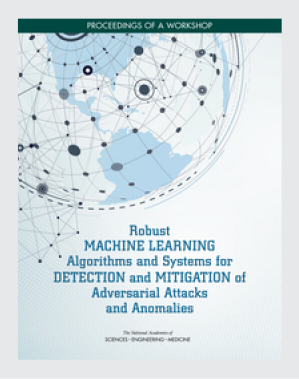


Robust Machine Learning Algorithms and Systems for Detection and Mitigation of Adversarial Attacks and Anomalies: Proceedings of a Workshop (2019)

DETAILS

82 pages | 8.5 x 11 | PAPERBACK
ISBN 978-0-309-49609-4 | DOI 10.17226/25534

CONTRIBUTORS

Linda Casola and Dionna Ali, Rapporteurs; Intelligence Community Studies Board; Board on Mathematical Sciences and Analytics; Computer Science and Telecommunications Board; Division on Engineering and Physical Sciences; National Academies of Sciences, Engineering, and Medicine

SUGGESTED CITATION

National Academies of Sciences, Engineering, and Medicine 2019. *Robust Machine Learning Algorithms and Systems for Detection and Mitigation of Adversarial Attacks and Anomalies: Proceedings of a Workshop*. Washington, DC: The National Academies Press. https://doi.org/10.17226/25534.

Robust MACHINE LEARNING Algorithms and Systems for DETECTION and MITIGATION of Adversarial Attacks and Anomalies

PROCEEDINGS OF A WORKSHOP

Linda Casola and Dionna Ali, *Rapporteurs*

Intelligence Community Studies Board

Board on Mathematical Sciences and Analytics

Computer Science and Telecommunications Board

Division on Engineering and Physical Sciences

The National Academies of
SCIENCES • ENGINEERING • MEDICINE

THE NATIONAL ACADEMIES PRESS
Washington, DC
www.nap.edu

THE NATIONAL ACADEMIES PRESS 500 Fifth Street, NW Washington, DC 20001

This activity was supported by Contract 2014-14041100003-019 with the Office of the Director of National Intelligence. Any opinions, findings, conclusions, or recommendations expressed in this publication do not necessarily reflect the views of any organization or agency that provided support for the project.

International Standard Book Number-13: 978-0-309-49609-4
International Standard Book Number-10: 0-309-49609-8
Digital Object Identifier: https://doi.org/10.17226/25534

Additional copies of this publication are available for sale from the National Academies Press, 500 Fifth Street, NW, Keck 360, Washington, DC 20001; (800) 624-6242 or (202) 334-3313; http://www.nap.edu.

Printed in the United States of America

Suggested citation: National Academies of Sciences, Engineering, and Medicine. 2019. *Robust Machine Learning Algorithms and Systems for Detection and Mitigation of Adversarial Attacks and Anomalies: Proceedings of a Workshop.* Washington, DC: The National Academies Press. doi: https://doi.org/10.17226/25534.

The National Academies of
SCIENCES • ENGINEERING • MEDICINE

The **National Academy of Sciences** was established in 1863 by an Act of Congress, signed by President Lincoln, as a private, nongovernmental institution to advise the nation on issues related to science and technology. Members are elected by their peers for outstanding contributions to research. Dr. Marcia McNutt is president.

The **National Academy of Engineering** was established in 1964 under the charter of the National Academy of Sciences to bring the practices of engineering to advising the nation. Members are elected by their peers for extraordinary contributions to engineering. Dr. John L. Anderson is president.

The **National Academy of Medicine** (formerly the Institute of Medicine) was established in 1970 under the charter of the National Academy of Sciences to advise the nation on medical and health issues. Members are elected by their peers for distinguished contributions to medicine and health. Dr. Victor J. Dzau is president.

The three Academies work together as the **National Academies of Sciences, Engineering, and Medicine** to provide independent, objective analysis and advice to the nation and conduct other activities to solve complex problems and inform public policy decisions. The National Academies also encourage education and research, recognize outstanding contributions to knowledge, and increase public understanding in matters of science, engineering, and medicine.

Learn more about the National Academies of Sciences, Engineering, and Medicine at **www.nationalacademies.org**.

The National Academies of
SCIENCES • ENGINEERING • MEDICINE

Consensus Study Reports published by the National Academies of Sciences, Engineering, and Medicine document the evidence-based consensus on the study's statement of task by an authoring committee of experts. Reports typically include findings, conclusions, and recommendations based on information gathered by the committee and the committee's deliberations. Each report has been subjected to a rigorous and independent peer-review process and it represents the position of the National Academies on the statement of task.

Proceedings published by the National Academies of Sciences, Engineering, and Medicine chronicle the presentations and discussions at a workshop, symposium, or other event convened by the National Academies. The statements and opinions contained in proceedings are those of the participants and are not endorsed by other participants, the planning committee, or the National Academies.

For information about other products and activities of the National Academies, please visit www.nationalacademies.org/about/whatwedo.

PLANNING COMMITTEE ON ENSURING THE QUALITY OF MACHINE-GENERATED ANALYTIC PRODUCTS FROM MULTI-SOURCE DATA: A WORKSHOP

RAMA CHELLAPPA, University of Maryland, College Park, *Chair*
TODD BORKEY, Alion Science and Technology
JULIE BRILL, Microsoft Corporation
LISE GETOOR, University of California, Santa Cruz
ANTHONY HOOGS, Kitware, Inc.
ANITA JONES, NAE,[1] University of Virginia
YUNYAO LI, IBM Corporation
JOYSULA RAO, IBM Corporation
SAMUEL VISNER, MITRE Corporation

Staff

GEORGE COYLE, Senior Program Officer, *Workshop Director*
CHRIS JONES, Financial Officer
MARGUERITE SCHNEIDER, Administrative Coordinator
DIONNA ALI, Research Associate
NATHANIEL DEBEVOISE, Senior Program Assistant

[1] Member, National Academy of Engineering.

INTELLIGENCE COMMUNITY STUDIES BOARD

Staff

[1] Member, National Academy of Engineering.
[2] Member, National Academy of Sciences.
[3] Member, National Academy of Medicine.

[1] Member, National Academy of Engineering.
[2] Member, National Academy of Sciences.
[3] Member, National Academy of Medicine.

[1] Member, National Academy of Engineering.
[2] Member, National Academy of Sciences.

Acknowledgments

This Proceedings of a Workshop was reviewed in draft form by individuals chosen for their diverse perspectives and technical expertise. The purpose of this independent review is to provide candid and critical comments that will assist the National Academies of Sciences, Engineering, and Medicine in making each published proceedings as sound as possible and to ensure that it meets the institutional standards for quality, objectivity, evidence, and responsiveness to the charge. The review comments and draft manuscript remain confidential to protect the integrity of the process.

We thank the following individuals for their review of this proceedings:

Terrance Boult, University of Colorado, Colorado Springs,
Dianne Chong, NAE,[1] Boeing Research and Technology (retired),
Anita Jones, NAE,[2] University of Virginia, and
Yunyao Li, IBM Corporation.

Although the reviewers listed above provided many constructive comments and suggestions, they were not asked to endorse the content of the proceedings nor did they see the final draft before its release. The review of this proceedings was overseen by Ellen W. Clayton, NAM,[3] Vanderbilt University Medical Center. She was responsible for making certain that an independent examination of this proceedings was carried out in accordance with standards of the National Academies and that all review comments were carefully considered. We also wish to thank Michelle Schwalbe, National Academies, for her guidance in the drafting of this manuscript. Responsibility for the final content rests entirely with the rapporteurs and the National Academies.

[1] Member, National Academy of Engineering.
[2] Member, National Academy of Engineering.
[3] Member, National Academy of Medicine.

Contents

1

Introduction

The Intelligence Community Studies Board (ICSB) of the National Academies of Sciences, Engineering, and Medicine convened a workshop on December 11–12, 2018, in Berkeley, California, to discuss robust machine learning algorithms and systems for the detection and mitigation of adversarial attacks and anomalies. With funding from the Office of the Director of National Intelligence, the ICSB established a Planning Committee on Ensuring the Quality of Machine-Generated Analytic Products from Multi-Source Data: A Workshop (biographical sketches provided in Appendix A) to develop the workshop agenda (see Appendix B). The workshop statement of task is shown in Appendix C.

Workshop speakers and participants discussed research challenges related to the following topics:

- Critical analysis of the current state of machine learning and artificial intelligence (AI) algorithms and systems that are used to generate analytic products from disparate structured and unstructured data types and to detect anomalies;
- Statistical methods that can be used to evaluate confidence hierarchies, model uncertainty, and error propagation, and manage risk as a function of time and complexity;
- Approaches for ensuring that machine-generated products compare favorably with those of trained human analysts; and
- Techniques for responding to adversarial manipulation of input data to influence analytical products by exploiting weaknesses in machine learning and AI algorithms and vulnerabilities in their implementation.

During the presentations and discussion sessions, attendees were asked to address the following questions, with particular emphasis on their role for the Intelligence Community:

- What are the key technical objectives and performance measures needed for success?
- What are the current and "next level" key performance metrics?
- What is the "level after next" of expected research and development performance?
- What is the research knowledge base?
- How can the government best prepare the scientific workforce to enhance discovery in this area?
- What are the requisite enabling technologies?

This proceedings is a factual summary of what occurred at the workshop. The planning committee's role was limited to organizing and convening the workshop. The views contained in this proceedings are those of the individual workshop participants and do not necessarily represent the views of the participants as a whole, the planning committee, or the National Academies of Sciences, Engineering, and Medicine.

2

Plenary Session

SPONSOR REMARKS AND EXPECTATIONS OF THE WORKSHOP

David Isaacson, Office of the Director of National Intelligence
Rama Chellappa, University of Maryland, College Park
George Coyle, National Academies of Sciences, Engineering, and Medicine

After welcoming participants and expressing his gratitude to the workshop planning committee, David Isaacson, Office of the Director of National Intelligence, explained that this 2-day workshop would present technical advances relating to adversarial attacks and anomalies, which have high-level national security implications. He said that Congress has taken notice of these technical advances, especially in light of the prevalence of fake videos and images created using machine learning techniques. Deepfake technology blurs the lines of fact and fiction and could undermine public trust in recorded videos and images, Isaacson explained. This increased presence of false information is accompanied by a volume of available digital information that exceeds the Intelligence Community's (IC's) current capabilities for human vetting and processing. Isaacson hoped that this workshop would present the current state of the art as well as provide insight into forthcoming innovations so that the IC and the nation are prepared to retool and be better equipped in both the current environment and future scenarios. Rama Chellappa, University of Maryland, College Park, and George Coyle, National Academies of Sciences, Engineering, and Medicine, noted that, to help guide the IC's future technology investments, workshop speakers would identify both near- and long-term enabling technology capabilities (see Appendix D).

ON COMPUTATIONAL THINKING, INFERENTIAL THINKING, AND DATA SCIENCE

Michael I. Jordan, University of California, Berkeley

The rapid growth in the size and scope of data sets in science and technology has created a need for novel foundational perspectives on data analysis that blend the inferential and computational sciences. That classical perspectives from these fields are not adequate to address emerging problems in data

science is apparent from their sharply divergent nature at an elementary level—in computer science, the growth of the number of data points is a source of "complexity" that must be tamed via algorithms or hardware, whereas in statistics, the growth of the number of data points is a source of "simplicity" in that inferences are generally stronger and asymptotic results can be invoked. On a formal level, the gap is made evident by the lack of a role for computational concepts such as "runtime" in core statistical theory and the lack of a role for statistical concepts such as "risk" in core computational theory. I present several research vignettes aimed at bridging computation and statistics, discussing the problem of inference under privacy and communication constraints, the problem of the control of error rates in multiple decision-making, and the notion of the "optimal way to optimize."

MACHINE LEARNING ON PERCEPTION: HYPE VS. HOPE

Ruzena Bajcsy, University of California, Berkeley

Ruzena Bajcsy, University of California, Berkeley, discussed model- versus data-driven analysis and prediction in the context of human movement. She first provided an overview of the Human-Assistive Robotic Technologies Laboratory's[1] human modeling and presented a series of important definitions.

- *Accuracy*. The difference between the measurement and the actual value.
- *Precision.* The variation observed when measuring the same part repeatedly.
- *Reliability*. The degree to which a test or an instrument consistently measures.
- *Reproducibility*. A component of the precision measurement or test method.
- *Repeatability*. The degree of agreement between tests.
- *Robustness.* The ability of a system to tolerate perturbations that might affect the system's functional body.
- *Measurement*. The mapping of qualitative empirical relations to relations among number, which is only a reflection of reality.
- *Data.* Processed measurement.

Bajcsy explained that understanding limits—accuracy, precision, sensitivity, detectability—of measurements is important, especially when analytics are used to make life or death decisions and in the design and implementation of safety-critical systems. It is essential that systems are safe and that their behavior is repeatable, reliable, stable, and robust. It is also important to provide a system user with guarantees and limits of system performance. To do this, one must know the limits of the sensors that provide measurement.

Bajcsy provided examples of the use of visual sensors in research. The first example was a collaboration with orthopedic surgeons on a study of postural stability in post-surgical sit-to-stand transitions. There is a high rate of failure in subjects undergoing spinal fusion, and this project aimed to shed light on when the surgery would not work based on biomechanical changes. Her team sought to identify kinematic, dynamic, and muscular changes pre- and post-spinal fusion surgery, as well as effects on balance and standing strategies. By understanding the limits of measuring devices, it was possible to detect the individual sit-to-stand transitions (Matthew et al., 2018). However, she noted that simplifying assumptions were needed, including that sit-to-stand could be modeled solely in the sagittal plane and that an allometrically scaled musculoskeletal model could represent the subject.

This research utilized three-dimensional cameras (Kinect 1, Kinect 2) and several different types of modeling methods including inverse kinematics, inverse dynamics (i.e., an estimate based on a patient's overall height and weight), and muscular model. The results of the research were used to inform decisions about surgical interventions and have shown a positive improvement. However, Bajcsy emphasized that

[1] To learn more about the Human-Assistive Robotic Technologies Laboratory at the University of California, Berkeley, see http://hart.berkeley.edu/, accessed February 19, 2019.

decisions about medical diagnostics are inherently personal, and individual situations may vary. She also stressed that differences in models can dramatically impact the data analysis results; accuracy is particularly important in healthcare decision-making. While this example focused on diagnostic devices, the future goal is to develop techniques for intervention.

The next study that Bajcsy discussed was on gross and fine motor synergies in reach-to-grasp movements, which is a work in progress. The objective of the work was to incorporate biomechanical models of function and impairment into individualized, assistive control schemes. To do this, Bajcsy's team built a robotic system for assessment and assistance of the planar reach-to-grasp motion. The system controls wrist position with elbow support and an assisted pinch grasp, measures the position and force at the wrist, and operates under admittance (i.e., force-velocity) and position control modes. This system illustrates that smoothness of the motion varies from healthy subjects to those who have had strokes. This concept relates to the ability for coordination of human–machine movement. The preliminary results of control and upper limb impairment subjects (i.e., stroke patients) showed differences in linearity and smoothness of trajectory, velocity profile, and ability to stabilize.

Another study considered enabling multi-degrees-of-freedom musculoskeletal dynamics models via ultrasound and acoustic myography. Unlike the other studies, which addressed full body analysis, this study began to address human dynamics more locally to better understand the strength of the muscles. Ultrasounds help to better understand the mechanical properties of the muscle because they give a reasonable spatial, temporal resolution and show cross-sections of muscles, Bajcsy explained. The objective of the study was to construct low-dimensional models mapping noninvasive sensor data to the in vivo muscle force of the arm. Bajcsy's team observed substantial deformation under changes in muscle load and kinematic configuration and is currently employing statistical shape analysis to extract low-dimensional deformation models.

Bajcsy commented that model-based approaches for human modeling research are useful because interpretability is critical for medical and human-interfacing applications, many of which are safety-critical and in need of guarantees. Thus, it is essential that the medical community understands, or at least trusts, the system. Domain knowledge is plentiful and important to scoping models appropriately; while models are also abstractions, they serve as a guide about what to measure. She said that biology obeys constraints (e.g., laws of physics, musculoskeletal attachments and interactions) but that there are many niche applications with their own interests and constraints—specific tasks (e.g., sit-to-stand, sports) and specific pathologies (e.g., stroke, spinal cord injury). It is critical to map the results to the existing body of literature because model-free approaches alone do not allow for easy incorporation of constraints and do offer limited interpretability. Big data are useful in exploratory analysis because they can be used to establish a healthy baseline for a particular task or measurement to classify pathology or activity or configuration. They can also be used to perform high-level analysis (e.g., extract through principal component analysis) to inform the development of future models, she continued, particularly for non-critical but time-consuming tasks. Data-driven approaches work best when the scope is limited and the training data sufficiently cover the space of the inputs.

In conclusion, Bajcsy said that model- and data-driven approaches are complementary for modeling human movement, although it is critical to determine the right tool for the right job. She reiterated that no two humans are alike and highlighted the promise of precision medicine that could provide more customized diagnostics and treatments. She highlighted several important questions to consider when thinking about human modeling, including the following:

- Are we interested in behavior/performance or the system structure to predict behavior at different scales?
- What simplifying assumptions can be made?
- What trade-offs must be made between accuracy/performance and generalizability?
- How can we provide guarantees for robustness and safety? Even if we consider techniques such as reachable sets in control systems, the boundaries depend on the accuracy and precision of the estimated parameters of the dynamical system, so it is important to pay attention to the estimators.

She closed by explaining that deep, fundamental results take time to achieve. In response to a question from an audience participant, Bajcsy said that her planar reach-to-grasp systems are cost-conscious as well as robust and easy to use. As an example, she explained that children with muscular dystrophy have decreased ability to reach as the disease progresses (i.e., children will not be able to comb their hair or feed themselves). She gave her reachability software to colleagues to use. Although it takes time to get approval from the Food and Drug Administration, she and her team have been fairly successful in transferring technology. Devices for estimating kinematics can be bought off the shelf (e.g., motion capture systems), and, in the case of the "sit-to-stand" experiment, such a device/system is now being used in at least two orthopedic departments. She emphasized that the role of academia is to develop ideas and demonstrate feasibility, and industry can take it from there.

3

Adversarial Attacks

MEDIA FORENSICS

Matthew Turek, Defense Advanced Research Projects Agency

Matthew Turek, manager of the media forensics "MediFor" program at the Defense Advanced Research Projects Agency, provided an overview of how the program detects manipulations of media assets, such as image and video. He started the presentation by giving the background of the program and by telling the MediFor story by showing a slide with several images and the question "Do you believe what you see?" He highlighted mantras such as "a picture is worth a thousand words" and "seeing is believing" that make up the current visual media narrative and noted that it will need to evolve. Turek then described several slides containing images that were either authentic or manipulated.

It is useful to understand if and how an image has been manipulated and be able to describe how assets could be related. Turek noted that the manipulation of media assets is not a new problem, but the dramatic increase of digital content from a host of sources such as social media, YouTube, and surveillance video over the past 20 years has changed the landscape. The changes associated with the rise of digital media include the ability to manipulate media assets with less skilled human intervention, effort, and resources. He pointed out that although an individual with a gaming-level personal computer can generate a convincing manipulation, we are not at the point where it could be done reliably and with targeted intent—but we are heading in that direction.

Prior to MediFor, there was the problem of not being able to assess the integrity of image and video assets at scale, even with sufficient expertise. The majority of assessments were manual, and there is more software available to create manipulation than detect it. He hopes that MediFor will raise awareness in the research community to develop software and techniques for automatically detecting manipulations in media.

To highlight the process of manually assessing an image, he showed the example of a manipulated image of a hovercraft landing and described how one can use the manual indicators such as the wakes, shadow consistency, and sun angle to determine the integrity of the image. Turek described the following underlying challenges of assessing manipulation: (1) There is not sufficient time or expert personnel to

manually analyze individual images or videos, and (2) no one technique will work. He then pointed out that there is a need for broad-based capabilities for detecting manipulations in media.

MediFor detects media manipulations by looking at three essential elements of media integrity: digital integrity, physical integrity, and semantic integrity. He pointed out that semantic and physical integrity indicators are needed because digital integrity indicators are strong but can break down under certain circumstances—e.g., high levels of compression. Turek described a fourth element, integrity reasoning, which combines the digital, physical, and semantic integrity elements into one integrated assessment. The MediFor system was created to conduct the integrated assessment of all of the digital, physical, and semantic analytics and indicators and ultimately produce an integrity report. The integrity report consists of the media integrity indicators, an integrity score (composed of each of the media integrity indicators), and the integrity basis (e.g., camera geometry). The MediFor system also has a console where the user or an analyst can interact or upload different media assets.

Turek described important digital, physical, and semantic integrity and integrity reasoning indicators. Each description consisted of the underlying questions, examples of manipulation detection, the previous state of the art, work of the MediFor researchers, and challenges.

- *Digital integrity indicators.* The underlying questions of digital integrity are as follows: Are the pixels/representations inconsistent and is the metadata consistent? Examples that indicate manipulations include blurred edges from object insertion, different camera properties, replicated pixels, and mangled compression. MediFor researchers' work includes looking for color-shifts, copy-paste, and composites. Turek highlighted that one researcher observed that computers can pick up on low-level signals that humans cannot, such as rounding differences from different implementations from compression. Prior to MediFor, manipulations were not classified or characterized in detail; this is a manually intensive process and is hard to analytically combine evidence from multiple sources. However, the ability to classify real versus simulated (e.g., deepfakes) and composite imagery is challenging. Currently, MediFor researchers are gathering evidence to determine how an image or video was manipulated and where the manipulation occurred.
- *Physical integrity indicators.* The main underlying question of physical integrity is as follows: Are the laws of physics violated? Examples that indicate manipulation include inconsistent shadows and lighting, elongation/compression, and multiple vanishing points. Previously, the majority of the detection technology was focused on the digital level, and processes were manual and time consuming. MediFor analyzes shadows and highlights and scene dynamics such as abrupt inexplicable changes in motion. Turek highlighted the difficulty that MediFor researchers have working on different aspects of reflection analysis such as automatic detection, understanding the geometry, and consistency. The automatic detection of shadows and reflection and scene consistency remains challenging.
- *Semantic integrity indicators.* The main underlying question of semantic integrity is as follows: Are the hypotheses about a visual asset supported or contradicted? Questions that help determine manipulation include the following: Is there contradicting evidence in associated images? Was the media asset repurposed? Are dates, times, and locations verifiable? Examples of work by MediFor researchers are combining information of different assets and media provenance. Previously, association and classification would only work well with dense coverage and unique features such as text. MediFor researchers are working on the difficult problem of media provenance to include the analysis of fewer images, less overlap, and subtle image cues. Challenges include verifying spatial temporal assumptions such as the location of an image on Earth at a particular time, uncovering provenance of an asset, and analyzing and space time aspects of video.
- *Integrity reasoning indicators.* The underlying question of integrity reasoning is "How can one assimilate digital, physical, and semantic integrity indicators into an integrated assessment?" Previously, most measurements were taken manually, were subjective, and had few quantitative

standards. MediFor researchers are working on fusing integrity indicators using contextual information because algorithms are looking for specific elements. Establishing confidence measures for various analytics and combining scores of different integrity modules are challenges.

The MediFor program collects data from annual challenges with the goal to target particular applications. The data set was comprised of a few thousand images at the initiation of the program. The National Institute of Standards and Technology, an evaluator of the MediFor program, conducted the nimble challenge in 2017. The data set contained 13,500 images, 1,151 videos, and focused on problems such as overhead imagery, commercial satellite data, and images embedded in scientific documents. Turek highlighted the auto journaling tool, which enables the ability to capture the history of the manipulations in a graphical format. Some of the journaling tool steps can be automated. In 2018, the media forensic challenge data set included 151,200 images and 3,628 videos. Problems extended to computer-generated imagery, green-screen manipulations, Photoshop manipulations, audio manipulations, and generative adversarial network (GAN)-altered images and video—the technology that underlies deepfakes. Future challenge problems will include the camera ID challenge and event recognition. The camera ID challenge will identify methods to tell if an image was taken by a particular cell phone; event recognition is the ability to sort images into known events. Turek suggested these capabilities will assist provenance, the ability to understand the content of images and video.

Annual evaluations comprehensively assess the MediFor program capabilities. The evaluations focused on image manipulation detection in 2017. Evaluations in 2018 were scaled to include video manipulation detection and image manipulation detection and localization that could enable detection of pixels that have been manipulated in an image. Turek explained that future evaluations may include work from the challenge problems and extend to video manipulation localization (both temporally and spatially), which could enable identification of specific ranges of frames that have been manipulated.

Turek presented a series of receiver operating characteristic[1] (ROC) curves to show the quantitative results of the MediFor program. He assessed media forensic challenge algorithms using nimble challenge data (4,000 images) on a variety of manipulations to produce the ROC curves. Results show significant improvement in detecting image manipulations over time. Turek also introduced ROC curves that use the opt-in feature, which allows algorithms to choose images to assess based on the features they were designed to evaluate. He explained that the fusion of media integrity indicators results in a significant performance gain. Turek noted that detection of video manipulations is a challenge.

Turek presented the GAN challenge that looked at manipulation detection and localization task on modified data. The challenge used three data sets: the full image set, containing images partially altered by GANs; the crop image set, which are completely generated by GANs or are real; and the video set, including deepfakes. Turek showed the full score and opt-in ROC curves for the full and crop image data sets. He highlighted that only one researcher had an ideal result if the data were fully manipulated by GANs.

Turek showed a GAN video example of the face swap, when one person's face is posted onto another person. He described that in order to have a convincing face manipulation, the faces of the original subject and manipulated subject should have similar characteristics. The characteristics described include similar face shapes, features, skin color, lighting, and background. The MediFor research groups observed that automated manipulations work best when the training and target data are well aligned. Turek showed the ROC curves with full scoring and opt-in feature and noted that the results showed credible performance levels and the need to reduce the false-alarm rate. Turek believes that future manipulations will become more robust and shared, promising unpublished work from a group that hopes to attribute specific GANs to specific manipulations.

In response to a participant's assertion that a manipulation-detection arms race exists, Turek agreed and added that the MediFor program has been open and publishing the results of its work. He further emphasized

[1] The receiver operator characteristic curve plots the true positive rate of detection of media manipulation against a false positive rate. Ideal results have a true positive rate of 1 and false positive rate of 0, which is the upper left corner of the plot.

that defenses consist of a broad range of capabilities—the digital, physical, and semantic layers—and the burden will be on the manipulators to access sufficient training data to produce a convincing manipulation. He also remarked that MediFor is investigating how easy it would be to take a detector and build that into the GAN framework and noted that GANs are difficult to optimize and train. He also responded to a question about whether downstream editing or other tools degrade sensor signatures; Turek responded yes, but sometimes they leave their own "fingerprints." He said that MediFor researchers are investigating how to identify software packages that have manipulated images or video. He added that the camera ID can often be recovered, even under certain levels of compression and manipulation; however, it does degrade with the removal of the noise artifacts.

Another participant asked why there is a big drop in the accuracy for detection of manipulation of videos. Could the same technology that works well for images be applied to videos? Or, is it a scalability or computational complexity issue? Turek responded that video detection capabilities are newer, less robust, and less mature.

FORENSIC TECHNIQUES

Hany Farid, Dartmouth College

Hany Farid, Dartmouth College, explained that owing to the power of visual media, serious issues of media authenticity are prevalent. His presentation focused on three forensic techniques that can be used to determine the authenticity of digital videos. First, he shared a video of a man sliding off a ramp, flying through the air, and landing in a pool. In order to determine the video's authenticity, the physics of the scene has to be understood. Farid explained how to reason about that type of motion in a video. Based on what is known about ballistic trajectory, once something is launched in the three-dimensional (3D) world, it follows a parabola (assuming no significant effects of wind or other external forces). While this is evident in a two-dimensional (2D) video if the camera capturing the scene is perpendicular to the plane of motion, understanding the physics of the motion becomes more challenging if the camera is positioned differently, at an angle, because the motion in the 2D video will no longer be a parabola. However, assuming the camera is static, one can parameterize the projection of a 3D parabola into a 2D camera and determine if the trajectory of a purported ballistic motion is consistent with the physics of a 3D projectile. If the motion is physically plausible then the video is likely authentic (see Figure 3.1a); otherwise, it might be manipulated (see Figure 3.1b) (see Conotter et al., 2012). Although the initial assumption was of a stationary camera, Farid noted that these techniques could also be applied with moving cameras. Although this technique applies to a narrow set of videos, Farid said that researchers are developing a number of different forensics that each apply to a small number of cases, but combined they can be used to analyze a broad range of videos.

In the next example, Farid showed a video in which an eagle appears to swoop down and grab a child in a park. He explained that this motion is much more complex than the ballistic motion discussed in the previous examples because the motion of an eagle cannot be easily modeled. However, with outdoor images, reliable information exists about the lighting since the Sun is the dominant light source. To determine the authenticity of this video, one can reason about the physical plausibility of shadows in the video. In the 3D world, there is a linear constraint among the point on a shadow, the corresponding point on an object, and the light source. This constraint holds in each 2D frame of a video. As shown in Figure 3.2, the linear constraints in blue for five objects in the scene are inconsistent with the constraints in purple for the eagle and the child. The eagle and the child are computer generated and were digitally inserted into a live-action scene.

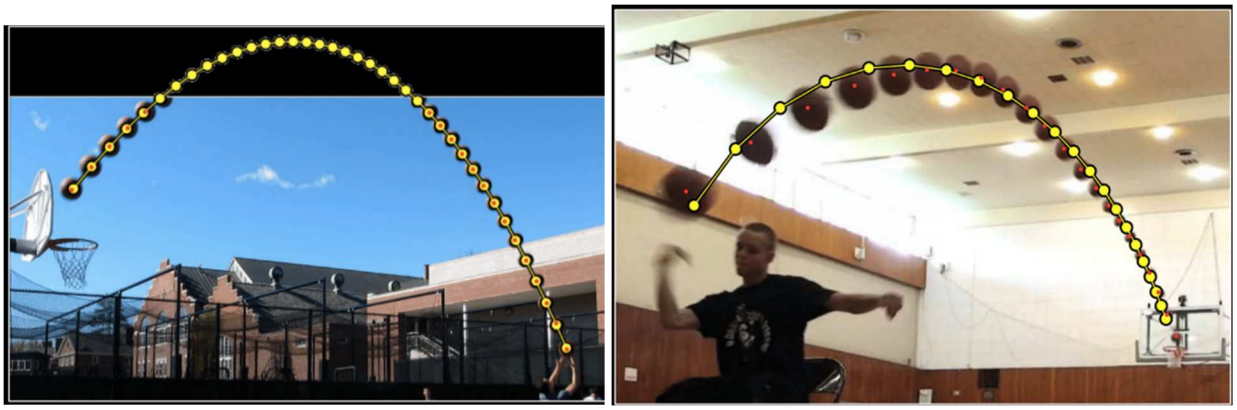


FIGURE 3.1 In the image on the left, the predicted trajectory (yellow) aligns with the tracked position (red), but in the image on the right, the predicted trajectory (yellow) does not align with tracked position (red). SOURCE: Hany Farid, Dartmouth College, presentation to the workshop, December 11, 2018.

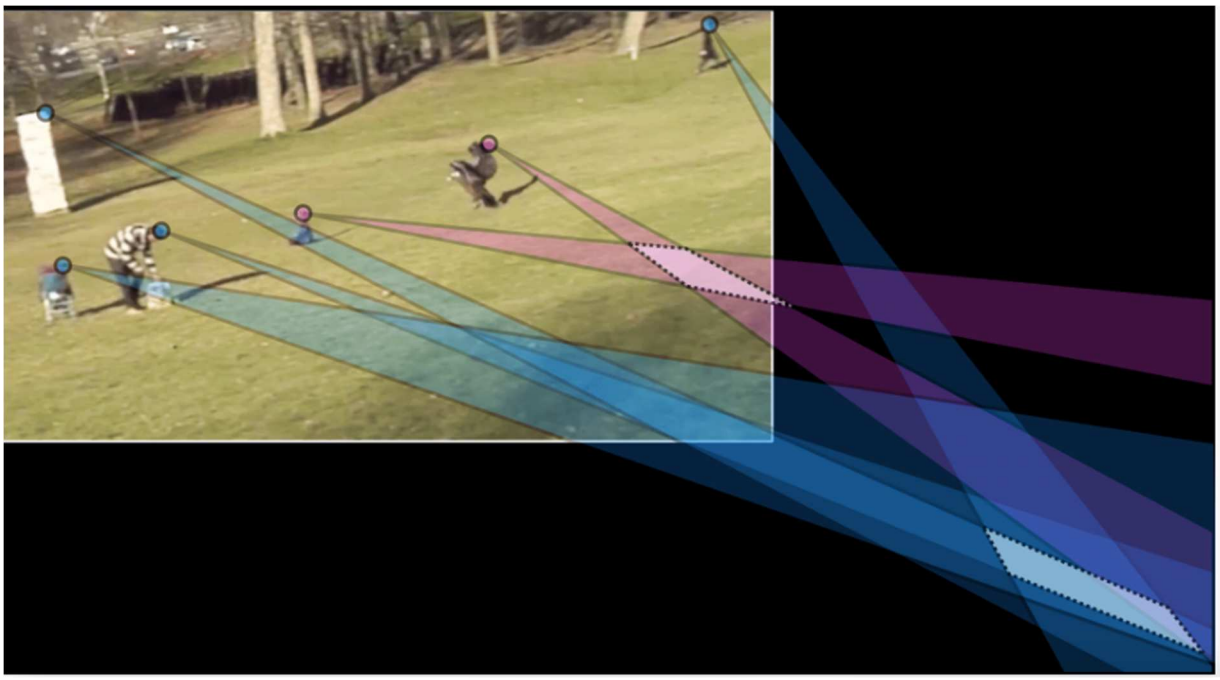


FIGURE 3.2 A study of the light and shadows in this scene confirm that the action is fake. SOURCE: Hany Farid, Dartmouth College, presentation to the workshop, December 11, 2018.

Farid pointed out that if adversaries understand the techniques for manipulation identification, they can create more authentic-looking imagery. This growing capability for manipulation is a societal concern.

In the final example, Farid showed a video in which former President Obama appears to be talking, but in reality actor Jordan Peele is talking. This is an example of a deepfake. In this instance, everything in the video of Obama is real, with the exception of his mouth, which was altered through use of deep neural networks to be consistent with Peele's speech patterns. Making videos like this could disrupt presidential elections and could give a politician plausible deniability in any video released that shows him/her in a poor light.

To detect this sort of fake, Farid and his team rely on soft biometrics. There is a relationship between what a person says and how he/she says it (e.g., in terms of facial expressions or head movements). Looking at these correlations and then building a probabilistic model reveals which video is likely to be authentic

and which is likely to be fake. His team is working on building these models for all individual world leaders, though they do not yet know if they will generalize to other people. Farid suspects the models will generalize, because human faces do not move in completely random ways during speech.

Although many good techniques exist to detect manipulation, Farid emphasized that there is still progress to be made in the near and long term. Many of the current techniques require a human in the loop, which does not scale. More automated processes are needed, as are faster approaches to handle the large volume of data available. Improved accuracy is also desirable; he noted that 99 percent accuracy is actually terrible because it indicates that 1 out of every 100 fake videos on the Internet is undetected. Secure-imaging pipelines, as developed by companies such as Truepic,[2] could help address this problem, Farid explained. Such technology could be especially useful for law enforcement agencies because it can prevent the manipulation of digital evidence. Farid even foresees that options for secure and unsecure photography will eventually be built into camera apps. Although there are still vulnerabilities with that approach, such as with a rebroadcast attack where a manipulated image is re-imaged to avoid detection, the risk could be more manageable. Farid asserted that better cooperation from social media companies is needed to reduce deepfakes. He believes that social media platforms have been weaponized and that most have not been responsive enough to misinformation campaigns on their platforms. He added that citizens have to become more informed in the digital age and that guidelines are needed for responsible deployment of technologies.

In response to an audience participant, Farid said that the spread of misinformation extends beyond a data ownership problem. U.S. citizens have to consider the appropriate balance between individual rights and societal harm, perspectives about which vary across the world. For example, Germany enacted aggressive legislation in an attempt to stop the spread of misinformation. In response to a comment from Ruzena Bajcsy, University of California, Berkeley, Farid said that U.S. citizens have given up privacy with social media without thinking carefully about it and are now trying to backtrack on those decisions. He added that it is preposterous that the majority of Americans take in their news from Facebook, a platform whose corporate interests may not align with users' interests. For these two reasons, it is crucial to better educate people. A workshop participant raised a question about the right approach toward open-source code and research. Farid responded that he has started to question his beliefs about publishing work openly and suggested that holding back code is a good compromise because new technologies can be assimilated into generative adversarial networks quickly. An audience participant expressed concern about finding the right balance with so many competing objectives in maintaining the integrity of communication, while another audience participant cautioned that too much obscurity of information could lead to an inability to see what our adversaries are doing.

[2] For more information about the company Truepic, see https://truepic.com, accessed April 3, 2019.

4

Detection and Mitigation of Adversarial Attacks and Anomalies

USING AI FOR SECURITY AND SECURING AI

Joysula Rao, IBM Corporation

Joysula Rao, IBM Corporation, opened his presentation by explaining that security attacks have been prevalent throughout the past 2 years.[1] New attacks use disruptive technology to create devastating results, and attackers are the first to exploit new technologies and platform shifts (e.g., to the cloud) to launch sophisticated attacks at scale. Those in the position of defense are always trying to catch up in order to protect the technologies. With artificial intelligence (AI), Rao continued, attacks are getting smarter. For example, Deeplocker is a type of AI-powered malware that avoids detection by most security controls in place today. With the cloud, attack surfaces are increasing in size and becoming more sophisticated. These systems were not originally designed to prevent micro-architectural attacks. And with the Internet of Things, attack targets have become physical—for example, vehicles, medical devices, and navigation systems. Automation enables attack campaigns to increase in speed and has contributed to the rise of both devastating malware and "fake news." Rao hypothesized that attackers will use AI to launch much more sophisticated attacks, at scale and on more targets, which will be more evasive. He said that it is difficult to anticipate what kind of AI tools attackers will use, so preparation for the future must begin now.

Rao explained that cybersecurity and AI is a dual problem, when considering the roles of both attackers and defenders (see Figure 4.1). Thinking about this dual problem enables one to better secure AI and to develop better security. AI can be used to bolster defenses and proactively disable AI-powered attacks. It is also necessary to counter adversarial AI and protect against adversarial environments, theft of data and models, corruption, and evasion.

[1] For more information on the sampling of security incidents, see IBM Corporation, "X-Force Threat Intelligence Index," https://xforceintelligenceindex.mybluemix.net/documents/IBM-X-ForceThreat-Intelligence-Index-2019.pdf, accessed August 7, 2019.

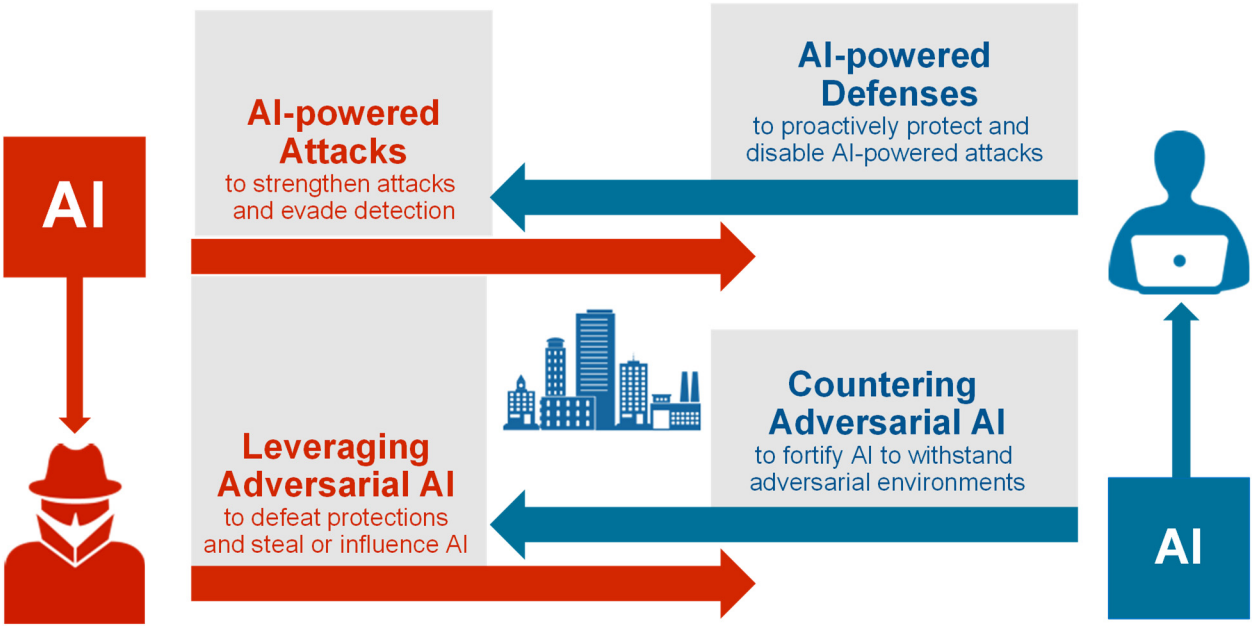


FIGURE 4.1 The dual problem of artificial intelligence. SOURCE: Joysula Rao, IBM Corporation, presentation to the workshop, December 11, 2018.

Rao noted that it is common for attackers to use AI and reinforcement learning techniques to craft emails that avoid enterprises' spam filters—similar to the work that has been done to mutate malware. In the Defense Advanced Research Projects Agency's Cyber Grand Challenge,[2] machines are used to find vulnerabilities of other machines and exploit them. Rao expects the future will be similar: AI versus AI. He next described Deeplocker, an example of an AI-embedded attack, which embeds a neural net within a piece of malware. A secret key within the neural net identifies certain attributes of a target and then embeds a concealed encrypted payload into software. Rao pointed out that no known antivirus software or other types of controls will currently pick up on such a targeted attack. This malware is equipped with the capability to show up only on the target's machine. This kind of targeted attack is extremely difficult to predict and protect against and will likely become more prevalent in the near future. Rao commented that the following series of techniques could be used to defend against AI-powered malware:

- Feature extraction and pattern recognition to improve decision making and detect unknown threats;
- Natural language processing to collect text on past and current breaches, consolidate threat intelligence, and increase security knowledge;
- Use of reasoning to locate evidence of breaches, remediate planning and outcomes, and anticipate new threats and next steps; and
- Automation of tasks to reduce the burden on the human analyst and decrease reaction time.

AI can also be used to address the following types of security needs:

- Improving modeling of behaviors to better identify emerging and past threats and risks. Relevant applications include the network, user, endpoint, app and data, and cloud. The examples are

[2] To learn more about the Cyber Grand Challenge, see Defense Advanced Research Projects Agency (DARPA), "Cyber Grand Challenge (CGC) (Archived)," https://www.darpa.mil/program/cyber-grand-challenge, accessed July 19, 2019.

beaconing detection, network anomaly detection, domain-name system analytics, and user behavior analytics.
- Consolidating intelligence sources, including structured and unstructured data sources. An example of this approach is Watson for Cyber Security.
- Integrating feedback from trusted advisors who can provide context and assess analyses. The applications are cognitive security operations center analyst, orchestration, automation, and digital guardian. An example is the Cognitive Threat Insights Platform.

Rao alluded to the success of image recognition; around 2015, human classification errors exceeded ImageNet's.[3] However, it is important to strive to improve further. He provided an example of a picture of an ostrich, which was labeled variously by different image recognition systems as a shoe shop, a vacuum, a safe, and an ostrich, to demonstrate that models are brittle and easily breakable. In cases of adversarial AI, Rao explained that problems arise not only with images but also with text and audio recordings. He explained that if predictions are inaccurate or systems are not understood completely, tragedies could occur, such as the self-driving Uber vehicle responsible for the death of a cyclist. In response to a question from an audience participant, Rao noted that these are all examples of human–machine issues.

He explained that people have been using different norms (e.g., L_1 and L_2) to represent the distance of pixel changes and therefore identify changes in images and adversarial attacks. L_2 norms do not reflect human perception; the machine can get easily confused and misclassify an image. Similar issues arise with malware. He said that it remains unclear what the right approach is to protect against adversarial attacks.

Privacy is another key issue in security. Rao noted that model inversion and membership inference could be useful and adversaries can learn a representative sample from a class as well as other sensitive information. Differential privacy, which protects individual privacy, does not address privacy protection in the aggregate.

Trained models can be poisoned, and their accuracy or performance decreased, through the insertion of a malicious payload or trigger. Adversaries can generate "backdoors" into neural networks by poisoning the training data. This type of manipulation is difficult to detect because the models perform well on standard training and validation samples but behave badly on backdoor keys. State-of-the-art detection methods require trusted training data. Backdoor detection using activation clustering is one way to protect against poisoning attacks.

He shared a historical overview of various defenses to adversarial techniques and how they were subsequently not robust. He began with models and defensive distillation; moved to other methods to add adversarial noise such as feature squeezing, principal component analysis, and adversarial training; and on to MagNet (see Meng and Chen, 2017, for example), and BReLU (see Liew et al., 2016, for example) and Gaussian. Moving forward, Rao said that some level of accuracy has to be reduced in order to achieve robustness, and some promising techniques are emerging. Rao reiterated that a duality exists between the attacker and the defender. With the application of a watermark in a model's training data that alters the final class, it is possible to leverage the ability to poison models, with negligible impact to the accuracy of the model. In this case, the adversary will not have access to the training data, and the watermark can be much more subtle, specific images.

Rao described IBM's open-source Adversarial Robustness Toolbox, which is continually updated with attacks and countermeasures; this can be a resource to developers of AI services. A workshop participant asked if a balance exists between research work in adversarial attacks and adversarial defenses and whether there is a reward system for doing one over the other. He noted that at least in academia, more publication opportunity and credit seems to be given to papers on attacks. Rao said that research in adversarial AI has gained momentum in the past 2 years, and that is reflected in the rate of publications. Rao agreed that although the attack papers are gaining more attention, the hard work is building systems that are secure against such attacks; some of the best minds are working on defense. A workshop participant disagreed and

[3] For more information on ImageNet, see http://www.image-net.org/, accessed February 19, 2019.

said that there is a huge bias toward publishing defense papers; many have been published despite their weak solutions and disingenuous messages. He emphasized that the need is for more effective attack papers.

Ruzena Bajcsy, University of California, Berkeley, asked Rao to explain his definition of AI, because it seemed to her that he was describing search and pattern recognition instead. Rao said that when industry talks about AI, it really refers to *augmented intelligence* (i.e., pattern recognition, machine learning, data mining, natural language processing, human–computer interfaces, and reasoning). He added that the goal is to offload the cognitive overhead that subject-matter experts would have and automate that with a machine, which is not general purpose AI. A workshop participant suggested that AI is practiced on an extremely narrow set of tools today, such as deep learning, which has many opportunities to fail. Rao agreed that many AI problems remain open and, even in narrow domains, deep learning may not be the right approach. Rama Chellappa, University of Maryland, College Park, said that no one is saying AI is working and added that deep learning should not be called AI because it has no statistical backing. However, Rao concluded that the attackers are not going to wait for the community to resolve these issues.

CIRCUMVENTING DEFENSES TO ADVERSARIAL EXAMPLES

Anish Athalye, Massachusetts Institute of Technology

Anish Athalye, Massachusetts Institute of Technology, opened his presentation by acknowledging that machine learning has enabled great progress in solving difficult problems in recent years (e.g., super-human classification, object detection, machine translation, game-playing bots, self-driving cars testing on public roads). Much of this has been enabled by deep learning. He elaborated that although machine learning systems achieve great "average-case" performance on these difficult problems, machine learning is fragile and has terrible performance in "worst-case" situations, such as adversarial or security-sensitive settings. An audience participant said that machine learning is performing well in test cases and wondered how Athalye defined average. Athalye responded that average performance indicates a level of comfort with deployment.

Athalye explained that imperceptible perturbations in the input to a machine learning system could change the neural network's prediction and dramatically affect the model and its output. A small adversarial perturbation can lead a state-of-the-art machine learning model to mislabel otherwise identifiable images. These attacks can be executed with just a small number of steps of gradient descent. An audience participant wondered whether when looking at the difference between the original image and the adversarial example and magnifying noise if it is possible to see anything interpretable. Athalye responded that while noise may be more interpretable with more robust machine learning models, in these cases only noise is perceptible.

Athalye described machine learning as not being robust for image classification and for many other problems such as semantic segmentation, reading comprehension, speech-to-text conversion, and malware detection. Attackers can tweak metadata to produce functionally equivalent malware that can evade a detector. An audience participant wondered if training models using adversarial examples could help build in a resistance to attacks. Athalye noted that this process has been shown to increase robustness to some degree, but current applications of adversarial training have not fully solved the problem, and scaling is a challenge.

Athalye noted that contradictory evidence exists as to whether real systems are at risk from adversarial examples. He described a study in which natural image transformations were applied to adversarial examples, thus breaking the adversarial example and classifying the image correctly in its true class. However, adversarial examples can be robust, and this approach does not always work.

He presented a basic image-processing pipeline in which an attacker gains control of an image, the image is fed into a machine learning model, and the resulting predictions are affected. However, the physical-world processing pipeline looks a bit different: a transformation with randomized parameters occurs between the image and the model. In this case, the attacker no longer has direct control over the model input (see Figure 4.2).

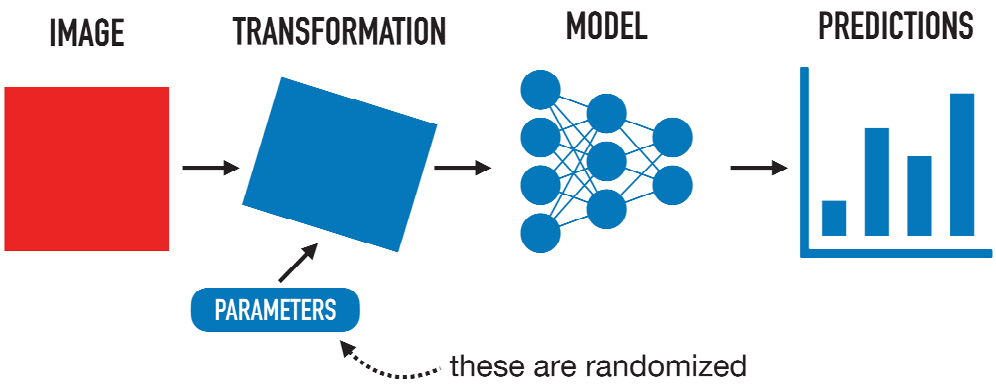


FIGURE 4.2 A physical-world image processing pipeline. SOURCE: Anish Athalye, Massachusetts Institute of Technology, presentation to the workshop, December 11, 2018.

An attack is still possible in this real-world setting because, even though one does not know what the exact transformation will be, the *distribution* of transformations is known. The transformation needs to be differentiable, he continued, and instead of optimizing the input to the model, one can optimize over all possible transformations to find a single point that no matter how it is transformed will confuse the model in all settings. This approach (i.e., expectation over transformation), which still uses gradient descent, can produce real-world robust adversarial examples.

Athalye then considered a three-dimensional (3D) processing pipeline, which is similar to the real-world pipeline. He explained that for any pose, 3D rendering is differentiable with respect to texture. He demonstrated a 3D adversarial object where no matter how the object is rotated, the machine learning model still classifies it incorrectly. In response to an audience participant, Athalye confirmed that, in this example, only the texture of the object was changed. He said that other researchers have also tested manipulating geometry rather than texture to construct adversarial objects. Athalye reiterated that machine learning is not robust in controlled and real-world settings, noting that even in a world filled with noise, the models can still be fooled.

Black-box models, too, are susceptible to adversarial attacks, Athalye explained. In a black-box threat model, the attacker has no visibility into the details of the model—the attacker can only construct an image, feed it in, and watch what emerges. Gradient descent is still used, but now so too is an estimate using queries to the classifier. Even more restricted settings, such as the Google Cloud Vision Application Programming Interface, are vulnerable to adversarial attacks, Athalye continued. He noted that in the hundreds of papers published on defenses for adversarial attacks, many of the proposed defenses lack mathematical guarantees. He added that many defenses submitted to the 2018 International Conference on Machine Learning (ICML)[4] were not robust, confirming that defending against adversarial attacks is a difficult problem. In closing, Athalye emphasized that robustness is a real-world concern because attacks are outpacing defenses. He said that it is crucial to understand the risks of adversarial attacks to current systems through rigorous evaluations and a principled approach that will lead to the construction of secure machine learning systems.

Chellappa suggested that Athalye revisit his conclusion that machine learning defenses are not robust, given that the MNIST database[5] (Modified National Institute of Standards and Technology database) of handwritten digits was labeled with 55 percent accuracy. He emphasized the value of describing both when

[4] For more information about the 2018 International Conference on Machine Learning, see https://icml.cc/Conferences/2018, accessed February 19, 2019.

[5] For more information on the MNIST database (Modified National Institute of Standards and Technology), see Y. LeCun, C. Cortes, and C. Burges, "The MNIST Database of Handwritten Digits," http://yann.lecun.com/exdb/mnist/, accessed April 8, 2019.

systems are working and when they are not. Chellappa also suggested that if a forensic examiner is working alongside a machine learning algorithm, many of the real-world problems Athalye described could disappear. He added that with knowledge of time series models from the 1970s, the concept of adversarial examples should not come as a surprise to anyone. Another audience participant discussed the black-box inversion that results from making many queries against a model and the characterization that is needed to produce an adversarial example. Athalye noted that although he is not aware of much research in this area, some are working on decreasing the number of queries required.

5

Enablers of Machine Learning Algorithms and Systems

IMPACT OF NEUROSCIENCE ON DATA SCIENCE FOR PERCEPTION

John Tsotsos, York University, Canada

John Tsotsos, York University, described that much has been written on how neuroscience and other brain sciences have inspired artificial intelligence (AI). In the context of human and non-human primate vision systems, he discussed the following: (1) how knowledge of the brain is being used to enable progress, (2) the notion that the predictive value of AI models has not been well exploited, and (3) whether modern AI has the right paradigm.

First, Tsotsos explained that human vision and brain sciences have inspired computer vision since its inception; a deep understanding of human vision helps better target and constrain solutions. He added that it is challenging to determine what level of abstraction to use to address a problem. Tsotsos provided a brief and selective history starting with Roberts (1963), who was inspired by Gibson (1950), whose work focused on the notion that if an object changes orientation or pose in a scene, a computer vision system must be able to recognize the object equally well. Rosenfeld and Thurston (1971) developed algorithms to employ grouping principles, which served as the foundation for perceptional organization. Julesz (1971) found stereovision to be a cooperative process, which led to the development of Marr and Poggio's (1979) classic stereo algorithm. Uhr (1972) proposed layered recognition cones for image processing, and Fukushima (1975) added the self-organizing component. Tsotsos (1987) demonstrated that the self-organizing architecture satisfies basic resource constraints in the brain as well as provides a resolution of the combinatorial problems of visual information processing. These layered hierarchies are now part of most of today's successful systems, Tsotsos explained. He said that behavioral studies also played an important role, beginning in the 1960s; Potter (1975) and Thorpe et al. (1996) later revealed that categorization could occur in very short presentation times. Marr's approach to visual processing, that discrimination had to be made in less than 160 milliseconds, was consistent with Potter and Thorpe's findings, although his work had been interpreted incorrectly by others as applying to everything instead of to the first segment.

Tsotsos shared a brief and selective timeline to AlexNet (see Figure 5.1). He noted, however, that the timeline does not end here; graphics processing units (GPUs) and increased memory enable the ability to

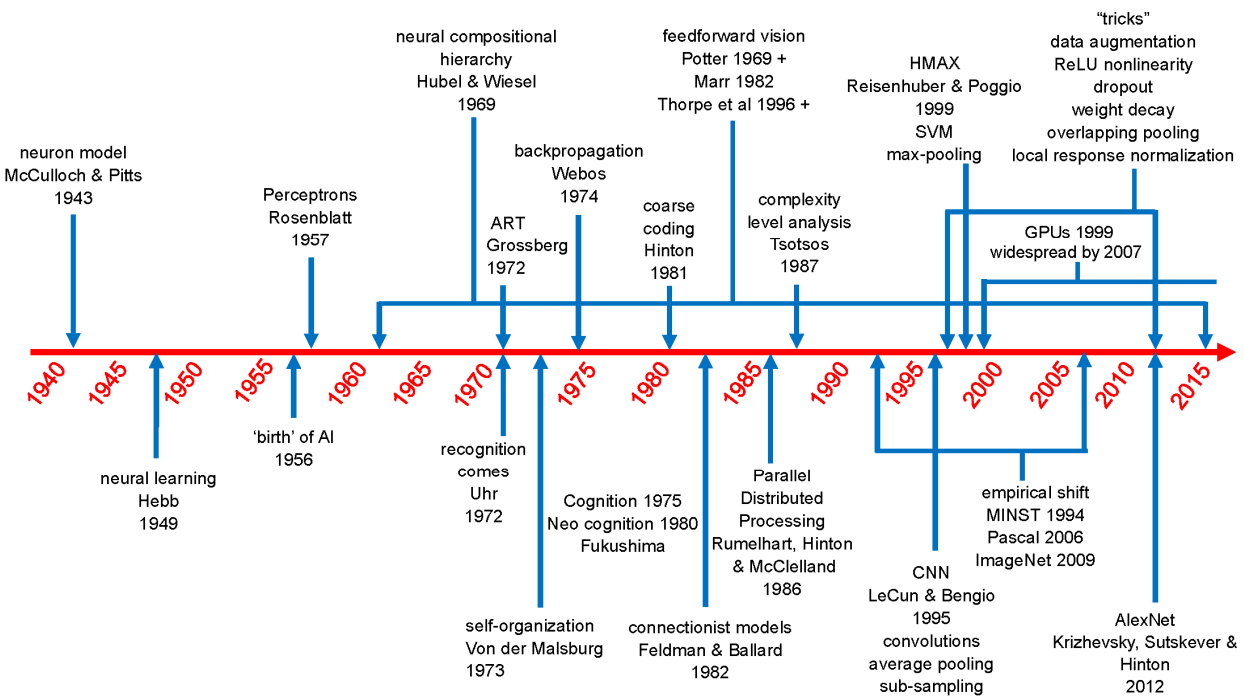


FIGURE 5.1 A brief timeline of the events that led to the development of AlexNet. SOURCE: John Tsotsos, York University, Canada, presentation to the workshop, December 11, 2018.

train on data sets. This indicates an empirical shift in how problems are solved. Previously, since answers were constrained by often limited resources, strategies followed a classic scientific method and, since large test sets were unavailable, solutions were generally independent of data. Modern AI, however, has adopted two premises in the meaning of solving a problem: (1) the solution should be "probably close to correct" (Valiant, 1984) and (2) the inductive learning hypothesis is used, which means that "any hypothesis found to approximate the target function well over a sufficiently large set of training examples will also approximate the target function well over other unobserved examples" (Mitchell, 1997). Tsotsos clarified that these can be formalized, but the solution strategies focus on data-driven learning, rejecting the classic scientific method, and measuring success using statistical uncertainty measures. Also, large test data sets are widely available and solutions are developed by incorporating statistical regularities from those data. Krizhevsky et al. (2017) said that "results can be improved simply by waiting for faster GPUs and bigger data sets to become available," but Tsotsos suggested that we are up against the Pareto Principle (i.e., 20 percent of the occurrences do not fit the abstraction) and that to succeed, the entire strategy has to change because the initial model was incorrect.

Tsotsos next discussed other aspects of the brain and human vision that have influenced the development of AI. He described the human retina, which is highly variable and does not generate homogenous and uniform representations, such as images. As he discussed the receptive field size in the visual cortex and spatial layout of brain neurons, he noted that this is a characteristic of machine vision system hierarchies that has not yet succeeded. Convolutional neural networks, he continued, do not mimic the architecture of the brain. The pathways in the brain are more top-down than feedforward, they are not strictly hierarchical, and the strength of connectivity is not uniform (see Figure 5.2).

The data structure more representative of the brain's pathways in Figure 5.2 is a lattice. Tsotsos called this a lattice of pyramids (P-Lattice) because each of the brain areas in Figure 5.2 is actually many sheets of neurons, where each has a particular retinotopic relationship to the others. This structure more closely models the network of the brain. He explained that due to the architecture and the connectivity, three problems arise: the crosstalk problem, the context problem, and the boundary problem.

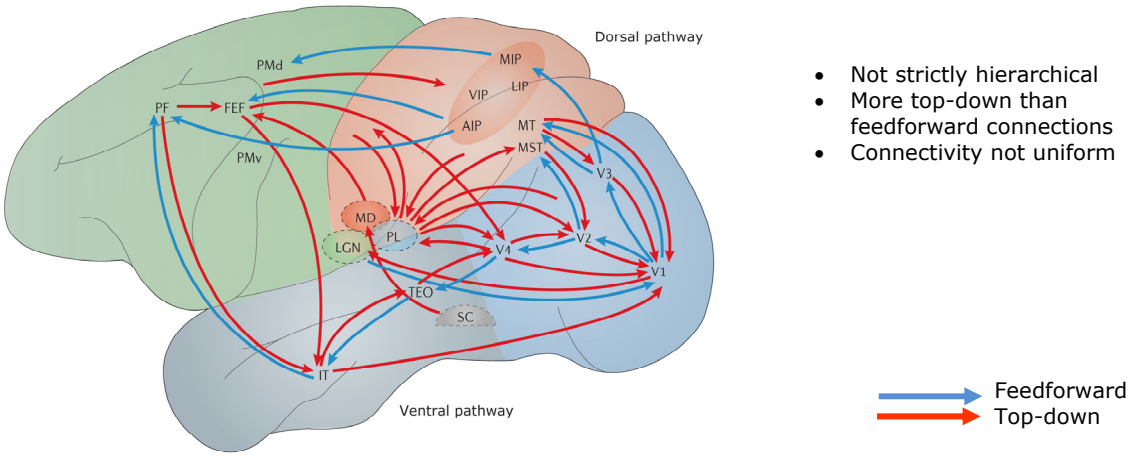


FIGURE 5.2 Neural pathways in the brain. SOURCE: John Tsotsos, York University, Canada, presentation to the workshop, December 11, 2018; reprinted with permission from Springer Nature: C.D. Gilbert and W. Li, 2013, Top-down influences on visual processing, *Nature Reviews Neuroscience* 14(5):350-363. Copyright 2013.

In terms of the crosstalk problem, he explained that the anatomical structure of connectivity leads to signal interference in overlap areas. To demonstrate an example of feedforward crosstalk, or entanglement, Tsotsos and his team manipulated images that were part of a training set to see how the best object detectors would perform. They used a state-of-the-art object detection method applied to an image of a living room from the Microsoft Common Objects in Context object detection benchmark. The image of an elephant placed in the room interfered with the network that represents the whole room and was thus not recognized unless it was to be moved slightly.

He explained that "predictive value" of AI models does not mean that a model predicts known observations—he defined that process as "explaining" instead. "Predictive value" means that a model makes a claim that needs to be tested. He said that a theory in the empirical sciences can never be proven, but it can be falsified. He explained that when one positively tests predictions, model believability increases.

Negatively tested predictions provide the impetus for model refinement, and better models enable an understanding of why failures occur. He stressed that this example of the elephant in the room showed that the underlying architecture of deep learning systems has inherent failures, and the only way to improve is to change the systems themselves. This is a more effective long-term solution than simply finding a defense for an attack on a system, he continued, as this example has often been criticized.

The second problem in network properties is the context problem. Tsotsos hopes in the future to see trained neural networks (with dynamic tuning) functioning correctly. He provided some history of developments in human learning relevant to this area. The understanding of how neurons change their behavior dynamically became more clear upon discovery of papers by von der Malsberg (1981) and Moran and Desimone (1985). Moran and Desimone (1985) performed an experiment with monkeys to identify neurons that responded strongly to stimuli; the monkeys were trained to attend to a response. This study led to Tsotsos' creation of a model of selective tuning in 1990 that showed a pattern of suppression around items attended all through the visual hierarchy. This led to an overall model of vision that uses visual hierarchy over time in both feedforward and feedback directions (Tsotsos et al., 2008). Another experiment (Cutzu and Tsotsos, 2003) asked subjects to fixate on a black dot and attend to a grey disk and then tested their ability to detect items near and far from that grey disk. This confirmed that Tsotsos' initial prediction about suppressive surround was true. Hopf et al. (2006) studied how attentional suppressive surround is used in neurophysiology, showing the same pattern around a central point of surround suppression. Boehler et al. (2009) confirmed that attentional suppressive surround was due to recurrent activations. Bartsch et al. (2017) revealed that attentional suppressive surround exists not only in the spatial dimension but also in the feature dimension.

Tsotsos explained that the third problem in such hierarchical networks, the boundary problem, is well known in computer vision. This is an important problem to consider because the human visual system overcomes it with eye movements—not image padding, as most current methods use. This will lead to a different eye fixation scheme than has been considered. Classically, one first makes a saliency map of an image, does a selection on the map, and then finally moves. That approach is insufficient because it lacks a good representation of the periphery. Tsotsos et al. (2016) developed a different fixation model: A number of different representations contribute to a priority map from which a selection is made. The first is a representation of the peripheral attentional map, and the second is a representation of the central attentional map. These are combined with what has been seen and what should have been seen, modulated by task information, which is then combined into a priority map for selection. The model does an excellent job of fixation in comparison to humans: accuracy is almost within human error. Selective tuning models can make predictions of new knowledge, and AI can help neuroscience in this way (instead of only providing analysis tools). Tsotsos summarized that it is unclear whether there have been any falsifiable predictions made by deep learning models. He added that attentional mechanisms are often ignored yet critical for generalization and that selective tuning offers an example of what is possible without machine learning.

In considering the question about whether AI has the right paradigm (i.e., a data-driven machine learning approach to AI), Tsotsos said that it may be too early to reject other approaches. Human learning is not based only on data: statistical information in a data set may not provide any insight into human learning processes. He explained that computer vision is task-directed, and single images under a variety of instructions are labeled differently; thus, it would not make sense to train a system with all possible instructions. He described visual processing as intractable; a single solution that is optimizable in all instances is impossible, so it is necessary to reframe the original problem. The space of all problem instances can be partitioned into subspaces where each may be solvable by a different method. Given that the brain is a fixed processing resource, the need to employ a variety of different solution strategies in a situation-dependent manner implies that those resources must be dynamically tunable to the current situation and that there exists an executive controller that orchestrates the process.

Tsotsos showed results that demonstrated how human visual capabilities are not fully present at birth, including those driven by the predictions of his Selective Tuning model. The suppression he showed was discovered to need 17-18 years to mature fully. He indicated that although it takes time for each human visual mechanism to mature, it is possible for humans to learn while these mechanisms are still maturing. This is important to consider when building learning algorithms, although he added that we do not fully understand the interplay between how we learn content while developing mechanisms. And, human learning is far more complex than current machine learning might capture.

In conclusion, Tsotsos said that neural and behavioral scientists need to be convinced that AI can contribute in the path to emulate human intelligence. He said that practical applications should not emulate human failings, but it is important to understand why those failings exist. AI cannot be superficial about the brain, and AI seems not to be moving in the direction of understanding or creating human-like intelligence: chasms have developed between current AI systems and what is known about human intelligence. Tsotsos emphasized that neural inspiration seems to have stalled decades ago, and the community is unsure of what it is trying to build. In response to a question from a workshop participant, Tsotsos said that there has not been work in his community about a system that has the ability to reason that something is not there. To solve that problem, he continued, active, inductive inference is needed to predict when something is there, to go and look for it, and to confirm whether it is there. In response to a question from Rama Chellappa, University of Maryland, College Park, Tsotsos said that difficulties arise in applications in which humans and machines work together because humans expect robots to work in ways they understand to be correct and rational (i.e., like a human). Both human and machine should know the goal of a joint task and develop an intention contract so that each would know what steps the other would take toward that goal in order to make things work out as each expected, he explained. A workshop participant wondered if the suppression observed in the experiment by Cutzu and Tsotsos (2003) was the same as the non-maximum suppression in neural networks, and Tsotsos responded that the suppression is not the same.

6

Recent Trends in Machine Learning, Parts 1 and 2

ON OPEN SET AND ADVERSARIAL ISSUES IN MACHINE LEARNING

Terry Boult, University of Colorado, Colorado Springs

Terry Boult, University of Colorado, Colorado Springs, said that the first part of his talk would explore issues with unknown inputs and open-set recognition in deep networks. He explained that human systems excel at both recognition tasks and noting when they do not know what an object is. However, researchers are not yet building open-set recognition problems and need to be considering how to deal with unknowns in the environment. Classic machine learning tends to deal with binary or multiclass classification (i.e., everything in training and testing comes from one of the known classes). However, other common problems include face verification, object detection, and open-set recognition. In this last case, there are multiple known classes and many unknown classes. Boult defined open-set recognition as a mixture of two very different problems: multiclass recognition and anomaly/novelty detection. Very different types of errors must be balanced, and risks associated with rare events must be considered. Handling unknown unknowns is especially important in the Intelligence Community (IC).

The problem of unknown unknowns can be formalized by balancing two risks: open space and empirical (Scheirer et al., 2013). Open space is defined as the space far from the known samples. The risk, then, is labeling anything other than the unknown. Boult cautioned against using the threshold classifiers' confidence to address this problem because this focuses on the boundary between the known and unknown classes, which is ill defined for open-set problems. Classic machine learning presumes all classes are known, and it classifies all of the feature space; this does not apply to open-set problems.

To address this problem, Boult's team developed an extreme value machine (EVM) in which the boundary between classes is described in every machine learning problem using a margin distribution theorem to derive extreme value theory-based non-linear models that are provably open set and can do incremental learning (Rudd et al., 2018). And, as new data are added to the classes, other classes can shrink and change shape efficiently. Boult next discussed how to use deep networks for open-set problems (Bendale and Boult, 2016). With a normal network, AlexNet makes a prediction and output comes from a SoftMax classifier. However, people started finding ways to attack deep neural networks with adversarial

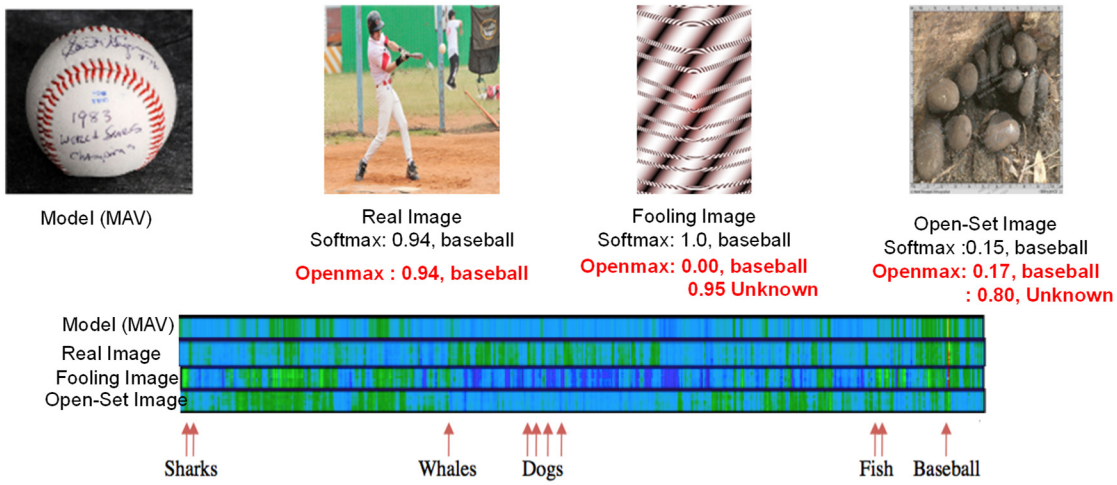


FIGURE 6.1 OpenMax detection of adversarial images. SOURCE: Terry Boult, University of Colorado, Colorado Springs, presentation to the workshop, December 12, 2018.

images. Boult's team set out to correct the high confidence errors and created OpenMax, an attempt at open-set deep networks. OpenMax takes one of the deep feature layers, using that to represent a particular class by taking positive instances of that class, and builds extreme value distribution to provably solve open-set problems. OpenMax looks at distance and representational space and provides probability-based estimates for image classes; the adversarial image is not consistent with other representations, so the OpenMax labels it as unknown (see Figure 6.1). He said that this could be done with open-set images: OpenMax can recognize something that it is not trained to handle, making this approach slightly better than the standard SoftMax.

Boult said that they expected the accuracy gap between SoftMax and OpenMax to be greater than 20 percent and the F-measure gap to be greater than 2 percent. So, they performed a test with known unknowns and unknown unknowns. With SoftMax, features for unknown inputs generally overlapped the known classes because the network was not designed to map things it did not know about. Boult and his team observed that there was a difference in entropy and magnitude: while the open set limited the response outside the ring of data, most of the unknowns had smaller magnitudes of deep feature representations. He described this as a natural outcome of looking at unfamiliar objects. However, OpenMax aims to represent unknown mappings with smaller magnitudes and the known mappings with larger magnitudes, with the feature vectors being pushed to the center. This was accomplished by introducing two new losses: entropic open-set loss, in which the entropy margin increases as do the SoftMax scores for the unknowns, and ObjectoSphere loss, in which the deep feature magnitude margin is increased, minimizing the Euclidian length of the deep representation for the unknowns. He explained that while the background classifier performed well, there were still many things outside the wedge. He pointed out that OpenMax performed poorly with a two-dimensional problem, while the entropic open set and the ObjectoSphere performed well. Classic open-set approaches do not really solve the problem of deep unknowns, Boult explained; using ObjectoSphere loss begins to address the learning of deep features that can handle unknown inputs.

Next, Boult's team used ObjectoSphere for open-set face recognition. Face recognition has to deal with unknown non-faces and unknown probes, so they trained a three-layer ObjectoSphere net with VGG2 input features and forced all unknowns toward the origin. ObjectoSphere performed better than the baseline and better than EVM and generally better than the gallery fine-tuned SoftMax. Boult said that in the near term, he would like to see open-set recognition algorithms used for well-behaved, low-moderate dimensional feature spaces. In the long term, he hopes to see better network models for open-set recognition as well as high-dimensional open-set algorithms. Research areas to improve include realistic open-set data sets/protocols and a better understanding of the image and feature relationships. Boult added that when

accuracy is used as a measure, performance is typically worse because it is impossible to preserve the robustness and there are few protocols set up to do so. Rama Chellappa, University of Maryland, College Park, interjected that the Intelligence Advanced Research Projects Activity's (IARPA's) Janus[1] had a protocol for open-set recognition. Boult acknowledged Chellappa's comment but pointed out that today's definition of an open set is different. Anthony Hoogs, Kitware, Inc., observed that Boult's capabilities list assumes a recognition problem as opposed to a detection and recognition problem, which is of particular importance in the IC. He suggested exploring detection when the spatial extent of the object is unknown because the class is unknown. Boult said that a background classifier was added to aid in detection; open set is more challenging because the only meaningful questions are detection and recognition. He explained that doing face recognition without counting scores for non-faces causes real problems—this is a key difference from the Janus protocol. An audience participant asked about the trade-off between generalization accuracy and open-set accuracy. Boult said that the first question to ask is how generalization accuracy is being defined. The next question to consider in open set is how distance is generalized: how far is too far? He emphasized the need to avoid generalizing more than is consistent with the data. Using the margin distribution theorem can be helpful in this instance.

Boult next discussed another type of unknowns for deep networks: the relationship between open-set and adversarial examples. Adversarial examples are image perturbations that are invisible to humans but can easily fool deep networks. An open set is naturally thought of in "image space," but the methods work in "feature space." Adversarial examples show that for current deep neural networks, the two spaces are often not well related. While open-set recognition tries to deal with inputs that are far from known training samples, adversarial examples are "imperceptibly" close to known data and different from hill-climbing adversarial images. Adversarial examples show that we do not understand how or what networks learn, but they can provide insights to improve deep network learning theory, he continued. He explained that adversarial perturbations are unnoticeable for humans. One can relate image and feature spaces by considering the following questions about the root causes of imperceptible adversarial images:

- Does adversarial training improve robustness to other types? For some adversarials, no amount of training has made any difference in robustness.
- Is it aliasing caused by aliasing and/or down sampling?
- Do adversarial examples develop because we "stop" near the boundaries?
- Is an adversarial example caused by SoftMax/strict classification layer in networks?
- Does our OpenMax approach improve robustness?
- Does threshholding distance in center loss protect from "adversarial images"?
- Can attributes help resolve adversarial issues for face recognition?
- Are all networks/classifiers subject to adversarials?
- Can we use adversarial examples to improve privacy?
- Can we build networks that estimate uncertainty that captures both open-set and adversarial properties?

Adversarials provide insights about the theories needed for networks and how they generalize, Boult explained.

In response to a question as to whether adversarial examples occur naturally, Boult said that adversarials do not have to be machine generated; they can occur if there is noise in a system. In order to attack deep features, Boult explained, one has to attack what is in the middle instead of only the output. Boult introduced a deep-feature adversarial approach called the layerwise origin-target synthesis (LOTS); using examples of object-recognition and face-recognition, he showed how LOTS' adversarial examples can successfully attack even open-set recognition systems, such as OpenMax, by matching features. In the study of adversarial system-level attacks, he noted that all prior work attacked just the network input image to change

[1] For more information on the Janus research program, see Intelligence Advanced Research Projects Activity, "Janus," https://www.iarpa.gov/index.php /research-programs/janus, accessed February 19, 2019.

the network output. His work matched an image's deep feature representation. He provided an overview of examples of full-system attacks using LOTS. Using 12 adversaries, LOTS attacked every single gallery template with a success rate over 99 percent. He also shared unpublished results on one-shot black-box attacks: there was no hill climbing on the system to be attacked; they used only a surrogate network and then a test attack. Untargeted attacks were weak on some systems but were somewhat successful and portable across multiple commercial systems. Targeted attacks vary, so Boult wanted to understand why and when they are portable. Boult found that minimal perturbations are rarely portable, and they are better when the surrogate network has the same structure.

Boult hopes that in the near term, capabilities will include an iterative LOTS system having the ability to attack all kinds of systems, with LOTS attacks being reasonably portable. He believes LOTS can be used to build physical attacks and camouflage. An open research question is whether manipulation feature learning, as addressed by ObjectoSphere, could improve the adversarial example issue. Boult emphasized the importance of having a systematic, scientific approach in the adversarial space. An audience participant asked about adversarial retraining and whether hardening the network by trying to build this type of attack will detect or defeat the attack and morph the image in a way that makes sense. Boult said that in his experiments, no matter how much training with adversarials was done against LOTS, it made no difference; neither accuracy nor robustness improved. In response to another audience participant's question, Boult acknowledged that there are many open questions about how to handle unknowns in high dimensions, and he encouraged more work in this area.

GENERATIVE ADVERSARIAL NETWORKS (GANS) FOR DOMAIN ADAPTATION AND SECURITY AGAINST ATTACKS

Rama Chellappa, University of Maryland, College Park

Rama Chellappa, University of Maryland, College Park, pointed out that for certain problems, such as face recognition, deep learning has worked well. He provided an overview of the many ways in which the world of computer vision has changed in the past 6 years. While the field used to be focused on physics and geometry, it is now focused on data. The field of computer vision has also witnessed impressive performance—notably on tasks such as face verification as well as object and face detection and recognition—and efforts continue (such as the IARPA Janus program) for unconstrained face verification and recognition. The field now uses multiple networks based on ResNet and InceptionNet, and researchers have achieved the true acceptance rate around 90 percent at the false acceptance rate of 1 in 10 million for the unconstrained face verification problem on a challenging face data set, according to Chellappa. Deep learning techniques are being used for the IARPA Deep Intermodal Video Analytics (DIVA) program,[2] and human analysts and computer vision systems continue to work together to achieve optimal performance, with the human making the final decision for the problem. Chellappa emphasized that adversarial examples are not actually a new problem; outlier detection has been an issue for years. He suggested that adversarial examples are much easier to combat with a human in the loop.

Chellappa discussed deep learning efforts for unconstrained face verification in more detail. Since 2014, this work has been supported by the IARPA Janus program and is a collaboration among the University of Maryland, Carnegie Mellon University, Columbia University, Johns Hopkins University, the University of Colorado, Colorado Springs, and the University of Texas, Dallas. This team is working on multitask learning in deep networks and network-of-networks problems. They have achieved state-of-the-art performance on face verification, search, and clustering tasks using a relatively small training data set, and their work has important implications for the field of forensics. Chellappa also discussed their work in

[2] For more information about the Deep Intermodal Video Analytics research program, see Intelligence Advanced Research Projects Activity, "Deep Intermodal Video Analytics (DIVA)," https://www.iarpa.gov/index.php/research-programs/diva, accessed February 19, 2019.

hyperface architecture, which utilizes deep neural networks. By sharing features, multiple tasks could be completed to address detection problems (e.g., pose estimation, age estimation). A good gallery of images is typically needed to build a three-dimensional (3D) model; however, for this project, they determined that 3D models were not needed. Performance increased with a focus on feature distribution. Chellappa presented face recognition and verification results on many publicly available challenge data sets made available by the IARPA Janus program. Hoogs observed that many programs rely on closed data sets that can only be viewed by the performer. As a result, papers are often rejected because the reviewers cannot evaluate the methods based on the data. If these programs' practices were more aligned with the open standard of the field, Hoogs continued, people would participate more often. Boult added that this lack of open data is not only impacting one's ability to publish but it is also limiting how and when problems could be solved. For example, if people who are not funded by these programs had access to the data, they might be able to provide better ideas to the performers and the government. He emphasized that the practice of sequestering data is bad for advancing science.

Chellappa next described an experiment with forensic examiners, human super-recognizers, and face recognition algorithms (Phillips et al., 2018). The experiment compared recognition capabilities of algorithms versus humans, and the results indicated that the best algorithm was performing much like a forensic examiner. Enabling human–machine collaboration combines the strengths and weaknesses of each and results in more accurate predictions. He noted the tremendous opportunity for learning of all kinds when humans and machines work together, especially in digital pathology and other health applications.

Chellappa's next topic was on adversarial learning using generative adversarial networks (GANs). He is most interested in GANs and domain adaptation to model the domain shift between the source and target domains in an unsupervised manner. GANs are excellent models that can be conditioned on learned embeddings and prompt the following questions: Can we extract knowledge of the target distribution using a generative process during training? Can we perform adaptation in a task-agnostic manner? With a conditional GAN, he continued, one works on the features themselves instead of on the data. Deep convolutional neural networks and GANs are popular because they perform well in domain transfers where many other approaches would fail.

Another problem in domain adaptation is semantic segmentation. Chellappa described an experiment aimed at utilizing synthetically generated labeled data and unlabeled real data to reduce the domain shift when evaluated on real data. Typically, learning algorithms perform poorly on samples drawn from a distribution other than training data, and deep models trained on synthetic data do not generalize well to real data. However, labeling real data is difficult and time consuming. Chellappa proposed a GAN-based training approach to learn a common feature space where the distance between source and target distributions is minimized. Contrary to several other approaches to domain adaptation, this addresses the large-scale semantic segmentation task, which increases the practical significance of the work. The pipeline consists of two main components: a supervised component and an adversarial component. The loss term consists of three parts: (1) the supervisory loss from the classification net, (2) the auxiliary classification loss from the discriminator, and (3) the domain prediction loss from the discriminator. All of the networks are trained in an end-to-end fashion. During test time, the network is deployed as a simple standalone network without the GAN components (see Figure 6.2).

Chellappa and his team also performed quantitative experiments over two large-scale adaptation settings: SYNTHIA[3] to CITYSCAPES[4] and Grand Theft Auto-5 to CITYSCAPES. The results on SYNTHIA to CITYSCAPES shows a mean intersection over union gain of 9.1 over a stronger source-only baseline network compared to previous approaches. Similarly, results on GTA-5 to CITYSCAPES also revealed a significant improvement in the baseline performance and as compared to previous approaches. It can be clearly observed that the improved generator quality results in an improved performance. He said that the field is improving in the creation of synthetic data and added that because GANs have hundreds of thousands of data, they are useful tools.

[3] The SYNTHIA data set is available at http://synthia-dataset.net/, accessed February 19, 2019.

[4] The CITYSCAPES data set is available at https://www.cityscapes-dataset.com/, accessed February 19, 2019.

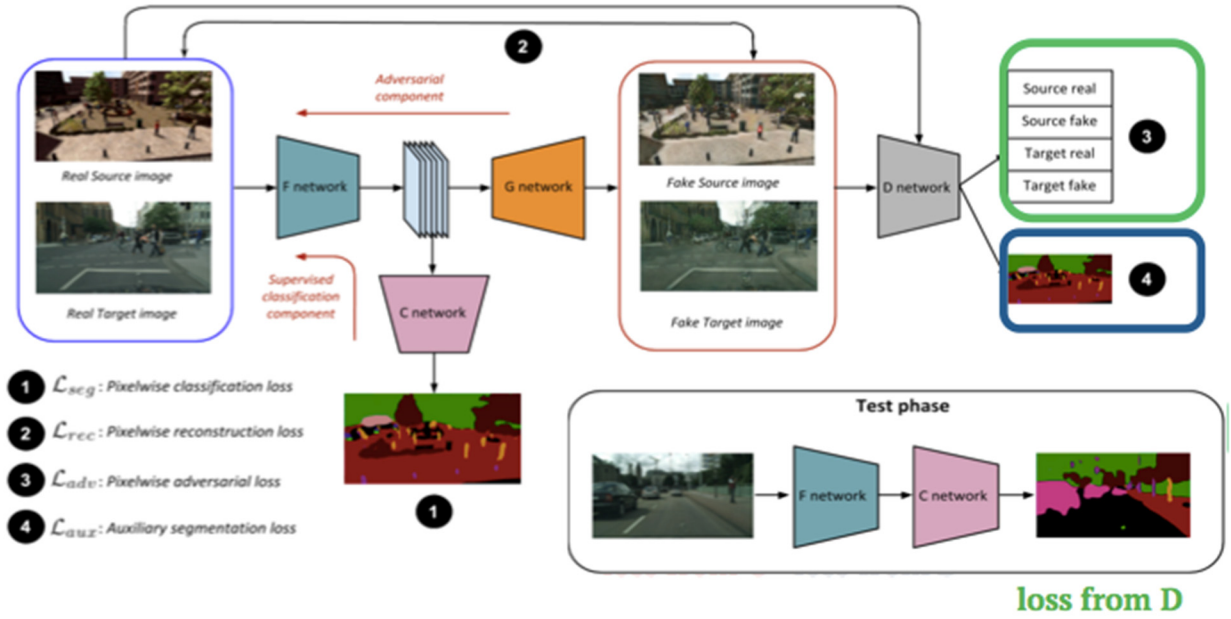


FIGURE 6.2 Overall approach to adaptation for semantic segmentation. SOURCE: Rama Chellappa, University of Maryland, College Park, presentation to the workshop, December 12, 2018. © 2018 IEEE. Reprinted, with permission, from S. Sankaranarayanan, Y. Balaji, A. Jain, S. Nam Lim, and R. Chellappa, 2018, "Learning from Synthetic Data: Addressing Domain Shift for Semantic Segmentation," pp. 3752-3761 in *Proceedings of IEEE/Computer Vision Foundation (CVF) Conference on Computer Vision and Pattern Recognition*, https://conferences.computer.org/cvpr/2018.

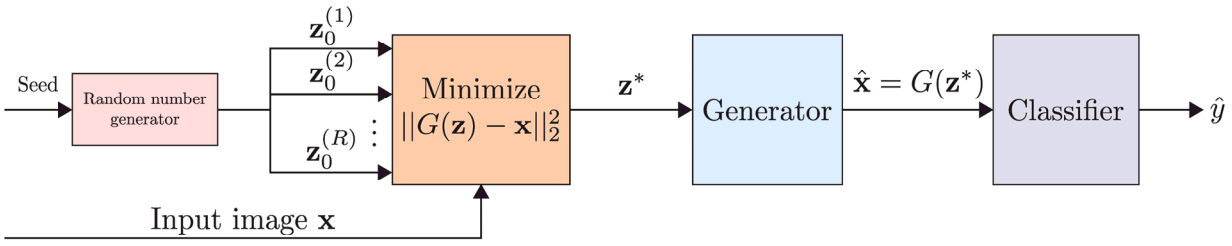


FIGURE 6.3 A generative adversarial network for defense. SOURCE: Rama Chellappa, University of Maryland, College Park, presentation to the workshop, December 12, 2018.

He explained that, given our robust statistics literature, it should not be a surprise that even small perturbations can break deep learning networks, which are hierarchical nonlinear regression models. He next discussed two current defense approaches: Defense-GAN (i.e., an algorithm to detect outliers; see Figure 6.3) and (2) Network-based solution (based on compact convolutional and L_2-SoftMax loss function). He reiterated that the output of the deep network should be looked at by the human in the loop. He said that the Defense-GAN is robust to Carlini-Wagner attacks; however, he noted that there is still work to do since no universal solution exists. The Defense-GAN is only a partially robust approach at 55 percent accuracy. Chellappa added that there is a theoretical justification for the Defense-GAN in the context of bounds (Fawzi et al., 2018) and said that bounds are important.

Chellappa suggested that networks could be fortified by hiding layers in a deep network to identify when the hidden states are off the data manifold and map these hidden states back to parts of the data manifold where the network performs well (Lamb et al., 2018). He said that because GANs are data-sensitive, not problem-specific, additional work is needed. He suggested that people in adversarial research put things in perspective in terms of how the work will be used and what issues will arise. With compact convolution, another way to structure the network, it is possible to constrain features to lie on a hypersphere manifold. Humans can distinguish the differences in adversarial examples with compact convolutional network sample perturbations. Medicine is another domain worth further consideration in the context of deep learning and artificial intelligence, according to Chellappa. At the 2019 Digital Pathology and Artificial Intelligence (AI) Congress: USA, he continued, GANs and domain adaptation for clinical research and health care will be an important topic.

Chellappa closed his presentation by discussing the medium- and long-term research opportunities for the field of computer vision. In the medium term, he hopes researchers will explore the robustness of deeper networks; work with multi-modal inputs; increase theoretical analysis; investigate how humans and machines can work together to thwart adversarial attacks; and demonstrate on more difficult computer visions problems (e.g., face verification and identification, action detection, detection of doctored media). He suggested that approaches based on MNIST (Modified National Institute of Standards and Technology database) data might not be effective for real computer vision problems. In the long term, research opportunities include adaptive networks (i.e., changing the network configuration and parameters in a probabilistic manner with guaranteed performance); humans and machines working together; and design networks that incorporate common sense reasoning. Boult noted that the fundamental problem, which is the existence of attackers, has to be addressed at some point. Hoogs said that Kitware, Inc., has a large group in medical image analysis. The medical community has been slower to implement new imaging technologies than other communities, but a 2018 conference on AI for radiologists emphasized deep learning. He described this as evidence that a revolution is impending and noted that regulatory and acceptance issues central to the medical field might also be of interest to the IC.

RECENT ADVANCES IN OPTIMIZATION FOR MACHINE LEARNING

Tom Goldstein, University of Maryland

Tom Goldstein, University of Maryland, explained that his talk would focus on theoretical perspectives on adversarial examples. He said that most people are familiar with evasion attacks in which malicious samples are modified at test time to evade detection. These attacks will fail in situations when one cannot control the properties of test time data. Another type of threat model is a poison attack in which the adversary controls the training data and manipulates the classifier during training. Poison attacks present substantial security risks and typically are initiated by bad actors or inside agents or enabled by harvesting system inputs or scraping images from the web.

Goldstein described the poisoning process, using an image that has been classified by ImageNet as an airplane, which is then attacked by an image of a "poison" frog. An image of a frog is attacked with a perturbation, and the new image is added to the training set. At test time, the classification of the airplane changes to a different class without anyone ever manipulating the test image itself (see Figure 6.4).

Goldstein explained that these attacks could be created in one of two contexts: transfer learning or end-to-end re-training. In transfer learning, a standard pre-trained net is used, the feature extraction layers are frozen, and the classification layers are re-trained. He described this as a common practice in industry in which it is possible to make a "one-shot kill." In end-to-end re-training, a pre-trained net is used and then all layers are re-trained. This approach requires multiple poisons.

Goldstein provided an example of the transfer learning approach to creating a poison image via a collision attack. With the same base image (frog) and target image (airplane) as before, the goal is for these terms to collide in the feature space. When solving the optimization problem, the objective is low. A poison

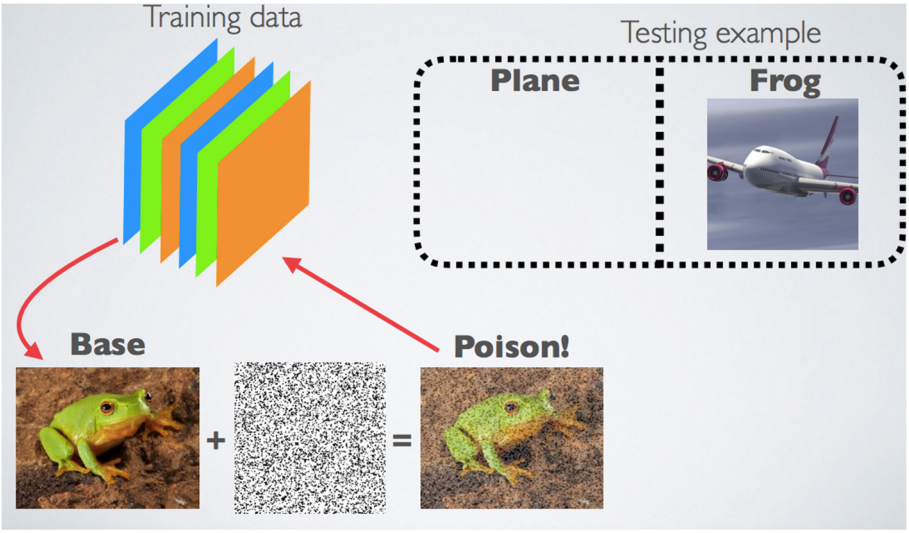


FIGURE 6.4 Example of a poison attack. SOURCE: Tom Goldstein, University of Maryland, presentation to the workshop, December 12, 2018.

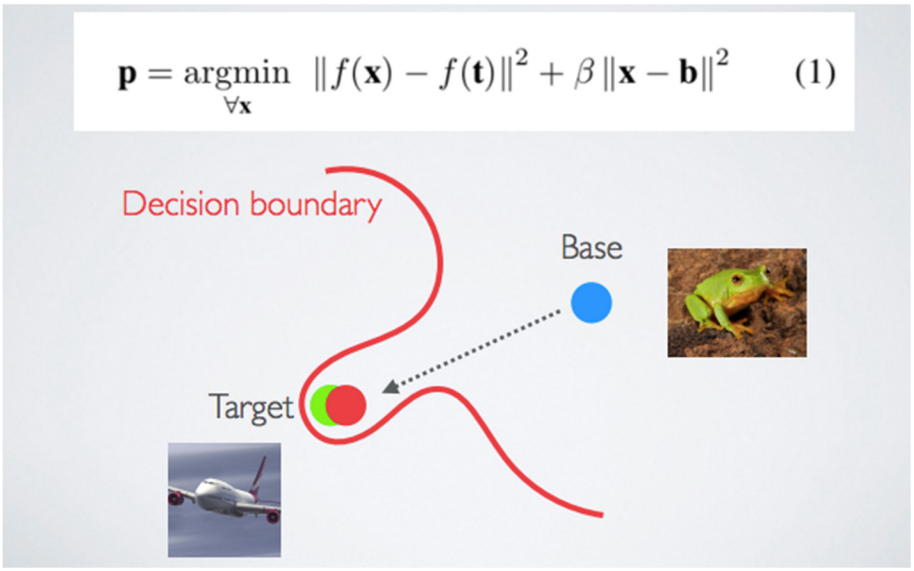


FIGURE 6.5 Example of a collision attack. SOURCE: Tom Goldstein, University of Maryland, presentation to the workshop, December 12, 2018.

image that looks like a frog collides with the airplane image in the feature space and is correctly labeled as a frog. The decision boundary moves and classifies objects in the feature space as frogs, so then the airplane will be classified as a frog at test time (see Figure 6.5).

In another example, Goldstein showed how he used an ImageNet classifier to classify fish images as fish. He then added a poison image to change the classification of the test images, thus creating poison dogs. After adding that image to the data set, the fish images are classified as dogs at test time. In this instance, it only took one poison image to change the behavior of the classifier at test time. Terry Boult, University of Colorado, Colorado Springs, asked if recomputing happens at every stage, since the feature representations are evolving, and Goldstein responded that since one is only training the deeper layers of

the network, one only needs to fool the feature representation that comes out of the feature extractor. He added that adversarial end-to-end networks are more complicated than this transfer learning case. He continued that these are called "clean label" attacks because the poisons are labeled "correctly"; the performance only changes on selected targets. These attacks are easily executed by outsiders and are difficult to detect (i.e., the picture of the base frog and poison frog look identical, so an auditor would not notice any differences). A workshop participant asked how broadly the attack works if one has a slightly different image of an airplane, for example, and if it would still be misclassified. Goldstein said that his team is working on getting this approach to work with a broader class of objects. In response to another question, Goldstein noted that detecting poison examples is similar to detecting adversarial examples. An audience participant asked if there is a way to leverage and protect the privacy of people's images online. Goldstein replied that when a person posts their picture online, it is labeled. Depending on the context, the picture could be targeted with either an evasion attack or a poison attack. Goldstein added that if a person does not want someone to train a classifier on him, they could add perturbations to the image to make it impossible to train a classifier on the image.

Goldstein turned to a discussion of end-to-end training, in which multiple poisons are required. In this case, the feature extractors learn to ignore adversarial perturbation. He showed a plot from multiple experiments that compared the number of poisons in terms of success rate and concluded that watermarking plus poisons leads to success. With about 50 poisons, the rate of successful attack was approximately 70 percent. In another example, it took 60 poison dogs to cause a bird to be misclassified in end-to-end training (i.e., without a pre-trained feature extractor, many more poisons are needed).

Speaking specifically of adversarial examples from a theoretical perspective, Goldstein explained that there has been constant back and forth between attacks and proposed fixes; ultimately, the state-of-the-art defenses are only partially robust. So, the fundamental question becomes whether adversarial examples are inevitable. To answer this question, most people assume that human perception is not exploitable and that high-dimensional spaces are not too strange—two assumptions with which he disagrees. Goldstein presented a toy problem to better understand adversarial examples: Using a sphere, one can make a simple classifier that slices the sphere into two areas, Class A and Class B. Class A covers 50 percent of the surface area of the sphere. The question of interest is whether there are adversarial examples for objects in Class A. To answer that question, an ε has to be defined (i.e., how powerful is the attacker?) and one computes the ε expansion of Class B (i.e., epsilon equals 0.1). Next, Class B is expanded by 0.1 units. The ε boundary covers Class A. Areas that are covered have adversarial examples and can be moved into Class B with small perturbations. The ε expansion covers 55 percent of the sphere, which means that most areas remain safe. In higher dimensions (e.g., 100 dimensions), even though ε is still 0.1, with random sampling, 84 percent of Class B is covered, which means most things have adversarial examples. At 1,000 dimensions, 99.8 percent of things sampled are within epsilon expansion, so almost everything is an adversarial example. This random sampling process produces adversarial behavior, and this is the most robust classifier one could construct. Levy and Pellegrino's 1951 isoperimetric inequality theorem states that the ε-expansion of any set that occupies half of the sphere is at least as big as the ε-expansion of a semi-sphere. So, in 1,000 dimensions, more than 99 percent of things will have adversarial examples. It does not matter how the class is designed; this assumes that the data are uniformly distributed and that the data live on a sphere, neither of which are true in the real world.

Thus, Goldstein shared a more realistic example of image classifiers. He defined an image as points in a unit cube; a class as the probability density function on the cube, bounded by U_c (i.e., a distribution that can be randomly sampled); and a classifier as the entity that partitions the cube into disjoint sets/labels every point in the cube (which is measurable). Using this method, it is possible to prove, in certain situations, that most things are adversarial. One would assume that in higher dimensions, things are more susceptible to adversarial attacks, but that is not actually the case. He also noted that sparse adversarial examples result when perturbations are hidden in high-frequency images. Goldstein explained that people believe high dimensionality is responsible for adversarial examples because it is hard to fool low-dimensional images in MNIST, for example. However, high dimensionality is not responsible for adversarial examples. There are 1 million ways in which MNIST is different from ImageNet; these are

completely different problems, which are more complicated than what people have assumed. Goldstein performed an experiment in which the image class is kept constant while the dimensionality is changed so as to isolate dimensionality as a variable. He clarified that adversarials are not necessarily worse on the larger MNIST. The dimensionality does not affect adversarial susceptibility, and this holds true in the real world—experiments with real neural nets agree with the theoretical results.

Goldstein explained that the CIFAR data set is more susceptible than the MNIST data set, and ImageNet is more susceptible than CIFAR. He conducted another experiment to show that adversarial susceptibility does not depend on dimensionality, but it does depend on the complexity of the image class. Complex image classes have low density. Returning to the question of whether adversarial examples are inevitable, Goldstein asserted that such a question is not well posed because the answer depends on the epsilon and p-norm chosen, as well as a number of other things.

Goldstein summarized key takeaways from his presentation: (1) adversarial robustness has fundamental limits, and those limits cannot be escaped; (2) adversarial robustness is not specific to neural nets, so clever approaches do not offer an escape; and (3) the robustness limit for neural nets might be far worse than intuition suggests. He noted that the community still has a long way to go in designing better defenses, there are no "certifiable" defenses, and the best defense is still a black-box model. Chellappa asked a question about how Goldstein's work relates to the work of Fawzi et al. (2018), who claim that all classifiers are prone to attack. Goldstein noted that Fawzi et al.'s work uses a different set of assumptions, so it is difficult to compare the two.

FORECASTING USING MACHINE LEARNING

Aram Galstyan, Information Sciences Institute, University of Southern California

Aram Galstyan, University of Southern California, opened his presentation with a discussion of a 1.5-year-long project on hybrid forecasting of geopolitical events before moving to a brief discussion about machine learning. Geopolitical forecasting begins with an individual forecasting problem (IFP). An example of an IFP is "Will the United Nations Security Council adopt a resolution concerning the Democratic Republic of Congo between 15 October 2018 and 15 December 2018?" Each forecasting question should have well-defined resolution criteria, Galstyan explained. The goal of forecasting is to try to solicit probability distribution on possible outcomes, not to try to determine whether something happened. The accuracy of a given forecast can be evaluated using a Brier score, which measures the accuracy of probabilistic predictions.

He described IARPA's Aggregative Contingent Estimation (ACE) program[5] from a few years ago that enhanced forecasting accuracy by combining the judgments of many forecasters who were not expert analysts. The ACE competition revealed that people with no prior forecasting experience could be trained to be good forecasters. The leader of the group who won the competition wrote a book entitled *Superforecasting: The Art and Science of Prediction*, which stated that the untrained were able to generate forecasts as accurate as those of trained analysts. The Hybrid Forecasting Competition[6] succeeds the IARPA ACE program and is developing and testing hybrid geopolitical forecasting systems. These systems integrate human and machine forecasting components to create maximally accurate, flexible, and scalable forecasting capabilities.

Galstyan's project, Synergistic Anticipation of Geopolitical Events (SAGE), is a collaboration among multiple universities and companies. The SAGE hybrid forecasting framework/pipeline is the fairly

[5] For more information on the ACE research program, see Intelligence Advanced Research Projects Activity, "Aggregative Contingent Estimation (ACE)," https://www.iarpa.gov/index.php/research-programs/ace, accessed February 19, 2019.

[6] For more information on the hybrid forecasting competition, see IARPA, "HFC Hybrid Forecasting Competition," https://www.hybridforecasting.com/, accessed February 19, 2019.

complicated system used to generate an automated forecast, with five interrelated tasks: (1) platform engineering (i.e., a web-based platform where users study forecasting questions and generate forecasts), (2) retention and recruitment (i.e., keeping volunteers engaged and interested in this time-consuming activity), (3) machine-based forecasting, (4) human-machine interactions and hybridization, and (5) diagnostic and feedback (see Figure 6.6). He emphasized that this research is interdisciplinary, drawing on expertise in forecasting, human judgment, machine learning, decision making, human–machine interfaces, and textual analysis.

Galstyan noted that the most challenging questions are those without historical precedent and for which statistical models cannot be used. He also believes that there should be no human in the loop for forecast generation. Instead, models should run out-of-the-box after the forecasting question is generated. Tailored models are not built; the fully automated pipeline is used, generating a reading list for users, which can be used to help make an informed decision. For the sake of simplicity for users, Phase 1 of the project has only one model, the AutoRegressive Integrated Moving Average (ARIMA) model for time series forecasts, and it is not a machine learning model. Once the data are received, the system automatically gets the parameters for ARIMA, generates forecast and confidence intervals, and shows the visualization of the model output to the users. A knowledge graph ontology is needed for this process because most of these questions are about events (i.e., all data are represented as events or time series). One issue in this process is scalability; even after some filtering, 150,000 articles from Lexis Nexis are tagged. Using smart sentence segmentation, it is possible to overcome this magnitude of information and find relevant content for users. He explained that Phase 1 had three different experimental conditions in order to see what the machine models added to the forecasting process. Condition A is a baseline (i.e., the user is shown no machine outputs); Condition B is a base rate (i.e., the user is shown historical data); and Condition C shows both the historical data and machine forecast. Next, a general aggregation approach is used to aggregate the individual forecast from the user with the machine forecast into a single forecast. He noted that much literature on crowdsourcing exists for how to aggregate this information. Essentially, a weighted average (where weights dynamically adapt based on performance of different forecasters on different questions) is used. The following findings emerged from the first year of the project: (1) the best accuracy was achieved in Condition B; (2) there is

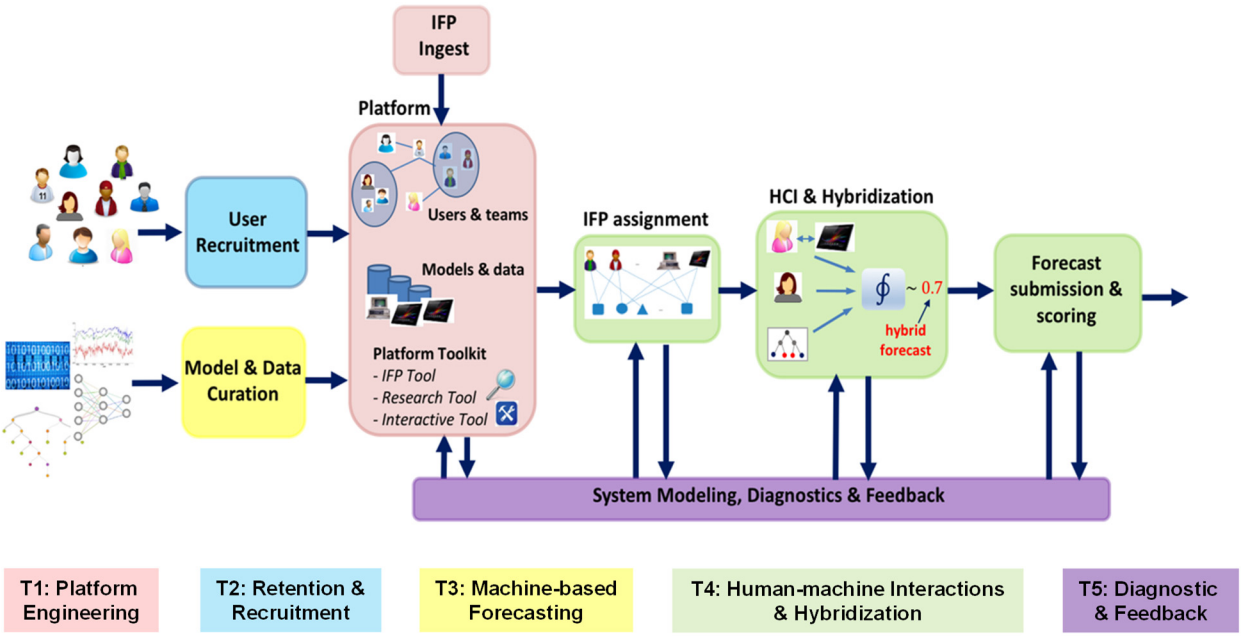


FIGURE 6.6 The Synergistic Anticipation of Geopolitical Events hybrid forecasting framework. SOURCE: Aram Galstyan, University of Southern California, presentation to the workshop, December 12, 2018.

statistically significant improvement in performance in the second half of the randomized controlled trial and improved accuracy of the machine models; (3) something is always going to break (i.e., data will sometimes have missing values or the right data will be unavailable); and (4) machine model accuracy improved over time as the pipeline became more stable.

Galstyan next discussed SAGE and algorithmic aversion; in other words, to what extent do human forecasters trust predictions made by machine models? For example, in an example of misplaced trust, forecasters in Condition C were affected by machine predictions. They adjusted their forecasts toward the forecast of the machine models, even when the machine forecast was inaccurate. This explains why the accuracy in Condition C was poor. An audience participant wondered if the forecasters choose the model prediction because there are no consequences associated with an incorrect decision. Galstyan explained that even though the forecasters were volunteers, the opportunity was considered prestigious and thus people still did their best work. However, he said that the participant's point of concern is valid, and discussions have begun about offering prize money to the best-performing forecasters. In response to a question from another audience member, Galstyan replied that there is a positive effect from the confirmation bias: when the machine forecast agrees with the human forecast, the human moves even further in that direction. In response to another question, Galstyan said that training materials explain to the user how ARIMA works, but users are not told how the model came up with a particular forecast. One of the hypotheses people have explored is that humans do not trust machine predictions because they do not trust a black box and do not know how things work. He does not know how much that trust would change if the process for generating the forecast were to be explained, because even if a user does not understand how a forecast was generated, there is still an influence of the machine model.

Finally, Galstyan discussed SAGE and forecasting rare events. An example IFP could be "Will Austrian journalist Max Zirngast be released by Turkish authorities by October 1, 2018?" The objective is to generate machine forecasts without any historical data upon which to build a meaningful model. Similarity-based forecasting is used. To do this, a bank of similar historical IFPs is leveraged to make a forecast. He noted that it is clear that forecasting questions are not independent; outcomes of IFPs can be interdependent. So it is necessary to explore dependencies and correlations and come up with a predictive model. Galstyan described two geopolitical event data sets: (1) Global Database of Events, Language, and Tone[7] and (2) Integrated Crisis Early Warning System.[8] Each includes coded interactions between sociopolitical actors, and both use the conflict and mediation event observation (CAMEO) coding scheme. These data sets and their coding schema can be used to do zero-shot forecasting, in which the question is parsed into a CAMEO code and mapped. The related CAMEO codes are leveraged to build a model for a particular event. Past events contain useful information for predictions of future events, so the past history of events can be leveraged to come up with better-than-random accuracy. Carefully constructing the training data set allows performance better than the baseline, according to Galstyan.

He emphasized that the best performing methods combine the efforts of humans and machines. He presented a roadmap for the Intelligence Community, including using hybrid sensemaking systems (i.e., analysts working synergistically with different machine models to come up with situational awareness). In the final section of his presentation, Galstyan discussed covariance estimation at scale. He said that it is difficult to understand how two variables are related when the number of samples is less than or equal to the number of variables. He listed a series of broad approaches to estimating covariance: Bayesian methods, sparse methods, and factor methods. He introduced a new factor method, Total Correlation Explanation (CorEx), which finds hidden factors that are uniquely informative about relationships. CorEx outperforms GLASSO, used in the sparse method, in the undersampled regime. And, Temporal CorEx performs significantly better than the state of the art and could also be used to detect anomalies. The most important aspect of this new method is the scalability and its ability to run on a large data set, Galstyan concluded.

[7] The Global Database of Events, Language, and Tone data set is available at https://www.gdeltproject.org/, accessed February 19, 2019.

[8] The Integrated Crisis Early Warning System dataverse is available at https://dataverse.harvard.edu/dataverse.xhtml?alias=icews, accessed February 19, 2019.

7

Plenary Session

TOWARD TRUSTWORTHY MACHINE LEARNING

Dawn Song, University of California, Berkeley

Dawn Song, University of California, Berkeley, explained that deep learning has advanced rapidly—for example, AlphaGo, a computer program, beat the world champion Go player, a human, and deep learning now powers everyday products. With this exponential growth in artificial intelligence (AI) and deep learning comes an increase in attacks, both in terms of scale and sophistication. These attacks are now occurring in new landscapes of the security arena, such as in the power grid and the banking system. It is important to consider the presence of attackers when thinking about machine learning, Song said. History reveals that attackers always follow in the footsteps of technology development or sometimes even lead it. The stakes are even higher with AI: as AI controls more systems, attackers will have greater incentives. And, as AI becomes more capable, the consequence of misuse by attackers will become more severe, she continued.

Song noted that attackers might try either to attack AI directly or to misuse it. When attackers attack the integrity of a system, they prevent the learning system from producing the intended or correct results and instead produce a targeted outcome, which they have designed. Attackers can also attack the confidentiality of a learning system in order to learn sensitive information about individuals. Song emphasized that better security in learning systems is needed to address these problems. When attackers misuse AI, they find vulnerabilities in other systems, target attacks, and devise attacks.

In order to prevent adversarial examples, the integrity of systems has to be protected. For example, self-driving cars need to recognize signs correctly in order to make safe decisions. If an attacker manipulates a stop sign with perturbations, thus creating an adversarial example, an image classification system can be fooled into thinking it is a speed limit sign instead, for example. Although most adversarial examples have arisen in the digital world, it is now possible to produce adversarial examples in the physical world (with physical perturbations). These real-world adversarial examples remain effective under different viewing distances, angles, and other conditions. This highlights the need to protect the integrity of a learning system so that it still generates the correct predictions or labels, even when under attack.

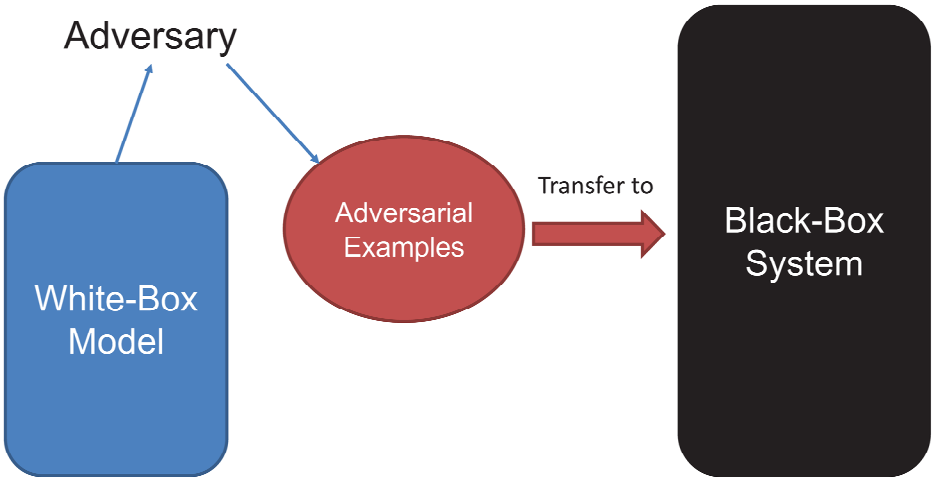


FIGURE 7.1 Transferability attack. SOURCE: Dawn Song, University of California, Berkeley, presentation to the workshop, December 12, 2018.

Song noted that adversarial examples are also prevalent in deep learning systems. Most existing work in this field is in image classification tasks and when the target model is known. Song's team is investigating adversarial examples through generative models, deep reinforcement learning, and visual question answering (VQA)/image-to-code. Her team also studies threat models—mostly white-box attacks in which the attacker knows the parameters of the neural networks, although adversarial attacks can also be effective on black-box models when the attacker does not know anything about the architecture. Her team discovered that state-of-the-art VQA models suffer from targeted adversarial attacks, with image perturbations that are typically undetectable by humans. Adversarial examples can also fool deep reinforcement learning agents (e.g., game-playing agents in Atari and the MuJoCo environment). Song emphasized that all of these examples indicate that adversarial attacks are prevalent in a variety of domains, tasks, and models.

Song explained that one way to generate adversarial examples is with generative adversarial networks. She noted that most of the examples she had discussed thus far were based on white-box attacks, in which it is assumed that the attackers know the details of the learning model (e.g., the architecture and parameters). Even when attackers do not know any details about the model, called a black-box attack, their attacks can be very powerful. Song described two types of black-box attacks. In a *zero-query attack*, which includes a transferability-based attack (see Figure 7.1), the attacker does not have query access to the target model and uses local information to generate a successful attack.

For the transferability-based attack, the attacker's objective is to attack a remote model. The attacker has local access to another learning system (i.e., white box) and generates adversarial examples to this local model. These adversarial examples can then succeed in transferring to the black-box system. To make a transferability attack especially effective, an attacker can use an ensemble targeted black-box attack based on transferability: the attacker can have an ensemble of different local white-box models (e.g., using AlexNet, ResNet, VGGNet, etc.) and generate targeted adversarial examples, fooling all of these different models. If an attacker generates adversarial examples in this way, there is increased likelihood that the examples will be transferred to the remote system to create a successful targeted attack on the remote model. She called this the most effective type of zero-query black-box attack. In a *query-based attack*, the attacker has query access to the target model. In this case, finite-difference gradient estimation and query-reduced gradient estimation are used to generate results that have similar effectiveness to white-box attacks. Using a query interface, an attacker can gain more information and attack more effectively, using finite-difference gradient estimation.

Song said that adversarial machine learning is about learning in the presence of adversaries. This can happen at inference time, when the adversarial example fools the learning system (i.e., gives the wrong

prediction) with evasion attacks (i.e., evades malware detection, fraud detection). It can also happen at training time, when the attacker poisons the data set to fool the learning system to learn the wrong model. The attacker selectively shows the learner the training data points (even with correct labels) to fool the learning system to learn the wrong model. Defending against data poisoning is particularly challenging with crowdsourcing and insider attacks. Overall, it is very difficult to detect when a model has been poisoned.

Adversarial machine learning is particularly important for security-critical systems. Song said that more than 100 defenses have been proposed over the past 18 months, but no sufficient defense exists today. Strong, adaptive attackers can easily evade today's defenses, and obfuscated gradients give a false sense of security. An ensemble of weak defenses does not, by default, lead to strong defenses. She explained that image classification is susceptible to adversarial examples just by the nature of the task specification, and she added that human vision is more complex than image classification.

Song's team proposed a new defense. She said that it is possible to characterize adversarial examples based on spatial consistency information for semantic segmentation. The learning system tries to segment an image into different objects. It is still simple to fool this learning system for segmentation with adversarial examples. The defense is based on spatial consistency (i.e., the consistency of segmentation results for randomly selected patches from an image). Such spatial consistency information from benign and adversarial instances is distinguishable. They apply mean intersection over union to compare the segmentation results between patches (Xiao et al., 2018). When one adds spatial and temporal constraints, it becomes difficult for attackers to generate adversarial perturbations, so this approach works for detecting adversarial examples.

Song emphasized that security will be one of the biggest challenges in deploying AI. She said that it is imperative to think about security at the software, learning, and distributed levels. It is challenging to ensure that no software vulnerabilities (e.g., buffer overflows and access control issues) exist; attackers can exploit such vulnerabilities and take control of learning systems. At the learning level, it is crucial to evaluate systems under both normal events and adversarial events. It is important to do both security testing and regression testing.

Another challenge at the learning level is to be able to reason about complex and non-symbolic programs. Currently, although reasoning techniques exist for symbolic programs, there are no sufficient tools to reason about non-symbolic programs, such as deep learning systems. Additionally, it is important to design new architectures and approaches with stronger generalization and security guarantees. Neural program synthesis is an exciting area of work, but the domain has had a number of challenges including the fact that neural programs do not generalize well and that they provide no proof of generalization. Song's approach to address these problems is to introduce recursion and to learn recursive neural programs. Recursion enables provable guarantees about neural programs and the ability to prove the perfect generalization of a learned recursive program via a verification procedure (i.e., explicitly testing on all possible base cases and reduction rules). This recursion approach also enables faster learning and generalization. This work revealed that neural program architecture impacts generalization and provability; recursive, modular neural architectures are easier to reason with, prove, and generalize; and designing new architectures and approaches enabling strong generalization and security properties for broader tasks is desirable and is a promising direction to explore.

Another challenge for security at the learning level is the ability to reason about how to compose components. Building large, complex systems requires compositional reasoning (i.e., each component provides abstraction; hierarchical, compositional reasoning proves properties of whole system). A question remains as to how to do abstraction and compositional reasoning for non-symbolic programs.

At the distributed level, where each agent makes local decisions, a question remains about how to make good local decisions that will lead to good global decisions.

In addition to attacking the integrity of a system, attackers can also attack its confidentiality, Song explained. Neural networks have high capacity, and attackers can exploit them to extract secrets in training data by querying learned models. For example, by simply querying a trained language model on an email data set that has users' credit card and social security numbers, an attacker could automatically extract the original social security numbers and credit card numbers. When training deep learning models, one has to

be careful with privacy protection for the training data set. Song's team proposed a solution to prevent such instances of memorization by training a differentially private neural network. In this case, the exposure is lessened and the attacker is unable to extract secrets.

Attackers can also misuse AI for large-scale automated, targeted manipulation. Thus, many questions remain about the future of machine learning systems and security: How do we better understand what security means for AI and learning systems? How do we detect when a learning system has been fooled or compromised? How do we build more resilient learning systems with stronger guarantees? How do we build privacy-preserving learning systems? How do we democratize AI? She emphasized that security will be one of the biggest challenges in deploying AI, and it requires a community effort.

A workshop participant said that the perturbations were visibly noticeable in Song's black-box attack examples and asked if she thought black-box attacks would get more sophisticated, with perturbations as indistinguishable as those in white-box attacks. Song reiterated that there are two types of black-box attacks: zero query and query based. The zero-query attack will always be much more challenging—the generated adversarial perturbations are larger in scale than in white-box attacks. With query-based attacks, the question becomes how many queries are allowed. With a larger number of queries, it is possible to generate better adversarial examples with less noticeable perturbations. A workshop participant asked why there is emphasis on defense from attacks rather than understanding why the underlying technology is flawed in the first place. Song said that deep learning systems are not learning the right representations. Although deep learning has issues, she continued, it is an approach that is currently working for vision tasks. Meanwhile, the community is continuing to search for better solutions that can address issues more fundamentally. In response to a question about Song's discussion of differential privacy, Song said that her recent work proposed a specific type of measurement to ascertain how much the neural network has remembered. Tom Goldstein, University of Maryland, asked if people have a moral responsibility in this field to propose a defense whenever proposing an attack. Song said that there are ethical standards about responsible disclosure when proposing attacks and added that proposing an attack without a defense is still progress.

8

Recent Trends in Machine Learning, Part 3

DOMAIN ADAPTATION

Judy Hoffman, Georgia Institute of Technology

Judy Hoffman, Georgia Institute of Technology, said that both the machine learning and computer vision communities believe that their models work well, particularly when they have access to a lot of data and the ability to use deep learning approaches to train the models. Looking at things such as challenge performance over time reinforces that notion. In the ImageNet classification challenge, Hoffman's team found that performance started to approach near perfect on the test set; this improvement came as a result of leveraging millions of training examples effectively and expanding capacities of models over time. While people expect similarly strong performance in real-world applications, such stellar performance is unlikely. This is disconcerting because the problem may be bias in the training data rather than any perturbations (as with adversarial examples). Classifiers are often trained with biased data that can be very different from what is found in real-world settings. For example, images scraped from social media sites are limited by what people choose to post on social media (e.g., images tend to have good lighting, composition, and resolution). Another example is image segmentation for autonomous vehicles, where bias can impact the interpretation of images (e.g., a model trained on images of roads from a sunny location may struggle to recognize snow-covered features).

To address these types of problems, domain adaptation can take a model that was trained under a set of biases and generalize or adapt it for use in new visual environments, without requiring significant human intervention. Domain adaptation requires access to a large, unlabeled data set representing a new environment that does not quite match the visual statistics of the original data set used to train the model. The model can adapt to this new data and learn to overcome differences and biases between the two visual environments.

Before deep learning was being used for domain adaptation, a common practice was to use a source representation, extract a high-dimensional vector that represented the data, and use that representation to learn a classifier on labeled source data. Classifiers were often blamed when there was a misalignment between the representation space on the source and the target. Earlier deep learning approaches relied on

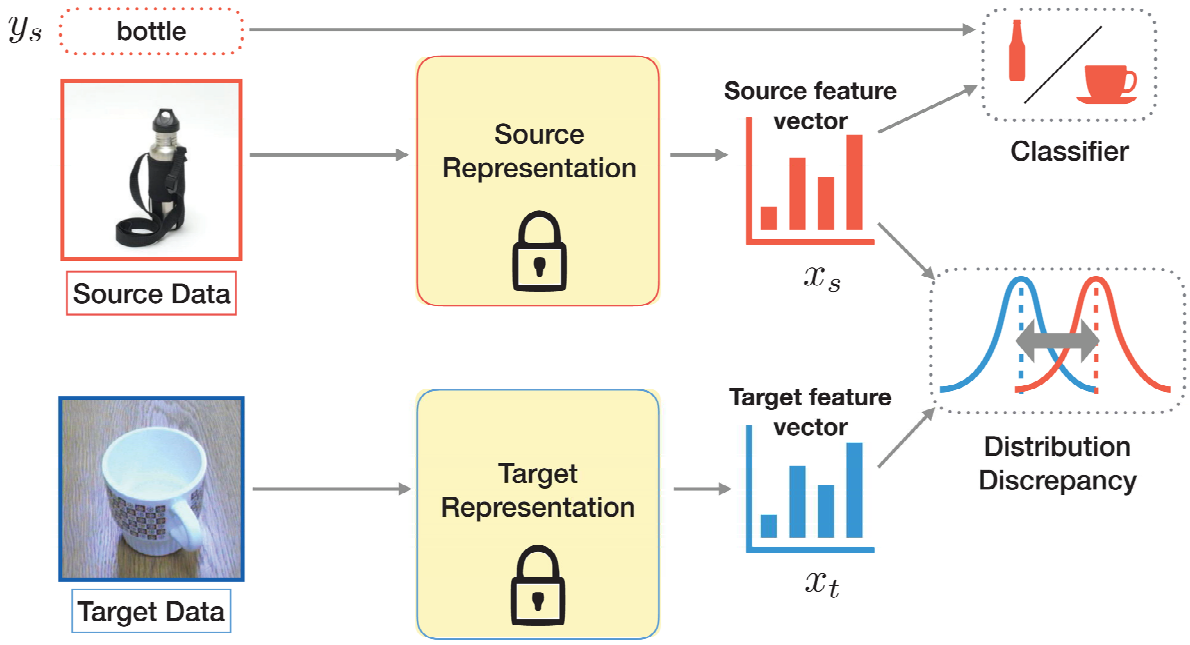


FIGURE 8.1 Using deep learning for domain adaptation. NOTE: CNN, convolutional neural network; x_s, source data; x_t, target data. SOURCE: Judy Hoffman, Georgia Institute of Technology, presentation to the workshop, December 12, 2018.

simple statistical techniques such as mean alignment; later, higher-order statistical alignment techniques were utilized. Today's technique bounds the target error by the discrepancy. The theoretical literature in domain adaptation allows for a probabilistic upper bound on the expected error in the deployment setting.

Now that deep domain adaptation is possible, images can be simultaneously represented and classified into relevant objects. This paradigm shift enables representation of images that are invariant to the differences between the first and the second domains. Instead of having a fixed function that separates the data into two distinct biases, end-to-end learnable functions can explicitly supervise those representations so as to be invariant to the differences between the two different collections of data. This can be done by defining a learning objective that aims to minimize the distribution mismatch (see Figure 8.1).

Hoffman emphasized that when trying to improve performance and reduce error in the deployment setting, it is possible to observe that the equation is dependent on how far apart the source and the target are, and it is possible to explicitly minimize that distance. The first challenge is trying to measure that distance; it is possible to empirically approximate the distance between the two distributions by looking at the error of the domain classifier. It is possible to estimate how far apart these two collections of data are by training a classifier in the representation space and evaluating the difference of the features of the two collections. Rama Chellappa, University of Maryland, College Park, asked how this is related to maximum mean discrepancy for domain adaptation, and Hoffman responded that this approach uses a different statistical alignment technique, different learning bounds, and a different algorithm but that the maximum mean discrepancy technique is also relatively easy to implement in practice.

By first training a classifier to differentiate between the source and the target, it is possible to create a minimization function; if it is possible to observe a difference, the collections are far apart in the feature space. Next, the underlying representation is updated by modifying the spatial distribution of the points. Because these data are labeled, these points can be relocated and a deep representation can be inferred. The question at hand is what happens when one tries to view a new collection of data under that same learned representation space. If the two collections are too far apart, the original classifier may not work as well on

the new unlabeled deployment set. In these cases, the distance between the two distributions can be approximated, the classifier can be trained to tell the difference between the target and source images, and the underlying representation can be updated. The object classifier that learned on the source images will now perform better on the unlabeled images from a new visual environment. This is iterated in high dimensions until it is no longer possible to learn a classifier to disambiguate the two sources of data. This gives a learning algorithm that does not require labels in the deployment setting but does make it possible to update the underlying representation so as to learn a domain-invariant feature space. She added that it is also possible to think about domain adversarial learning directly by aligning in pixel space, similar to the work of generative adversarial networks (GANs).

Hoffman shared an example of digit translation to demonstrate the key difference in image-to-image translation with GANs and domain adaptation. The goal is to recognize the different digits observed, assuming there are only labels in one paradigm of a Google Street View data set. This performance will be evaluated in the MNIST (Modified National Institute of Standards and Technology database) data set. With image-to-image translation techniques, it is possible to take the Google Street View and translate it to look like the MNIST digits. However, the standard GAN approach is difficult to optimize and there are issues to consider, such as mode collapse and production of blurry images. An alternate approach is the CycleGAN model, which introduces auxiliary losses that should improve upon the original optimization by adding a new learning objective. She said that it should also be possible to take the inverse stylization to reconstruct the original image, and then, ideally, there would be enough low-level structural information represented in the translated image to assume a better translation than that from the standard GAN model. This generates two learning objectives: a stylization and a reconstruction. However, there are failures with image-to-image translation techniques that can be catastrophic if the goal is to take an image, translate it, and use it to train a better classifier for a new visual environment.

Using the example of Google Street View digits again, Hoffman said that the learned translations are correctly stylized per MNIST, but the corresponding reconstructions have the wrong semantics. This means that it would not be possible to train a classifier on these images for the target setting. However, the end goal of this approach is not to produce the best possible looking image; rather, the goal is to use the image to perform some recognition task. In the large labeled data set, it is possible to train an initial classifier; this classifier does not have perfect performance on MNIST but has strong performance and can be a useful cue overall. Next, another objective is added specifying that the translation should retain the same semantics as judged by the weak classifier. The Google Street View images are translated into images that look like those from the MNIST data set. Since the translated images are not perfect, further feature space alignment is performed. Hoffman showed the performance results for digit recognition using different methods on the unlabeled target data set. A simple digital data set has a low accuracy at 67 percent. This shows how much performance can degrade across different visual environments. Hoffman's team's method, CyCADA, generated 90 percent accuracy. Even without using any labels from MNIST, it is possible to use statistical alignment techniques to improve performance from 67 to 90 percent.

Hoffman's team tried this same approach on a larger problem: semantic segmentation in driving settings. Collecting annotations of driving scenes is expensive and there is large potential for change in the scenes, owing to weather, location, and vehicle. She showed three different settings from these variations where one may opt to use automotive adaptation instead of requiring new labels to recognize similar but different scenes: (1) In the cross-city adaptation example, she trained on the CITYSCAPES data set and looked at performance in testing on a San Francisco data set. Differences exist in the signs, tunnels, and road sizes. Before adaptation, a building is confused with sidewalk/road; after adaptation, the building is recognized correctly. (2) In the cross-season adaptation example, she trained on a fall scene and tested on a winter scene, both synthetically rendered in the SYNTHIA data set. Before adaptation, there was confusion on the parts of the scene covered in snow; after adaptation, the overall statistics of the image were used to adapt the model to understand the environmental change. (3) In the synthetic-to-real pixel adaptation example, she was learning directly from simulated driving imagery using a Grand Theft Auto scene to train to learn to recognize things in a real scene from CITYSCAPES. She took images from the simulated imagery, learned to make them look more realistic, and then performed feature space alignment

to get a real-world model. Adaptation makes it possible to recognize the images more effectively by using overall marginal statistical techniques. Using the mean intersection over the union of the data, it is possible to improve raw pixel accuracy using unsupervised alignment techniques. The algorithm looks at the pixel level and the feature level and examines how to best make changes to the representation so as to still perform original classification tasks but also to learn biases in invariance space.

Hoffman said that the objective of adversarial domain adaptation is to minimize the discrepancy between the source and the target. This can be computed in feature space and/or image space, but she cautioned against letting the domain classifier confuse the model.

Next she turned to a brief discussion of deploying similar techniques in a setting that continues over time and in which a visual scene changes when the task does not (e.g., a camera pointed at a traffic intersection). It is possible to apply domain adversarial learning, but new research questions surfaces. The goal is to have a model that generalizes across different settings, remembers things previously known, reacts when new situations arise, learns quickly from few examples, and proposes scalable solutions. Aram Galstyan, University of Southern California, asked if connections exist to tools for invariance representation learning. Hoffman said that there are connections, and others have also looked at these settings. She said that her research is unique in that she is examining the classifier and the invariance at the same time. She described this as a semi-supervised learning paradigm where performance is important. She said that the most relevant paradigms are those that come from the adaptation community. Tom Goldstein, University of Maryland, wondered if the adaptation is responsible for greying-out the images. Hoffman said yes and noted that image-to-image translation techniques do not have vibrant colorization. He suggested combating that problem by changing the loss function, and Hoffman noted that there are likely additional steps that need to happen. Chellappa asked about short- and long-term issues in the field. Hoffman emphasized the value of rethinking the notion of domains in adaptation literature: What happens before going to continuous? What happens if someone provides a collection of data sets? What is the best way to combine all of this information? She said that there is also a need to address the overall robustness in the adaptation literature, with a focus on better controlling the initial model so it is less susceptible to natural or artificial changes.

EXPLAINABLE MACHINE LEARNING

Anna Rohrbach, University of California, Berkeley

Anna Rohrbach, University of California, Berkeley, discussed techniques that can be useful to expose and combat bias and other problems faced by analysts. Although artificial intelligence (AI) can make fairly accurate predictions, questions arise about whether those predictions can and should be trusted. AI often relies on black boxes, about which the user has no information. She suggested that techniques would be adopted more widely if AI could explain its predictions, thus increasing trust and improving interpretability. Rohrbach and her team's goal is to improve the interpretability of deep learning by making more introspective modular architectures and by building models that can justify their predictions for users through "storytelling." Explainability and performance both stand to be improved owing to these measures. Deep models may perform better when augmented with explainable AI (XAI). Target domains include image and video analysis, vehicle control, and strategic games.

Rohrbach discussed three of her team's innovations during her presentation. The first project is developing more modular, introspective architectures to aid in reasoning. When humans are confronted with a question, they take a series of reasoning steps and analyze relationships before attempting to answer the question. A system can learn to do such step-wise reasoning and deliver human-understandable outputs. She commented on the history of work in this area that was based on supervised layouts for processing the questions. These relied on an external parser to analyze the question and build "expert" layouts to supervise the module choice. As a comparison, the current model has differentiable module choices that are inherently explainable, and the question is represented through recurrent neural networks. The model is end-to-end

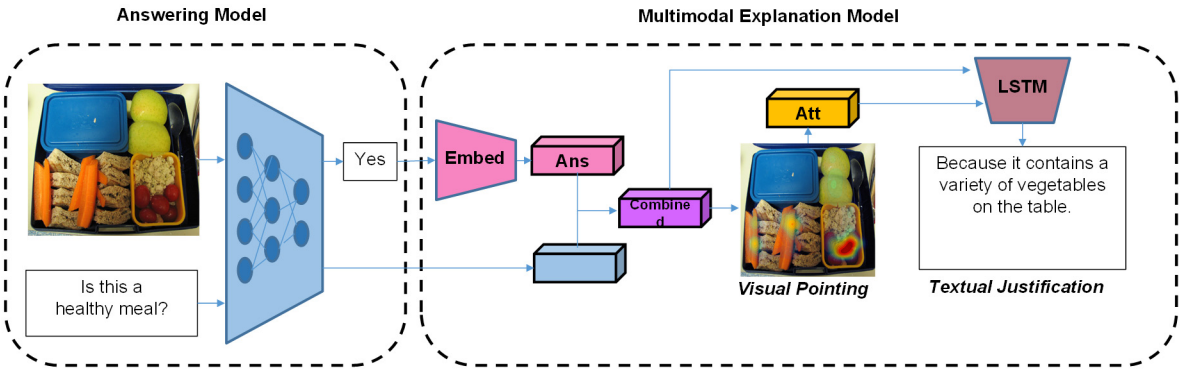


FIGURE 8.2 Example of textual justification with grounding and rational pragmatics. NOTE: LSTM, long short-term memory; Att, attention. SOURCE: Anna Rohrbach, University of California, Berkeley, presentation to the workshop, December 12, 2018. Photo of lunch from the Microsoft Common Objects in Context (MS-COCO) data set (https://farm4.staticflickr.com/3195/3078492451_35988d2062_z.jpg) under the license https://creativecommons.org/licenses/by-nc-sa/2.0.

trainable with gradient descent so that no reinforcement learning is needed. It can also work on multiple tasks and has compositional behavior without requiring layout supervision. All of the modules contribute to the answer but in a different degree. In a human evaluation, the human users rate the model as "clear" in terms of understandable reasoning; the flexible architecture is inherently interpretable by design.

The second project Rohrbach discussed was about textual justification, which is a post-hoc explanation. Providing justifications for responses and pointing to evidence are trivial tasks for humans but not for AI. She explained that any given visual question answering (VQA) architecture or answering model could be augmented with a multi-modal explanation model, which will condition on the visual and question representations, as well as the answer, to predict the textual justification, which will be grounded through the visual pointing mechanism. The only supervision comes from the text that was provided by the human. By doing this, the system can learn to provide human-like justifications (see Figure 8.2).

Rohrbach and her team also found that humans are better at predicting a model's success or failure when given an explanation; humans benefit from the machine-generated explanation because it helps them decide whether to trust the machine.

The third project Rohrbach presented was about causal factors for visuomotor policies. The objective of this project was to predict vehicle motion and communicate to the human in natural language why a certain driving behavior occurred (i.e., generating textual explanation of how visual evidence is compatible with a decision). The team collected data from the Berkeley DeepDrive data set (a large, crowdsourced collection of driving videos from human drivers) and augmented those with descriptions and explanations. The desired outcome is to have explanations that align with whatever the vehicle controller is seeing in that moment as opposed to providing post-hoc explanations. This could be conveyed as explanations with strongly aligned attention or weakly aligned attention, with the latter working best because it allows the system the freedom to find other evidence to mention, to which the controller does not necessarily need to pay attention. The goal is to develop explanations that are introspective and conditioned on the internal representation of the controller. However, there is currently no temporal reasoning across all of the video segments, which can lead to mistakes in explanations. In the future, to get the system to focus on more interesting but perhaps less common events or objects (e.g., pedestrians), it is important to bias the system toward categories that are more significant for humans.

Summarizing the results of these three projects, Rohrbach said that the multi-step introspection model learns to reason and allows users to better predict model performance via intermediate visualizations. She said that this is a goal for all systems but is not trivial to adapt to other domains. She said that the textual justification model can predict expert rationales in a variety of domains and is useful for understanding

category differences. She added that causal XAI provides introspection into driving behavior and helps obtain grounded human understandable explanations.

A workshop participant asked Rohrbach if her team has tried applying the justification to naturally occurring or intentionally generated adversarial images. Rohrbach said that although they have not tried that yet, they have done something related for VQA. They have performed an adversarial attack and visualized the attention to see where the model is "looking." The attack forces it to find evidence to justify the incorrect answer (i.e., it often fools the model in an interpretable way). A workshop participant observed that the explanations Rohrbach's system gave appear to be based on things one can see in the image at the time, as opposed to common sense knowledge. Rohrbach said that both types of information are used and the explanation tries to be relevant in context. Although the systems do not get external common sense, the machines pick up patterns from training data. The ultimate goal is for the system to be relevant to the image/video instead of only providing a plausible common sense explanation. Although additional external knowledge has not yet been injected, that could be an interesting future direction.

Rohrbach then turned to a discussion of diagnosing and correcting bias in captioning models. Explanations could be useful to expose and measure bias in systems. She said that useful priors are necessary for learning (i.e., all machine learning methods learn by exploiting correlations and patterns present in training data). Failures can occur if the task domain changes or if the domain remains the same but the data distribution changes. This sometimes leads to incorrect or even offensive predictions. She described how captioning models exaggerate imbalances in training data and amplify bias—for example, an image of a woman can be labeled as a man, owing to the presence of a desk and a computer in the scene. In this case, the system captured biases and spurious correlations. One goal of Rohrbach's work is to improve gender bias in deep captioning models. The prediction of gender in captions needs to be based on appropriate evidence. In order to make an accurate gender prediction using a standard cross entropy loss for image captioning, the captioning pipeline is retained while additional losses are introduced to correct the problem of gender misclassification. The results are categorized by error rate (i.e., misclassification of women as men and vice versa), gender ratio (i.e., women/men in predicted captions), and being right for the right reasons (i.e., the model is looking at the people in the image). The Equalizer model reduces the distance to the gender ratio (i.e., the difference between the ratio of man to woman predicted and the ratio in ground truth captions) and error rate (i.e., percent of gendered words predicted incorrectly) and is right for the right reasons more often than the baselines. If evidence is weak and gender is uncertain, the label "person" will be used in the prediction.

A workshop participant suggested that such biases are a property of the data set, and any combination of properties would likely have this same flaw. Rohrbach said that in the case of gender, there are societal implications pertaining to fairness. Here, the goal is to avoid the amplification of biases, to avoid spurious correlations that lead to errors, and to be fairer to the actual data set distribution. Another audience participant asked if randomness could be injected into the process to make the neural network more powerful. Rohrbach said that vision and language each carry responsibility for these kinds of errors and that it is difficult to uncover where the error originates. One will always have to deal with multimodality, so it is difficult to determine whether that would be a solution. She said that there is a way for the process to propagate priors or biases in non-obvious ways: people have tried rebalancing the training data, but Rohrbach does not think this is a scalable and reasonable solution because there are many different biases and correlations, and they cannot all be balanced. An audience participant asked if the labels are more accurate with her approach, and Rohrbach responded that they have become more accurate in terms of gender while maintaining caption quality. In response to a follow-up question as to whether there is an advantage to using this approach over simply rebalancing the data, Rohrbach said that, yes, this approach is advantageous over the baseline because the baseline carries and amplifies all of the training biases and does not adapt to new distributions. A workshop participant asked how the system could dynamically rebalance the bias. Rohrbach reiterated that she trains a single system on the given training data, which has its biases, but evaluates on different test sets. She said that the system adapts well to different test distributions. She explained that the system is now looking at the person rather than relying on context correlations. She emphasized that correcting this behavior alone is an important advance. A workshop

participant said that this goes beyond an issue of bias in a particular data set during training or test and has more to do with training using only one objective (i.e., the overall captioning loss, which combines multiple sources of evidence in the data). The right thing to do would be to see if, for example, an umbrella always corresponds with a woman; that could be a valuable cue. Hoffman added that the goal is to have more modular models that pay attention to different cues and report on each of them independently. She said that it sounds like Rohrbach is introducing an auxiliary loss that allows one to focus on relevant details of what should be used to determine gender. Rohrbach said that while in some cases the correlations are desired and it is acceptable to use them, there could be ramifications from the public about such stereotypes.

Rohrbach provided a very brief overview of change captioning before concluding her presentation. She noted that although data analysts want to use AI, there are obstacles across multiple domains in understanding changes over time. Although not everything that changes would be relevant to an analyst, it is important to recognize semantic changes: semantic change detection and captioning is a way to detect changes between two different images of the same scene taken at different times. The goal is to be able to summarize the change that occurred using the tools from image captioning and other vision or language techniques to discover and explain the change over time. Changes in scenes can be nontrivial to recognize. Qualitative results reveal that one cannot rely on single attention, because if something is no longer present, the system has a difficult time understanding that change and associating the same object in two scenes. So, the approach she described has two attention mechanisms to discover the change and highlight in an explainable manner where the right evidence is located. In summary, the Dual Dynamic Attention Model is used to localize and describe changes between images. She noted that her team's model is robust to viewpoint changes and its dynamic attention scheme is superior to the baselines.

9

Machine Learning Systems

BUILDING DOMAIN-SPECIFIC KNOWLEDGE WITH HUMAN-IN-THE-LOOP

Yunyao Li, IBM Corporation

Yunyao Li, IBM Corporation, focused on systems that IBM has built that leverage the human in the loop. She noted that IBM is building technologies spanning multiple steps in the artificial intelligence (AI) life cycle: knowledge representation, creation, and refinement. These steps include the following:

- Document ingestion, in which the document is converted into a machine-consumable form;
- Capture of domain-specific knowledge, including vocabulary, constraints, logical expressions, rules, and domain schema;
- Document understanding, during which knowledge is extracted from individual documents;
- Integration of (un/semi-structured) information across data, using entity understanding and resolution;
- Evaluation of knowledge quality to validate the model's performance over in-domain and out-of-domain areas; and
- Knowledge base consumption when building predictive models and query abstraction for programmatic/human access to knowledge bases.

The objectives for creating industry-specific knowledge applications are to (1) build AI using scalable and explainable tools to cover these life-cycle steps by including human-in-the-loop to capture knowledge from domain experts, knowledge workers, and end users, with systems supporting provenance, debugging, and error analysis and (2) to construct and refine domain knowledge using machine learning and deep learning techniques. These technologies can be applied in a variety of industry domains, including finance, health care, compliance, and national security.

Li described two IBM systems that assist in knowledge representation and creation with a human in the loop: SystemT[1] for enterprise text understanding and SystemER[2] for entity understanding and resolution. She shared a use case for building financial entity profiles from various companies' Security and Exchange Commission and Federal Deposit Insurance Corporation filings from over a 20-year span. First, information must be extracted from the individual documents; then one must construct an entity-centric view with information about the individual companies to understand how different financial companies deal with different learning risks. Domain modeling is the first building block that enables the creation of this domain-specific knowledge base. This is important because it is impossible to learn everything from data only—a domain expert is needed to help understand the domain, including the domain schema (i.e., key entities and relationships that will help address the question at hand), hard and soft constraints, and vocabulary (e.g., similar names might have different metrics, while similar metrics might have different names).

Backend analytics is the second building block that enables the creation of a domain-specific knowledge base; this is about building and maintaining the models and algorithms to understand the individual documents and link them to each other. IBM takes a layered approach to accomplish this: the application layer, models and algorithms layer, computing platform layer, and hardware layer, one built upon the next. The main focus is to attain key categories of tasks to solve in the models and algorithms layer (e.g., natural language processing) and define a domain-specific language that captures the primitives to such tasks so that it is easier to program and apply automatic optimization. This also enables platform independence.

One backend analytics system uses domain-specific language for natural language processing. The SystemT architecture has Annotation Query Language to specify extractor semantics declaratively, a compiler and optimizer to choose a different execution plan that implements semantics, and operator run time. The fundamental results and theorems show that the language is expressive and performs well. Given that language understanding is the foundation of many AI applications, an alternative view of this system shows the layered semantic linguistic abstractions. The abstraction operators can be very simple; the basic operators are at a syntactical level. She added that it could also be more complex, when the semantic meaning of every sentence is understood. The cross-lingual semantic analysis capability is embedded as part of the system. More sophisticated operators allow people with domain knowledge to easily build domain-specific extractors. For example, to categorize each sentence in a contract for IBM, a domain expert can use this semantic layer to write highly precise algorithms with abstraction that are more portable than doing a deep learning model. IBM has also done some work on using deep learning to automatically learn transparent domain-specific analysis that enables domain experts to co-create meaningful interpretable models.

Li emphasized that IBM works on important real-world problems by doing long-term, well-funded research, which makes an impact in business, science, and education. SystemT, for example, ships with more than 10 IBM products, has resulted in more than 50 research papers, and is taught in universities and online courses. The ultimate goal is to use machine learning to help more customers to use IBM's systems faster and for a broader range of use cases.

She then explained SystemER, another backend analytics system that uses a high-level integration language for entity resolution. She explained that entity resolution is used to build curated knowledge from hard data that can be loosely structured, heterogeneous, and sparse. Its goal is to infer explicit links among entities that otherwise would remain hidden in the data. As an example, in one application, Li constructed author profiles from publication records in medical research by identifying and linking all of the occurrences of the same author across different publications. In another example of using an entity resolution system to integrate all of the information about an entity, Li reiterated that matching across

[1] The website for SystemT is https://researcher.watson.ibm.com/researcher/view_group.php?id=1264, accessed February 19, 2019.

[2] For more information, see IBM Corporation, "Medium Energy Ion Scattering" https://researcher.watson.ibm.com/researcher/view_group.php?id=2171, accessed May 31, 2019.

sparse heterogeneous data sets is difficult and ambiguous. The goal is to understand the structure of various entry attributes, learn different combinations of attributes whenever available, and leverage new types of matching functions, which may exploit additional knowledge. To express, reuse, and relearn these entity resolution operations, a domain-specific, high-level integration language is needed. It is possible to generate algorithms that can do entity resolution automatically, which are comparable to a manually curated data set but with less effort.

The third building block to enable the creation of this domain-specific knowledge base is the human in the loop. The human can be involved in data labeling, model development, and deployment and feedback. As an example, high-quality, expert-level labeled data can be produced at low cost using auto-generation and crowdsourcing and then used for semantic role labeling, which indicates who did what to whom, when, where, and how. Li said that the goal is to develop a cross-lingual capability for this kind of representation, but obtaining labeled data for semantic role labeling is challenging. There are difficulties with both generalization and models: linguistic expertise as well as language or domain expertise is required, and semantic role labeling models are often black-box models with high complexity, so a traditional active learning framework does not fit. To generate high-quality labeled data for semantic role labeling, researchers have leveraged the preexisting high resources for English and parallel corpora through automatic generation, but this is still insufficient. Expert curation with active learning has also been used, although that process still generates some errors. Because not all of the tasks are equally difficult, crowdsourcing has also been employed to create high-quality training data. By developing classifiers that can automatically determine the level of difficulty of a task, the expert effort is reduced and results can be improved because experts curate difficult tasks and the crowd curates easy tasks.

If humans are in the loop for model development, it is possible to develop self-explaining models in a target language that humans can manipulate as well as reduce the amount of labeled data required with transfer learning and active learning. For instance, having a human in the loop to learn extraction by example is particularly useful in creating rules with higher accuracy more quickly. The user would load a document, highlight text, label positive or negative examples, and then ask the system to suggest some rules. The system has automatically learned rules and makes suggestions. The user can browse each rule and either accept a rule or provide feedback to the system to refine the rule further. With this tool, SystemT creates extractors that are comparable in quality to those of a human expert. Another example she shared used active learning for entity resolution. For a typical machine learning method (i.e., creation of training data, feature engineering, and entity resolution model learning), all steps require a significant amount of human effort; the goal is to reduce the labeling effort as much as possible by automating and learning some of the complex features. Active learning significantly reduces human effort in training data creation, learning matching and normalization functions, and learning high-accuracy entity resolution models. This active learning–based approach enables the completion of use cases in a few days or weeks instead of in a few months. During the deployment/feedback stage, the end user provides feedback on the AI services and influences the entire AI life cycle, from data acquisition to model development. Li explained that utilizing AI accelerates model improvement.

Li provided a brief overview of how extraction and entity resolution are used to build a financial knowledge base using data sources (both structured and unstructured) from the Security and Exchange Commission and the Federal Financial Institutions Examination Council. The first challenge to overcome is that it is difficult to identify and integrate information distributed across a document. Ultimately, it is possible to build a fairly large financial content knowledge base and help answer questions of importance to financial domain experts. This financial knowledge base can help in solving other business problems, such as comparing industry key performance indicators and understanding counterparty relationships and loan exposure. In closing, Li reiterated that having a human in the loop is very important. All of the tools available today (e.g., deep learning, reinforcement learning) can be leveraged, but the concept of human-in-the-loop needs to be explored further. Because the human is the customer, it is important to leverage domain expertise instead of turning everything into a labeling problem. Aram Galstyan, University of Southern California, asked how much effort it would take to generalize this to multilingual settings, and Li responded that the effort could be divided into first building the foundational technology (which would

require years of effort) and then doing the language adaptation (which should only take a few months). She said IBM will test this next year, and the ultimate goal is to do all the difficult work for the consumers.

ROBUST DESIGNS OF MACHINE LEARNING SYSTEMS

Anthony Hoogs, Kitware, Inc.

Anthony Hoogs, Kitware, Inc., opened his presentation by giving the audience a broad perspective on deep learning, open-source software and data, and machine learning systems. He then moved to a discussion of case studies for interactive machine learning, before concluding with his thoughts about relevance to and prognostications for the Intelligence Community (IC). The commercial industry has made substantial investments in AI and has had great successes, but it can be difficult to transfer these technological successes into the IC domain. Project Maven was one effort to energize machine learning and get it quickly fielded into operations in the Department of Defense (DoD), but one of the problems faced over and over again is that commercial data (e.g., from Internet videos) does not look like military data (e.g., from surveillance videos), so methods do not transfer across domains. How these differences are dealt with has been a long-running theme, Hoogs mentioned. Because mission-critical life and death situations are more apparent in defense and intelligence communities, robustness in AI is essential.

Hoogs described open source software and data as "the unsung hero of deep learning." Everyone focuses on graphical processing units, but open-source software, specifically Caffe,[3] from the University of California, Berkeley, is truly responsible for the deep learning revolution. Open-source practice in the deep learning world enables a model to operate within industry or government within a few months. This facilitates transition across domains, has fueled more development, and allows engineering-level applications to be released more quickly.

Kitware is an open-source company with several toolkits including Kitware Imagery and Video Exploitation and Retrieval (KWIVER)[4] and Video and Imagery Analytics for the Marine Environment (VIAME),[5] which is a specialized version of KWIVER for underwater environments. KWIVER has a number of toolkits loosely coupled together; one feature is interactive machine learning. He reiterated that most of what he shared in this presentation is openly available at the source code level.

He noted that deep learning enables engineering and operations for machine learning problems; these systems need to be assured and robust. The success of adversarial attacks does not mean that machine learning is brittle, he continued. As applied research moves toward operations, there is an increased focus on data provenance, model assurance, guaranteed performance, and explainability, which are especially challenging problems for complex models. He encouraged funding research in those areas, especially considering the need for robust systems.

With the VIAME toolkit, the goal was to develop an open source software platform for the National Marine Fisheries Science Centers' image and video analysis in close coordination with the National Oceanic and Atmospheric Administration (NOAA). He explained that VIAME is having great impact across NOAA. There are many analogues for the IC, as this is a good use case of interactive machine learning with a straightforward and simple method. NOAA has six National Marine Fisheries Science Centers in the United States, and each was doing some form of underwater data collection with camera rigs at different depths for fisheries stock assessment and then developing algorithms to analyze these data. However, they had a problem trying to classify fish from hours of footage. NOAA's Strategic Initiative (2013-2018) on Automated Image Analysis had a mission to develop guidelines, set priorities, and fund projects to develop broad-scale, standardized, efficient automated analysis of still and video imagery for

[3] To learn more about Caffe, see University of California, Berkeley, "Caffe" http://caffe.berkeleyvision.org/, accessed February 19, 2019.

[4] For more information about KWIVER, see http://www.kwiver.org/, accessed February 19, 2019.

[5] For more information about VIAME, see http://www.viametoolkit.org/, accessed February 19, 2019.

use in stock assessment. This work began with a National Academies of Sciences, Engineering, and Medicine-hosted workshop in 2014, which convened computer vision experts and produced a summary report (NRC, 2015). Organizations such as NOAA have important societal problems, which are often more appealing to academic researchers then IC problems, but NOAA's types of problems are not well funded.

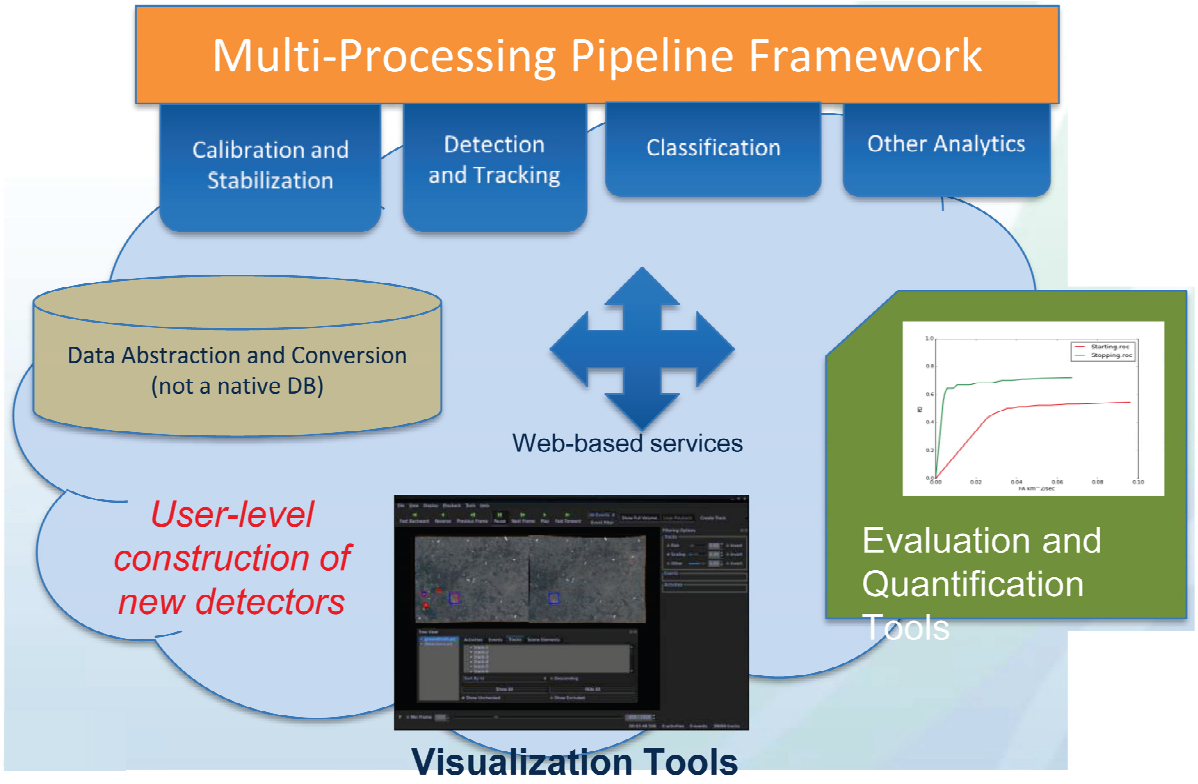
Hoogs gave an overview of NOAA's example data collectors; data streams exceed the capabilities of human analysts, so automated tools must be developed to increase the speed of analysis, reduce costs, and improve assessments. These data collectors include towed-camera or towed-diver benthic surveys, remotely operated vehicle fish surveys, net camera platforms, stereo-camera platforms, and animal body cameras. As impactful as deep learning is, Hoogs continued, counting and classifying all fish still exceeds the state of the art; algorithms still cannot do what humans can do, even with the stereo-camera platforms. Kitware built, installed, and trained users on VIAME at all six fisheries centers to help with their analyses of all of these data collection videos. VIAME utilizes deep learning and a variety of other methods. It has databases to store data, user interfaces, and tools for quantification (see Figure 9.1a). What is most challenging is that this system has to fit in and around systems that the fisheries centers already use.

Hoogs next described the components of the VIAME system (see Figure 9.1b). He said that there are two deep learning ways to build classifiers: classifiers can apply to an entire frame or to a detection paradigm. Kitware has supplied and built a generic fish detector, training across a variety of fish instances from the cameras mentioned previously. The users have the ability to build new detectors and new classifiers on their own. Once they create those, they can run them on arbitrary amounts of new data. They also have the ability to do specialized things such as stereo measurements. Kitware has also supplied user interfaces to look at outputs.

Hoogs explained that, from the interactive machine learning perspective, there are two options. The first option is to label the data, train a deep learning model, run it, see how it works, add more training data, balance training data, and repeat. With the second option—the interactive classifier construction method—users can solve simpler problems easier and quicker for themselves. They start with an image archive, run a generic object detector generator to detect anything a biologist could be interested in, get many objects, perform classification on each object, and then each goes into an archive. The goal is to have a lot of data, and then it is possible to query and train on these data, as well as do analysis on these data. Once the archive is complete, queries begin (i.e., start with an image example, find similar ones, get results, give feedback on results, build a classifier online based on results, go back to archive to re-rank results, and improve). Hoogs mentioned that knowledge of deep learning is not necessary to use this system.

For image and video search process, the system produces bounding boxes around particular fish and are compared to similar fish imagery in the archive. The query runs interactively; the task is to compute the deep learning feature vector and use it to find the nearest neighbors in the archive. This process takes only a few seconds before top results are generated. The user will then select and label which images are correct and which are not. Offline and unknown to the user, Kitware builds a support vector machine to differentiate between the positive and negative examples to help refine results. The user can repeat this process until he/she is satisfied with the result. People have also used the VIAME system to detect seals through aerial imagery—for full-frame classification and for object detection adjudication and annotation. In this domain, the data provenance is known because the fisheries centers collect the images themselves, which makes the system more robust overall. This also gives the users adequate control over the training set. Users, however, create an interesting bias since they are creating classifiers interactively, and Kitware is working on methods with the fisheries centers to make sure they understand the problem of bias and the importance of validation.

NOAA's problems have relevance to the IC. They have similar problems with fine-grained classification, instance counting, aerial/overhead imagery with resolution challenges, camouflage, complex backgrounds, busy scenes, and a wide range of image and video types. Environmental monitoring is a compelling surrogate problem for IC and DoD problems, he continued. However, the only way that people will work on these problems is with increased funding.



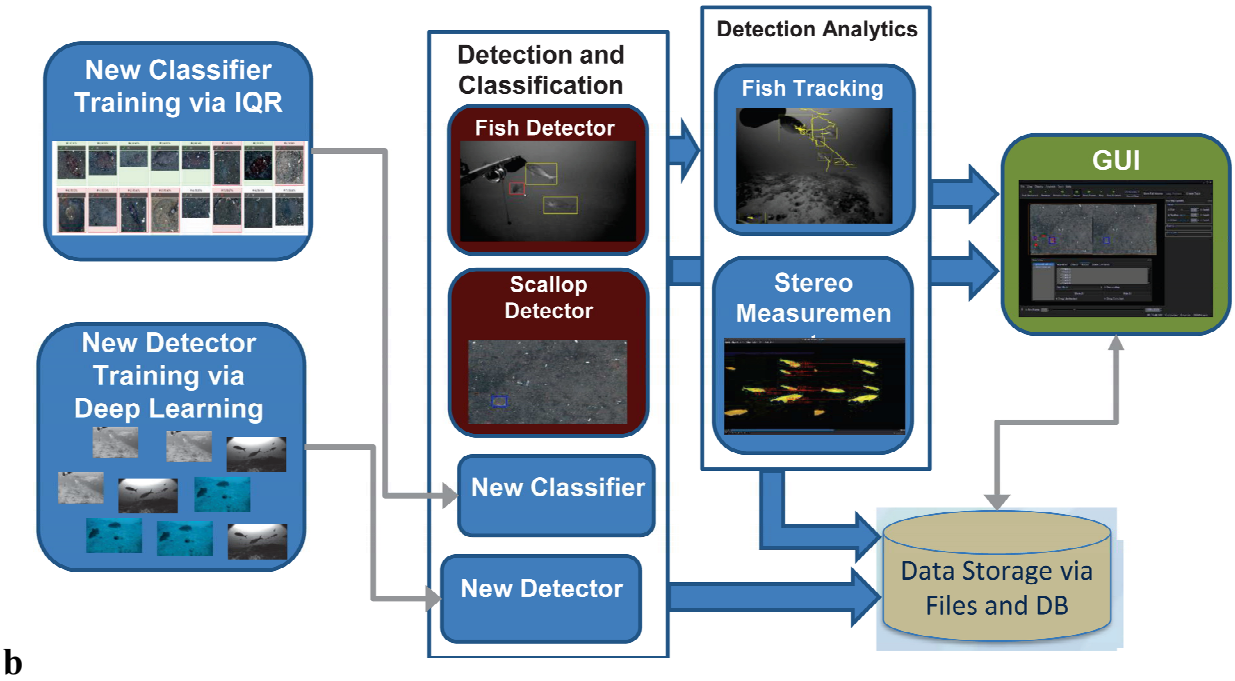


FIGURE 9.1 (a) The high-level architecture of VIAME's system and (b) VIAME's components. SOURCE: Anthony Hoogs, Kitware, Inc., presentation to the workshop, December 12, 2018.

Hoogs described another system, the Visual Global Intelligence and Analytics Toolkit, which is used for satellite imagery. The software is almost the same as that in VIAME, and the domains and problems are similar. The interactive system enables search for entities that do not have models (e.g., surface-to-air missile site detection). Hoogs also gave an overview of the Defense Advanced Research Projects Agency's Explainable AI (XAI) program.[6] In this program, researchers aimed to explain image querying while using a saliency algorithm designed for image classification and created a new algorithm for image matching. This algorithm is helpful when there are multiple objects in a scene and the user is running a query. The user can select images with correct matching and proper justifications. This technology is still a work in progress and does not work well if scene content is more salient than the desired object. There are a number of robust machine learning systems for the IC. User adaptation, extension, and specialization is a critical capability since IC analysts have highly specialized, unique problems. A question arises as to whether machine learning introduces significant robustness vulnerabilities for the IC, but Hoogs is not convinced it does. He said, however, that there are user biases, unvetted data, and validation challenges to consider.

In summary, Hoogs said that open-source software, models, and data have enabled the deep learning revolution and should continue to be supported. He stressed that IC funding should leverage (rather than compete with) commercial development and transfer expertise, models, and data to IC problems. The machine learning and computer vision research communities are experiencing typical growing pains as engineering and operations expand exponentially. He reiterated the importance of engaging academic researchers who may be interested in IC/DoD problems with compelling challenges and complete data sets. While interactive AI for online specialization is very promising, he noted that deep learning has had less impact, thus far, on high-level reasoning and contextual problems than image recognition. Data provenance and assurance continue to be important. Open research questions include how best to defeat physical adversarial attacks, understand black-box attacks, and infer the relationship among unbalanced training sets, rare objects and events, anomaly detection, and adversarial attacks.

[6] For more information about the XAI program, see M. Turek, "Explainable Artificial Intelligence (XAI)," DARPA, https://www.darpa.mil/program/explainable-artificial-intelligence, accessed February 19, 2019.

References

Bartsch, M.V., K. Loewe, C. Merkel, H.-J. Heinze, M.A. Schoenfeld, J.K. Tsotsos, and J.M. Hopf. 2017. Attention to color sharpens neural population tuning via feedback processing in the human visual cortex hierarchy. *Journal of Neuroscience* 37(43):10346-10357.

Bendale, A., and T.E. Boult. 2016. "Towards Open-Set Deep Networks." Pp. 1563-1572 in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. https://www.cv-foundation.org/openaccess/CVPR2016.py.

Boehler, C.N., J.K. Tsotsos, M. Schoenfeld, H.-J. Heinze, and J.-M. Hopf. 2009. The center-surround profile of the focus of attention arises from recurrent processing in visual cortex. *Cerebral Cortex* 19:982-991.

Conotter, V., J. O'Brien, and H. Farid. 2012. Exposing digital forgeries in ballistic motion. *IEEE Transactions on Information Forensics and Security* 7(1):283-296.

Cutzu, F., and J.K. Tsotsos. 2003. The selective tuning model of visual attention: Testing the predictions arising from the inhibitory surround mechanism. *Vision Research* 205-219.

Fawzi, A., H. Fawzi, and O. Fawzi. 2018. "Adversarial Vulnerability for Any Classifier." https://arxiv.org/abs/1802.08686.

Fukushima, K. 1975. Cognitron: A self-organizing multilayered neural network. *Biological Cybernetics* 20(3-4):121-136.

Gibson, J.J. 1950. *The Perception of the Visual World*. Boston, MA: Houghton Mifflin.

Hopf, J.-M., C.N. Boehler, S.J. Luck, J.K. Tsotsos, H.-J. Heinze, and M.A. Schoenfeld. 2006. Direct neurophysiological evidence for spatial suppression surrounding the focus of attention in vision. *Proceedings of the National Academy of Sciences* 103(4):1053-1058.

Julesz, B. 1971. *Foundations of Cyclopean Perception*. Chicago, IL: University of Chicago Press.

Krizhevsky, A., I. Sutskever, and G.E. Hinton. 2017. ImageNet classification with deep convolutional networks. *Communications of the ACM* 60(6): 84-90.

Lamb, A., J. Binas, A. Goyal, D. Serdyuk, S. Subramanian, I. Mitliagkas, and Y. Bengio. 2018. "Fortified Networks: Improving the Robustness of Deep Networks by Modeling the Manifold of Hidden Representations." arXiv preprint arXiv:1804.02485.

Liew, S.S., M. Khalik-Hani, and R. Bakkteri. 2016. Bounded activation functions for enhanced training stability of deep neural networks on visual pattern recognition problems. *Neurocomputing* 216:718-734.

Marr, D., and T. Poggio. 1979. A theory of human stereo vision. *Proceedings of the Royal Society of London B* 204:301-328.

Matthew, R.P., S. Seko, J. Bailey, R. Bajcsy, and J. Lotz. 2018. "Tracking Kinematic and Kinetic Measures of Sit-to-Stand Using an Instrumented Spine Orthosis." Pp. 1-5 in *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. https://ieeexplore.ieee.org/xpl/conhome/8471725/proceeding.

Meng, D., and H. Chen. 2017. "Magnet: A Two-Pronged Defense Against Adversarial Examples." https://arxiv.org/abs/1705.09064.

Mitchell, T.M. 1997. Does machine learning really work? *AI Magazine* 18(3):11-20.

Moran, J., and R. Desimone. 1985. Selective attention gates visual processing in the extrastriate cortex. *Science* 229:782-784.

NRC (National Research Council). 2015. *Robust Methods for the Analysis of Images and Videos for Fisheries Stock Assessment: Summary of a Workshop*. Washington, DC: The National Academies Press.

Phillips, P.J., A.N. Yates, Y. Hu, C.A. Hahn, E. Noyesm K. Jackson, J.G. Cavazoa, et al. 2018. Face recognition accuracy of forensic examiners, superrecognizers and face recognition algorithms. *Proceedings of the National Academy of Sciences* 115(24):6171-6176.

Potter, M.C. 1975. Meaning in visual search. *Science* 187(4180):965-966.

Roberts, L.G. 1963. "Machine Perception of Three-Dimensional Solids." Doctoral dissertation. Cambridge, MA: Massachusetts Institute of Technology.

Rosenfeld, A., and M. Thurston. 1971. Edge and curve detection for visual scene analysis. *IEEE Transactions on Computers* (5):562-569.

Rudd, E., L.P. Jain, W.J. Scheirer, and T.E. Boult. 2018. The Extreme Value Machine. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 40(3):762-768.

Scheirer, W.J., A. Rocha, A. Sapkota, and T.E. Boult. 2013. Toward open-set recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 35(7):1757-1772.

Thorpe, S., D. Fize, and C. Marlot. 1996. Speed of processing in the human visual system. *Nature* 381(6582):520-522.

Tsotsos, J.K. 1987. "A 'Complexity Level' analysis of vision." Pp. 346-355 in *Proceedings of the 1st International Conference on Computer Vision* (M. Brady and A. Rosenfeld, eds.). Washington DC: IEEE Computer Society Press.

Tsotsos, J., I. Kotseruba, and C. Wloka. 2016. A focus on selection for fixation. *Journal of Eye Movement Research* 9(5).

Tsotsos, J.K., A. Rodriguez-Sanchez, A. Rothenstein, and E. Simine. 2008. Different binding strategies for the different stages of visual recognition. *Brain Research* 1225:119-132.

Uhr, L. 1972. Layered "recognition cone" networks that preprocess, classify and describe. *IEEE Transactions on Compute*rs C-21(7):758-768.

Valiant, L. 1984. A theory of the learnable. *Communications of the ACM* 27(11):1134-1142.

von der Malsburg, C. 1981. *The Correlation Theory of Brain Function*. Internal Report 81-82. Göttingen, Germany: Department of Neurobiology, MaxPlanck Institute for Biophysical Chemistry.

Xiao, C., R. Deng, B. Li, F. Yu, M. Liu, and D.X. Song. 2018. "Characterizing Adversarial Examples Based on Spatial Consistency Information for Semantic Segmentation." https://arxiv.org/abs/1810.05162.

Appendixes

A

Biographical Sketches of Workshop Planning Committee

RAMA CHELLAPPA, *Chair*, is a distinguished university professor and professor of electrical and computer engineering and an affiliate professor of computer science with the University of Maryland, College Park. He received a B.E. (honors) degree from the University of Madras, Madras, India, in 1975, an M.E. (Distinction) degree from the Indian Institute of Science, Bangalore, India, in 1977, and M.S.E.E. and Ph.D. degrees in electrical engineering from Purdue University in 1978 and 1981, respectively. He is also affiliated with the Center for Automation Research and the Institute for Advanced Computer Studies (permanent member). In 2005, he was named a Minta Martin Professor of Engineering. Prior to joining the University of Maryland, he was an assistant (1981–1986) and associate professor (1986–1991) and director of the Signal and Image Processing Institute (1988–1990) with the University of Southern California. Over the past 37 years, he has published numerous book chapters and peer-reviewed journal and conference papers. He has coauthored and coedited books on Markov random fields and face and gait recognition as well as collected works on image processing and analysis. He has served as a co-editor-in-chief of *Graphical Models and Image Processing*. His current research interests are machine intelligence; face and gait analysis; markerless motion capture; 3D modeling from video, image, and video-based recognition and exploitation; compressive sensing; and hyperspectral processing. Professor Chellappa has received several awards, including a National Science Foundation (NSF) Presidential Young Investigator Award, four IBM Faculty Development Awards, an Excellence in Teaching Award from the School of Engineering at University of Southern California, and two paper awards from the International Association of Pattern Recognition. He received the K.S. Fu Prize from the International Association of Pattern Recognition and the Inaugural Leadership Award from the Institute of Electrical and Electronics Engineers (IEEE) Biometrics Council. He received the Society, Technical Achievement, and Meritorious Service Awards from the IEEE Signal Processing Society. He also received the Technical Achievement and Meritorious Service Awards from the IEEE Computer Society. At the University of Maryland, he was elected as a Distinguished Faculty Research Fellow and as a Distinguished Scholar-Teacher, and he received the Outstanding Faculty Research Award and the Poole and Kent Teaching Award for the Senior Faculty from the College of Engineering, an Outstanding Innovator Award from the Office of Technology Commercialization, and an Outstanding GEMSTONE Mentor Award. In 2010, he was recognized as an Outstanding ECE by Purdue University. In 2016, he was also recognized as a Distinguished Alumni by the

Indian Institute of Science. He is a fellow of the IEEE, the International Association for Pattern Recognition, the Optical Society of America, the Association for the Advancement of Artificial Intelligence, the Association for Computing Machinery (ACM), and the American Association for the Advancement of Science. He has served as an associate editor for four IEEE publications and as the editor-in-chief of the *IEEE Transactions on Pattern Analysis and Machine Intelligence*. He served as a member of the IEEE Signal Processing Society Board of Governors and as its vice president of awards and membership. He has served as a general and technical program chair for several IEEE international and national conferences and workshops. He is a golden core member of the IEEE Computer Society and served a 2-year term as a distinguished lecturer of the IEEE Signal Processing Society. Recently, he completed a 2-year term as the president of IEEE Biometrics Council.

TODD BORKEY joined Alion Science and Technology as chief technology officer (CTO) in 2017. As CTO, Mr. Borkey manages the company's technology strategy, along with its development and operations. He is chartered to expand Alion's business by developing new technologies and solutions that meet the changing needs of our clients. Mr. Borkey came to Alion from Thales Defense and Security where he served for 5 years as both corporate CTO and vice president of the Thales System Solutions business unit. At Thales, Mr. Borkey was responsible for the roadmap and business operations to a wide range of products, which included radio frequency communications, C4ISR solutions, radars, sonars, and cyber/EW products. Prior to Thales, Mr. Borkey served for 6 years as CTO of DRS Defense Solutions, a $1.3 billion C4ISR defense enterprise. Earlier in his career, he performed a range of engineering and management assignments within Northrop Grumman and AT&T Bell Labs. As the senior technology leader at Alion, Mr. Borkey brings profit and loss (P&L) experience, program management, and business development experience into the role. He has a master of science in engineering management from Stevens Institute of Technology and holds an undergraduate degree in applied mathematics.

JULIE BRILL is the corporate vice president and deputy general counsel for Global Privacy and Regulatory Affairs at Microsoft Corporation. She is a former partner at Hogan Lovells. She was a commissioner of the Federal Trade Commission (FTC) from April 2010 to early 2016, where she worked actively on issues of critical importance to today's consumers, including protecting consumers' privacy, encouraging appropriate advertising substantiation, guarding consumers from financial fraud, and maintaining competition in industries involving health care and high-tech. Ms. Brill was named "the commission's most important voice on Internet privacy and data security issues," a "key player in U.S. and global regulations," "one of the top minds in online privacy," one of the top four U.S. government players "leading the data privacy debate," "one of the top 50 influencers on big data," a "game-changer," and "one of the 50 most powerful people in health care." She also focused on the need to improve consumer protection in the financial services arena. Ms. Brill has received numerous national awards for her work. In addition to the International Association of Privacy Professionals 2014 Privacy Professionals Privacy Leader of the Year Award, she also received the New York University School of Law Alumna of the Year Award, was named one of eight "Government Stars" among the "2015 Attorneys Who Matter," and was recently elected to the American Law Institute. Prior to becoming a commissioner, Ms. Brill was the senior deputy attorney general and chief of consumer protection and antitrust for the North Carolina Department of Justice. She also served as an assistant attorney general for consumer protection and antitrust for the State of Vermont for more than 20 years and was an associate at Paul, Weiss, Rifkind, Wharton and Garrison in New York. She clerked for Vermont Federal District Court Judge Franklin S. Billings, Jr. Ms. Brill graduated, magna cum laude, from Princeton University, and from New York University School of Law, where she had a Root-Tilden Scholarship for her commitment to public service.

LISE GETOOR is a professor in the Computer Science Department at the University of California, Santa Cruz, and the director of the University of California, Santa Cruz, D3 Data Science Center. Her research areas include machine learning and reasoning under uncertainty; in addition, she works in data management, visual analytics, and social network analysis. She has over 200 publications, is a fellow of the

Association for Artificial Intelligence, and is an elected board member of the International Machine Learning Society. She serves on the board of the Computing Research Association; has served as *Machine Learning Journal* action editor and associate editor for the *ACM Transactions of Knowledge Discovery from Data* and *Journal of Artificial Intelligence Research*, and has served on the Association for the Advancement of Artificial Intelligence Council. She is a recipient of an NSF Career Award and 11 best paper and best student paper awards. In 2014, she was recognized as one of the top 10 emerging researchers in data mining and data science based on citation and impact according to KD Nuggets. She is on the external advisory board of the San Diego Super Computer Center and the scientific advisory board for the Max Planck Institute for Software Systems, and has served on the advisory board for companies including Sentient Technologies. She received her Ph.D. from Stanford University in 2001, her M.S. from the University of California, Berkeley, and her B.S. from the University of California, Santa Barbara, and was a professor at the University of Maryland, College Park, from 2001-2013.

ANTHONY HOOGS is the senior director of computer vision at Kitware, Inc., a small scientific computing research and development firm based on open-source software. Dr. Hoogs leads Kitware's computer vision group, which he started in 2007 and now has 40 members. For more than two decades, he has supervised and performed research in various areas of computer vision including event, activity, and behavior recognition; motion pattern learning and anomaly detection; tracking; remote sensing imagery analytics; image segmentation; object recognition; and content-based retrieval. At Kitware, he has initiated and led more than two dozen projects in image and video analysis, involving more than 20 universities and sponsored by government institutions including the Defense Advanced Research Projects Agency (DARPA), the Air Force Research Laboratory, the Intelligence Advanced Research Projects Activity, the Office of Naval Research, the Office of the Secretary of Defense, the National Geospatial-Intelligence Agency, and others, ranging from basic academic research to developing advanced prototypes and demonstrations installed at operational facilities. Previously at GE Global Research (1998-2007), Dr. Hoogs led a team of vision researchers on projects sponsored by the U.S. government, Lockheed Martin, and NBC Universal. He has published more than 80 papers in computer vision and has served as general co-chair for the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2017; general co-chair for the IEEE Winter conference on Applications of Computer Vision (WACV) 2016 and 2018; area chair for CVPR (2009, 2010, 2012, 2018); workshops co-chair for CVPR (2012); corporate relations chair for CVPR (2009, 2010) and the International Conference on Computer Vision (2013); program co-chair for WACV (2009, 2011); organizer for various CVPR and ICCV workshops; member of the Computer Vision Foundation Advisory Board and Industrial Advisory Board; member of the steering committee for WACV. He regularly serves as a reviewer for major computer vision conferences and journals including CVPR, ICCV, the European Conference on Computer Vision, and WACV. He has served on technical panels for NSF and DARPA, including DARPA Information Science and Technology (ISAT) panels in 2007, 2009, and 2013 and has started a 3-year term as an ISAT member in 2017. In 2014, he served as an organizer of the National Academies of Sciences, Engineering, and Medicine's Workshop on Robust Methods for the Analysis of Images and Videos for Fisheries Stock Assessment, sponsored by the National Oceanic and Atmospheric Administration (NOAA), then joined the Steering Committee for the NOAA National Marine Fisheries Service Automated Imagery Analysis Strategic Initiative. In 2017, he served as an organizer of the National Academies' In-FoRM Machine Analytics Workshop, sponsored by the Intelligence Community to examine challenges and emerging opportunities in machine learning for intelligence analytics. Dr. Hoogs received a Ph.D. in computer and information science from the University of Pennsylvania in 1998; an M.S. from the University of Illinois, Urbana-Champaign, in 1991; and a B.A., magna cum laude, from Amherst College in 1989.

ANITA JONES is a university professor emerita in the School of Engineering and Applied Science at the University of Virginia. The Honorable Anita K. Jones served as director of defense research and engineering for the U.S. Department of Defense (DoD) from 1993 to 1997, overseeing the department's science and technology program. She has served as vice chair of the National Science Board and a member

of the Massachusetts Institute of Technology (MIT) Corporation Executive Committee. She is currently a senior fellow of the Defense Science Board, a trustee of InQTel, and a member of provost's advisory board for MIT Lincoln Laboratories. She was awarded the Founders' Medal by the IEEE, the Ada Lovelace Award by the Association of Women in Computing, the Arthur M. Bueche Award by the National Academy of Engineering, the Philip Abelson Award by the American Association for the Advancement of Science, and the award for Distinguished Public Service by the DoD. The U.S. Navy named a seamount in the North Pacific Ocean for her. Duke University, Carnegie Mellon University, and the University of Southern California have all awarded her honorary doctorates. She has published more than 50 technical articles and two books on computer software, cybersecurity, and science and technology policy. She is a member of the National Academy of Engineering. Dr. Jones holds an A.B. in mathematics from Rice University, an M.A. in literature from the University of Texas, Austin, and a Ph.D. in computer science from Carnegie Mellon University.

YUNYAO LI is a principal research staff member and a senior research manager member with IBM Research–Almaden, where she manages the Scalable Knowledge Intelligence group. She is a member of the New Voices program of the National Academies. She is also a master inventor and a member of IBM Academy of Technology. Her expertise is in the interdisciplinary areas of natural language processing, databases, human–computer interaction, and information retrieval. She is a widely recognized expert in these areas both within IBM and in the external research community, with more than 50 research publications and more than 20 patents granted/filed in these areas. She is a founding member of SystemT, a state-of-the-art information extraction engine currently powering 10+ IBM products, and numerous research projects and customer engagements. She is also a founding member of Gumshoe, a novel enterprise search engine that has been powering the IBM intranet and ibm.com search from 2010 to 2017. Her contribution to these projects has been recognized by multiple prestigious IBM internal awards. She received her Ph.D. degree and master's degrees from the University of Michigan, Ann Arbor, and undergraduate degrees from Tsinghua University, Beijing, China. Dr. Li is passionate about improving the diversity for the science, technology, engineering, and medicine field. She has been actively mentoring women and underrepresented minorities for more than 10 years. She has been serving on the MentorNet Mentor-Protégé Council since 2013 and the external advisory board for the Computer Science Department of the San Jose State University since early 2016. She also has cofounded and currently leads Almaden Women's Interest Network Group, aiming to provide a networking forum for technical women in IBM Almaden Research Center, advance women in technology, and enhance the diverse workforce.

JOYSULA RAO is an IBM fellow and the director of security research at IBM. Based in IBM's Thomas J. Watson Research Center, the global team comprises more than 200 researchers who work in the areas of cryptography, cybersecurity, cloud, mobile security, and secure platform technologies. Dr. Rao works closely with customers, academic partners, and IBM business units to drive new and innovative technologies into IBM's products and services and definitive industry standards. The goal of his research is to raise the bar significantly on the quality of security while simultaneously easing the overhead of developing and deploying secure solutions. Dr. Rao has published widely in premier security conferences and workshops. He holds numerous U.S. and European patents. He is a member of the prestigious IBM Academy of Technology and an emeritus member of the International Federation for Information Processing Working Group 2.3 (Programming Methodology) and the Industry Advisory Board of the Georgia Tech Information Security Center. Dr. Rao obtained his doctorate degree from the University of Texas, Austin, an M.S. in computer science from the State University of New York at Stony Brook, and an A.B.Tech. in electrical engineering from the Indian Institute of Technology, Kanpur.

SAMUEL VISNER is a senior executive at the MITRE Corporation, an adjunct professor of cybersecurity at Georgetown University, and the director of the National Cybersecurity Federally Funded Research and Development Center (NCF). MITRE manages the NCF in support of the National Institute of Standards and Technology's National Cybersecurity Center of Excellence). In this role, he oversees efforts to bring

together experts from industry, government, and academia to demonstrate integrated cybersecurity solutions that are cost-effective, repeatable, and scalable. The NCF is the first of its kind dedicated to cybersecurity. Mr. Visner joined MITRE having served as senior vice president and general manager for cybersecurity and resilience at ICF, where he managed the company's business unit (P&L) supporting a wide range of national and homeland security clients and programs, including cybersecurity research and development. He also held leadership positions at Computer Sciences Corporation and Science Applications International Corporation. In addition, he served as chief of signals intelligence programs at the National Security Agency, where he was awarded the agency's Exceptional Civilian Service Award in 2003. Mr. Visner has been a leader in public–private partnerships and collaborations, including the Intelligence and National Security Alliance, the Air Force Communications and Electronics Association, the Professional Services Council, and the National Academy of Sciences. Throughout his career, he has worked across multiple federal sponsors. Mr. Visner also serves as member of the Cyber and Domestic Security Councils of the Intelligence and National Security Alliance. As an adjunct professor of science and technology in international affairs at Georgetown University, Mr. Visner teaches a course on cybersecurity policy, operations, and technology. He is also a member of the Council on Foreign Relations and an Intelligence Associate of the National Intelligence Council and is a member of the Intelligence Science and Technology Experts Group, sponsored by the National Academy of Sciences and serving the Office of the Director of National Intelligence. Mr. Visner also served as a member of the board of directors of CVG/Avtec. Mr. Visner holds a bachelor's degree in international politics from Georgetown University and a M.S. in telecommunications from George Washington University. Mr. Visner has served twice on the Intelligence, Surveillance, and Reconnaissance Task Force of the Defense Science Board and has published articles on national and cybersecurity in *World Politics Review*, the *Georgetown Journal of International Affairs*, and the *Defense Intelligence Journal*.

B

Workshop Agenda

DAY 1: DECEMBER 11, 2018

Session 1: Plenary

8:00 AM	Sponsor Remarks and Expectations of the Workshop *David M. Isaacson, Office of the Director of National Intelligence*
8:15	Generation of Capability Technology Matrix Tables *Rama Chellappa, Planning Committee Chair, University of Maryland, College Park* *George Coyle, Study Director, Intelligence Community Studies Board, National Academies of Sciences, Engineering, and Medicine*
8:30	Recent Advances in Machine Learning *Michael Jordan, University of California, Berkeley*
9:30	Machine Learning on Perception: Hype vs. Hope *Ruzena Bajcsy, University of California, Berkeley*
10:30	Break

Session 2: Adversarial Attacks

11:00	Media Forensics *Matthew Turek, Defense Advanced Research Projects Agency*
11:45	Forensic Techniques *Hany Farid, Dartmouth College*

12:30 PM Lunch

Session 3: Detection and Mitigation of Adversarial Attacks and Anomalies

1:30 *Joysula Rao, IBM Corporation*

2:15 Circumventing Defenses to Adversarial Examples
Anish Athalye, Massachusetts Institute of Technology

3:00 Break

Session 4: Enablers of Machine Learning Algorithms and Systems

3:30 Impact of Neuroscience on Data Science for Perception
John Tsotsos, York University, Canada

5:30 Capability Technology Matrix Tables Preparation

6:00 Adjourn for the Day

DAY 2: DECEMBER 12, 2018

8:00 AM Sponsor Remarks
David M. Isaacson, Office of the Director of National Intelligence

Session 5: Recent Trends in Machine Learning—1

8:15 On Open Set and Adversarial Issues in Machine Learning
Terry Boult, University of Colorado, Colorado Springs

9:00 GANs for Domain Adaptation and Security Against Attacks
Rama Chellappa, University of Maryland, College Park

9:45 Break

Session 6: Recent Trends in Machine Learning—2

10:00 Recent Advances in Optimization for Machine Learning
Tom Goldstein, University of Maryland

10:45 Forecasting Using Machine Learning
Aram Galstyan, Information Sciences Institute, University of Southern California

Session 7: Plenary Session

11:30	Plenary Talk *Dawn Song, University of California, Berkeley*
12:30 PM	Lunch

Session 8: Recent Trends in Machine Learning—3

1:30	Domain Adaptation *Judy Hoffman, Georgia Institute of Technology*
2:15	Explainable Machine Learning *Anna Rohrbach, University of California, Berkeley*
3:00	Break

Session 9: Machine Learning System

3:15	Building Domain-Specific Knowledge with Human-in-the-Loop *Yunyao Li, IBM Corporation*
4:00	Robust Design of Machine Learning Systems *Anthony Hoogs, Kitware, Inc.*

Session 10: Capability Technology Matrix Tables

4:45	Discussion on Preparing the Capability Technology Matrix Tables
5:30	Adjourn Workshop

C

Workshop Statement of Task

To address quality concerns raised during the 2017 ODNI workshop on "Challenges in Machine Generation of Analytic Products from Multi-Source Data," a 2-day National Academies' workshop will explore methods for assessing the accuracy and veracity of machine-generated analytic intelligence products and techniques for addressing the potential impacts of adversarial manipulation of analytical inputs.

A planning committee will organize a workshop to discuss:

1. The current state of machine-driven approaches such as machine learning and natural language processing that can be used to generate and evaluate analytic products from disparate structured and unstructured data types and to detect anomalies;
2. Approaches for ensuring that machine-generated products compare favorably with those of trained human analysts;
3. Statistical methods that can be used to establish confidence hierarchies, model uncertainty, and error propagation, and manage risk as a function of time and complexity and;
4. Techniques for responding to adversarial manipulation of input data to influence analytical products by exploiting weaknesses in machine learning and other AI algorithms and vulnerabilities in their implementation.

A rapporteur-authored workshop proceedings will be prepared.

The planning committee will consider the following research questions:

- What are the technical objectives and metrics needed for success?
- What are the primary issues?
- What are the current and "next level" key performance metrics? What is the "level after next" of expected research and development performance?
- What is the research knowledge base?
- How can the government best prepare the scientific workforce to enhance discovery in this area?

D

Capability Technology Matrix

Workshop participants identified both near- and long-term enabling technology capabilities to help guide the Intelligence Community's future technology investments (Table D.1).

TABLE D.1 Short- and Long-term Technology Capabilities Described by Workshop Speakers

Participant	Short-Term Capabilities (3-5 years)	Long-Term Capabilities (5-10 years)
Matthew Turek, Defense Advanced Research Projects Agency	• Formalized requirements to test and assess models • Verification of whether existing models can run using new requirements • Hybridized models that incorporate contextual information in addition to data points	
Hany Farid, Dartmouth College	• Creation of more automated processes and faster work, owing to the sheer volume of data that is available • Improved accuracy • Use of secure-imaging pipelines to prevent the manipulation of digital evidence • Better cooperation from social media to reduce deepfakes • Increased education for citizen awareness to better recognize "fakes" in the digital age • Guidelines for responsible deployment of technologies • Consideration of the ethical and societal implications of algorithms that are advertised as artificial intelligence (AI), but are actually linear regression models, particularly in the	

Participant	Short-Term Capabilities (3-5 years)	Long-Term Capabilities (5-10 years)
	predictive space, (e.g., popular algorithms used to make decisions on university admissions and employment) • The weaponization, use and ethical implications of black-box AI algorithms	
Joysula Rao, IBM Corporation	• AI techniques used for defense • AI-powered attacks	• Approaches and methodologies for developing provable security
Anish Athalye, Massachusetts Institute of Technology	• The increase of the use of AI /machine learning to attack real systems • More realistic attacks by malicious actors • More principled evaluations of defenses • Provable/certifiable defenses	• Increased security of machine learning systems with answers to the following questions: — What is an adversarial example? — What is the specification of a machine learning system? • How should a machine behave given a particular input?
Terry Boult, University of Colorado, Colorado Springs	• Open-set recognition algorithms for well-behaved, low-moderate dimensional feature spaces • Realistic large open-set data sets/protocols • Better understanding of image–feature relationships • Ability of iterative Layer-wise Origin-Target Synthesis (LOTS) to attack all kinds of systems • LOTS attacks are reasonably portable • Use of LOTS to build physical attacks/camouflage • Problems with good representations to relate images to features	• Better network models for open-set deep recognition • High-dimensional open-set algorithms
Rama Chellappa, University of Maryland, College Park	• Explore the robustness of deeper networks • Work with multimodal inputs • Increase theoretical analysis • Investigate how humans and machines can work together to thwart adversarial attacks • Demonstrate on more difficult computer vision problems (e.g., face verification/identification, action detection, detection of doctored media)	• Keep changing the network configuration and parameters in a probabilistic manner with guaranteed performance (i.e., adaptive networks) • Humans and machines work together • Design networks that incorporate common sense reasoning
Aram Galstyan, Information Sciences Institute, University of Southern California		• Hybrid sensemaking systems
Judy Hoffman, Georgia Institute of Technology	• Develop effective uncertainty measures and confidence intervals	

Participant	Short-Term Capabilities (3-5 years)	Long-Term Capabilities (5-10 years)
	• Improve auto-calibration and understanding of when things have changed • Rethink the notion of domains in adaptation literature • Address the idea of overall robustness in the adaptation literature—find a way to improve control over the initial model to reduce susceptibility to natural or artificial changes	
Anthony Hoogs, Kitware, Inc.	• Generative adversarial networks effectively applied to video • Cataloguing of large, common objects and events • Large graphical processing unit farms required to keep up with video generation • Structure learning of deep networks on a large scale • Increased model transfer between video domains • Human-level accuracy for action recognition in single-action, temporarily clipped videos • Free-form, text-based queries with limited and open syntax and vocabulary • Adversarial attacks on video recognition problems (e.g., action recognition) • Cataloguing of all objects, scene elements, and events	• Human-level accuracy for action and complex activity recognition in surveillance video • Similar accuracy to humans but much faster for video search and retrieval for complex activities and abstractions in Internet videos
Nathalie Baracaldo, IBM Corporation	• Extraction attacks in learning models	
Yunyao Li, IBM Corporation	• Declarative systems that enable the building of more complex and interpretable models at scale • SystemT: a declarative text understanding system for the enterprise • SystemER: a declarative entity understanding and resolution system for the enterprise • Human-in-the-loop technologies that allow human and machines to co-create artificial intelligence algorithms/models • Emerging technologies to enable cross-lingual universal semantic understanding of natural language • Domain-specific knowledge base construction • Data need to be collected so they follow some procedure that the institutes, including commercial institutes, can use to develop the technology without issue (i.e., Institutional Review Board-type approval for data connections)	• Automatically build robust explainable machine learning models easily adaptable across languages and domains requiring minimal labeled data and human input

E

Acronyms

2D	two dimensional
3D	three dimensional
ACE	Aggregative Contingent Estimation
AI	artificial intelligence
ARIMA	AutoRegressive Integrated Moving Average
CAMEO	conflict and mediation event observation
CorEx	Total Correlation Explanation
DoD	Department of Defense
EVM	extreme value machine
GAN	generative adversarial network
GPU	graphical processing unit
IARPA	Intelligence Advanced Research Projects Activity
IC	Intelligence Community
ICSB	Intelligence Community Studies Board
IFP	individual forecasting problem
KWIVER	Kitware Imagery and Video Exploitation and Retrieval
LOTS	layerwise origin-target synthesis
NOAA	National Oceanic and Atmospheric Administration
ODNI	Office of the Director of National Intelligence
ROC	receiver operating characteristic
SAGE	Synergistic Anticipation of Geopolitical Events
VIAME	Video and Imagery Analytics for the Marine Environment
VQA	visual question answering
XAI	explainable artificial intelligence