source.

Submission Format. Authors are required to prepare manuscripts employing the on-line style files developed by IEEE, which include guidelines for abbreviations, mathematics, and graphics. All manuscripts accepted for publication will require the authors to make final submission employing these style files. The style files are available on the web at the **IEEE Author Digital Toolbox** under "Template for all TRANSACTIONS." (LaTeX and MS Word). Please note the following requirements about the abstract:

- The abstract must be a concise yet comprehensive reflection of what is in your article.
- The abstract must be self-contained, without abbreviations, footnotes, displayed equations, or references.
- The abstract must be between 150–250 words.
- The abstract should include a few keywords or phrases, as this will help readers to find it. Avoid over-repetition of such phrases as this can result in a page being rejected by search engines.

In addition to written abstracts, papers may include a graphical abstract; see http://www.ieee.org/publications_standards/publications/graphical_abstract.pdf for options and format requirements.

IEEE supports the publication of author names in the native language alongside the English versions of the names in the author list of an article. For more information, see "Author names in native languages" (http://www.ieee.org/publications_standards/publications/authors/auth_names_native_lang.pdf) on the IEEE Author Digital Toolbox page.

Refining the Use of English Language in Your Manuscript. English language editing services can help refine the language of your article and reduce the risk of rejection without review. IEEE authors are eligible for a 10% discount at American Journal Experts; visit <http://www.aje.com/go/ieee/> to learn more. Please note these services are fee-based and do not guarantee acceptance. You are also free to select another professional editing service, or to ask a colleague who is fluent in English to assist with editing. However, if the revised manuscript does not meet the English usage criteria, then you must use the designated editing service (AJE [<http://www.aje.com/en/>]) for a fee, or you must withdraw the manuscript.

Open Access. The publication is a hybrid journal, allowing either Traditional manuscript submission or Open Access (author-pays OA) manuscript submission. Upon submission of your final files, if you choose to have your manuscript be an Open Access article, you commit to pay the discounted OA fee if your manuscript is accepted for publication in order to enable unrestricted public access. As of 01 January 2017, the OA fee is \$1,950. Any other application charges (such as charge for the use of color in the print format) will be billed separately once the manuscript formatting is complete but prior to the publication. Any other application charges (such as overlength page charge and/or charge for the use of color in the print format) will be billed separately once the manuscript formatting is complete but prior to the publication. If you would like your manuscript to be a Traditional submission, your article will be available to qualified subscribers and purchasers via IEEE Xplore. No OA payment is required for Traditional submission.

Page Charges.

Voluntary Page Charges. Upon acceptance of a manuscript for publication, the author(s) or his/her/their company or institution will be asked to pay a charge of \$110 per page to cover part of the cost of publication of the first ten pages that comprise the standard length (two pages, in the case of Correspondences).

Mandatory Page Charges. The author(s) or his/her/their company or institution will be billed \$220 per each page in excess of the first ten published pages for regular papers. (**NOTE: Regular Papers accepted to IEEE TRANSACTIONS ON MULTIMEDIA in excess of 8 pages will be subject to mandatory overlength page charges, and correspondence papers accepted to T-MM in excess of 6 published pages will also be subject to overlength page charges.) These are mandatory page charges and the author(s) will be held responsible for them. They are not negotiable or voluntary. The author(s) signifies his willingness to pay these charges simply by submitting his/her/their manuscript to the TRANSACTIONS. The Publisher holds the right to withhold publication under any circumstance, as well as publication of the current or future submissions of authors who have outstanding mandatory page charge debt. No mandatory overlength page charges will be applied to overview articles in the Society's journals.

Color Charges. The color-in-print charge is a flat rate of \$275 per figure. The corresponding author of the article will have the opportunity to address the color-in-print option during an "Article Setup" step. All invoices and payments are handled through an automated payment portal system. The payment portal allows various payment types such as credit card, bank wire transfers, check, pre-approved waivers, special payment circumstances, and third party billing. Please note that split payments are not supported at this time. If you have any questions, please contact aoprocessing@ieee.org for Open Access processing and reprints@ieee.org for all other charges.

*EDICS Webpages:

IEEE TRANSACTIONS ON SIGNAL PROCESSING:
<http://www.signalprocessingsociety.org/publications/periodicals/tsp/TSP-EDICS/>
 IEEE TRANSACTIONS ON IMAGE PROCESSING:
<http://www.signalprocessingsociety.org/publications/periodicals/image-processing/tip-edics/>
 IEEE/ACM TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE / ACM:
<http://www.signalprocessingsociety.org/publications/periodicals/taslp/taslp-edics/>
 IEEE TRANSACTIONS ON INFORMATION, FORENSICS AND SECURITY:
<http://www.signalprocessingsociety.org/publications/periodicals/forensics/forensics-edics/>
 IEEE TRANSACTIONS ON MULTIMEDIA:
<http://www.signalprocessingsociety.org/tmm/tmm-edics/>
 IEEE TRANSACTIONS ON COMPUTATIONAL IMAGING:
<http://www.signalprocessingsociety.org/publications/periodicals/tci/tci-edics/>
 IEEE TRANSACTIONS ON SIGNAL AND INFORMATION PROCESSING OVER NETWORKS:
<http://www.signalprocessingsociety.org/publications/periodicals/tsipn/tsipn-edics/>

Join Now for 2018

The IEEE Signal Processing Society is the world's premier association for signal processing engineers and industry professionals, serving its nearly 17,000 members with highly-ranked publications, world class conferences, professional development resources, job opportunities, and more.

CONNECT

Network with other professionals through SPS conferences, workshops, Technical Committees, Special Interest Groups, and local events curated by more than 180 worldwide SPS Chapters.

SAVE

Access members-only discounts on SPS publications and conferences. Gain eligibility to apply for travel grants to our flagship conferences ICASSP, ICIP, and GlobalSIP.

ADVANCE

Further your career with world-class educational resources, including the new SPS Resource Center, opportunities for awards and recognition, and volunteer opportunities across society activities.



SCAN TO JOIN



@ieeeSPS



/ieeeSPS



signalprocessingsociety.org

JOIN SPS TODAY AND RECEIVE

Benefit and Package	Essential Membership	Preferred Membership
Inside Signal Processing eNewsletter	✓	✓
IEEE Signal Processing Magazine	Digital Electronic	Digital Electronic Print
IEEE Signal Processing Content Gazette	✓	✓
Signal Processing Digital Library <i>Electronic access to seven solely-owned SPS publications through IEEE Xplore®</i>		✓
SPS Resource Center	✓	✓
IEEE Professional Member Price <i>Membership through 31 December 2017</i>	\$22.00	\$39.00
IEEE Student Member Price <i>Membership through 31 December 2017</i>	\$11.00	\$20.00
Affiliate Member Price <i>Membership through 31 December 2017</i>	\$96.50	\$113.50

In addition, all SPS members receive:

- › **Networking** and **collaboration** opportunities with a global network of nearly 17,000 signal processing professionals
- › **Discounts** on SPS conferences and workshops, including our flagship conferences ICASSP, ICIP, and GlobalSIP
- › **Discounts** on print editions of SPS-sponsored publications
- › Eligibility to apply for **travel grants** to SPS conferences
- › **Connect** with members near you through local events curated by SPS' 180+ worldwide Chapters
- › Career growth and **professional development** tools and resources
- › Eligibility to join a **Technical Committee** or **Special Interest Group** to meet SPS members with similar technical interests to develop and strengthen technical communities within signal processing, having a voice in awards, conferences, publications, education, and more
- › **Volunteer** opportunities throughout Society activities, including publications, conferences, membership, public visibility, and more.
- › **Students** get exclusive access to competitions, job fairs, and networking events

signalprocessingsociety.org

