The science of deep learning

Richard Baraniuk[a], David Donoho[b,1], and Matan Gavish[c]

artificial intelligence | machine learning | deep learning | neural networks

Scientists today have completely different ideas of what machines can learn to do than we had only 10 y ago.

In image processing, speech and video processing, machine vision, natural language processing, and classic two-player games, in particular, the state-of-the-art has been rapidly pushed forward over the last decade, as a series of machine-learning performance records were achieved for publicly organized challenge problems. In many of these challenges, the records now meet or exceed human performance level.

A contest in 2010 proved that the Go-playing computer software of the day could not beat a strong human Go player. Today, in 2020, no one believes that human Go players—including human world champion Lee Sedol—can beat AlphaGo, a system constructed over the last decade. These new performance records, and the way they were achieved, obliterate the expectations of 10 y ago. At that time, human-level performance seemed a long way off and, for many, it seemed that no technologies then available would be able to deliver such performance.

Systems like AlphaGo benefited in this last decade from a completely unanticipated simultaneous expansion on several fronts. On the one hand, we saw the unprecedented availability of on-demand scalable computing power in the form of cloud computing, and on the other hand, a massive industrial investment in assembling human engineering teams from a globalized talent pool by some of the largest global technology players. These resources were steadily deployed over that decade to allow rapid expansions in challenge problem performance.

The 2010s produced a true technology explosion, a one-time–only transition: The sudden public availability of massive image and text data. Billions of people posted trillions of images and documents on social media, as the phrase "Big Data" entered media awareness. Image processing and natural language processing were forever changed by this new data resource as they tapped the new image and text resources using the revolutionary increases in computing power and newly globalized human talent pool.

The field of image processing felt the impact of the new data first, as the ImageNet dataset scraped from the web by Fei-Fei Li and her collaborators powered a series of annual ImageNet Large-Scale Visual Recognition Challenge (ILSVRC) prediction challenge competitions. These competitions gave a platform for the emergence and successive refinement of today's deep-learning paradigm in machine learning.

Deep neural networks had been developing steadily since at least the 1980s; however, their heuristic construction by trial-and-error resisted attempts at analysis. For quite some time in the 1990s and 2000s, artificial neural networks were regarded with suspicion by scientists insisting on formal theoretical justification. In this decade they began to dominate prediction challenges like ImageNet. The explosion of image data on the internet and computing resources from the cloud enabled new, highly ambitious deep network models to win prediction challenges by substantial margins over more "formally analyzable" methods, such as kernel methods.

In fact, the performance advantage of deep networks over more "theoretically understandable" methods accelerated as the decade continued. The initial successes famously involved separating pictures of cats from dogs, but soon enough successes came in full-blown computer vision problems, like face recognition and tracking pedestrians in moving images.

[a]Department of Electrical and Computer Engineering, Rice University, Houston, TX 77005; [b]Department of Statistics, Stanford University, Stanford, CA 94305; and [c]School of Computer Science and Engineering, Hebrew University of Jerusalem, 919051 Jerusalem, Israel

This paper results from the Arthur M. Sackler Colloquium of the National Academy of Sciences, "The Science of Deep Learning," held March 13–14, 2019, at the National Academy of Sciences in Washington, DC. NAS colloquia began in 1991 and have been published in PNAS since 1995. From February 2001 through May 2019 colloquia were supported by a generous gift from The Dame Jillian and Dr. Arthur M. Sackler Foundation for the Arts, Sciences, & Humanities, in memory of Dame Sackler's husband, Arthur M. Sackler. The complete program and video recordings of most presentations are available on the NAS website at http://www.nasonline.org/science-of-deep-learning.

Author contributions: R.B., D.D., and M.G. wrote the paper.

The authors declare no competing interest.

[1]To whom correspondence may be addressed. Email: donoho@stanford.edu.

First published November 23, 2020.

A few years following their initial successes in image processing, deep networks began to penetrate natural language processing, eventually producing in the hands of the largest industrial research teams systems able to translate any of 105 languages to any other, even language pairs for which almost no prior translation examples were available.

Today, it is no longer shocking to hear of deep networks with tens of billions of parameters trained using databases with tens of billions of examples. On the other hand, it may have been increasingly unsettling for scientists to witness human performance being dominated by empirically derived systems whose best-understood properties were simply their ability to prevail in gameplay and in prediction challenges like ImageNet.

In March of 2019, the National Academy of Sciences convened a Sackler Colloquium on "The Science of Deep Learning" in the Academy building in Washington, DC. The goal of the organizers was to advance scientific understanding of today's empirically derived deep-learning systems, and at the same time to advance the use of such systems for traditional scientific research.

To those ends, important figures from academia and industry made presentations over 2 d; audience members—who included many graduate students and postdocs from institutions nationwide, as well as research sponsors from the NSF, NIH, and Department of Defense (DoD), and US government scientists from Washington, DC-area laboratories—found much to discuss in the hallways between presentations. Two presentations that were strikingly successful with the audience included Amnon Shashua and Rodney Brooks.

Amnon Shashua, from Hebrew University and Intel Mobility systems, discussed computer vision research strategies to enable self-driving cars. He told the audience that error rates of vision systems for moving vehicles need to stay below one missed detection per trillion units of visual experience, and discussed modeling and test strategies that can someday produce verified systems with such low error rates.

Rodney Brooks of Massachusetts Institute of Technology (MIT) explained how, in his view, it would be hundreds of years before machine-learning systems exhibit fully general intelligence. In support, he pointed to the currently prodigious appetite of today's successful deep-learning systems for massive volumes of good data, and contrasted this with humans' ability to understand and generalize from very little data.

In the weeks prior to the colloquium, the White House released a national strategy document titled "Artificial Intelligence for the American People" (1), which called for new United States investment in artificial intelligence (AI). Because the colloquium took place in the Academy's building on the Washington Mall, the colloquium overnight became a perfect venue to discuss the new initiative. Representatives of funding agencies (NSF, NIH, and DoD), including some who were deeply involved in formulating the strategy, described their recent and coming research portfolios, and told the audience how deep-learning research fit into coming national research initiatives.

As part of the Sackler colloquia series, the event is accompanied by a special issue of PNAS, the one you are now reading, authored by some of the speakers and participants of the colloquium. The many interesting papers gathered together in this volume reflect the vitality and depth of the scientific work being carried out in this new and rapidly developing field.

The special issue begins with two general overview papers. Terrence J. Sejnowski of the Salk Institute discusses "The unreasonable effectiveness of deep learning in artificial intelligence" (2). Sejnowski's title stands in a tradition of similar titles that starts from Eugene Wigner's famous essay on "The unreasonable effectiveness of mathematics in the physical sciences" (3) and continued in this decade with "The unreasonable effectiveness of data" (4) by Alon Halevy, Peter Norvig, and Fernando Pereira of Google. In this tradition, authors generally point to a technology (e.g., Mathematics, Big Data, Deep Learning) that enjoys undoubted success in certain endeavors, but which we don't completely understand, and which, from a higher-level perspective, might seem surprising. Sejnowski (2) examines the paradox that, for a range of important machine-learning problems, deep learning works far better than conventional statistical learning theories would predict. Sejnowski suggests that, while today's deep-learning systems have been inspired by the cerebral cortex of the brain, reaching artificial general intelligence will require inspiration from other important brain regions, such as those responsible for planning and survival.

Tomaso Poggio, Andrzej Banburski, and Qianli Liao of MIT follow up nicely with "Theoretical issues in deep networks" (5), which considers recent theoretical results on approximation power, complexity control, and generalization properties of deep neural networks. Empirically, deep neural networks behave very differently under these three aspects from other machine-learning models. For approximation, the authors state formal results proving that certain convolutional nets can avoid the "curse of dimensionality" when approximating certain smooth functions. For complexity control and regularization, the authors consider the gradient flow of appropriately normalized networks under the exponential loss as a dynamical system. The authors point to implicit regularization properties of unconstrained gradient descent, to possibly explain the complexity control observed in overparameterized deep nets.

The idea that "deep learning keeps surprising us" was further developed by Christopher D. Manning, Kevin Clark, John Hewitt, Urvashi Khandelwal, and Omer Levy of Stanford University (6). They consider deep neural nets, trained via self-supervision, which predict a masked word in a given context without labeled training data. The authors challenge the dominant perspective in linguistics, which posits that statistical machine-learning predictive language models do not develop interesting emergent knowledge of linguistic structures. Striking empirical evidence is presented for syntactic, morphological, and semantic linguistic structures that emerge in deep neural networks during their self-supervised training. That such rich information emerges through self-supervision has tantalizing implications for human language acquisition.

Kyle Cranmer of New York University, with coauthors Johann Brehmer and Gilles Louppe, discuss another emergent surprise in their article "The frontier of simulation-based inference" (7). The article describes important scientific inference problems in particle physics that were until now viewed as intractable. Pointing to today's "Machine Learning Revolution," the author's identify new possibilities for attacking such inference problems, by fusing massive scientific simulations, machine-learning ideas such as active learning, and probabilistic modeling. In effect, machine learning can help us by training on measurements from scientific simulations to give us empirical models in place of classic analytical probabilistic models, which often are unavailable. They point to a range of scientific inference problems and conclude with these words: "... several domains of science should expect ... a significant improvement in inference quality ... this transition may have a profound impact on science" (7).

Our special issue also provides engaging articles on specific research questions. Peter L. Bartlett of University of California,

Berkeley and coauthors Philip M. Long, Gábor Lugosi, and Alexander Tsigler discuss "Benign overfitting in linear regression" (8). Many recent deep-learning models contain more parameters to be determined than there are data points to fit them. We say such models are overfit. Traditionally, this would have been considered inimical to good empirical science. As the authors say: "The phenomenon of benign overfitting is one of the key mysteries uncovered by deep learning methodology: Deep neural networks seem to predict well, even with a perfect fit to noisy training data" (8). The authors conduct a penetrating formal analysis of the situation in the simplified setting of linear regression.

Antonio Torralba of MIT, with coauthors David Bau, Jun-Yan Zhu, Hendrik Strobelt, Agata Lapedriza, and Bolei Zhou address an important concern: Deep neural networks contain billions of artificial neurons, but what are they doing? Their article "Understanding the role of individual units in a deep neural network" (9) starts as follows: "Can the individual hidden units of a deep network teach us how the network solves a complex task? Intriguingly, within state-of-the-art deep networks, it has been observed that many single units match human-interpretable concepts that were not explicitly taught to the network: Units have been found to detect objects, parts, textures, tense, gender, context, and sentiment." The authors describe quantitative tools to make such identifications. Building a second "annotation network," they develop a "dissection" framework identifying the concepts that drive the networks' neurons to respond. The technique applies to image-classification and image-generation networks and provides new insights into adversarial attacks and semantic image editing.

Doina Precup, of McGill University and DeepMind, and her coauthors André Barreto, Shaobo Hou, Diana Borsa, and David Silver discuss reinforcement learning, the variety of machine learning that gave us AlphaGo's world-beating gameplay systems. Reinforcement learning is famously data-hungry. Precup and colleagues suggest a way out. Their article "Fast reinforcement learning with generalized policy updates" (10) begins with: "The combination of reinforcement learning with deep learning is a promising approach to tackle important sequential decision-making problems that are currently intractable." To surmount the obstacles to such a combination with deep learning, the authors (10) propose "... generalization of two fundamental operations in reinforcement learning: Policy improvement and policy evaluation. The generalized version of these operations allow one to leverage the solution of some tasks to speed up the solution of others." Barreto et al. (10) find that "Both strategies considerably reduce the amount of data needed to solve a reinforcement-learning problem."

The special issue ends with two articles addressing emergent concerns about effects of machine learning on daily life. Anders C. Hansen of Cambridge University and coauthors Vegard Antun, Francesco Renna, Clarice Poon, and Ben Adcock identify a looming technical threat. Their article "On instabilities of deep learning in image reconstruction and the potential costs of AI" (11) calls attention to the important phenomenon of instability of deep neural network in computer vision. Instabilities in image classification, along with the potential safety and security issues they raise regarding the use of deep-learning vision systems in mission-critical systems, have been discussed extensively in the literature. The authors expose an analogous instability phenomenon in deep-learning–based image reconstruction, where a deep neural network is trained to solve an imaging inverse problem. They are concerned about potential safety issues in applications, such as medical imaging. Antun et al. propose a stability test to diagnose stability problems and describe software implementation of the test for inspecting such systems.

Jon Kleinberg of Cornell University and coauthors Jens Ludwig, Sendhil Mullainathan, and Cass R. Sunstein (12) close the special issue by addressing a fundamental new concern with the possible side-effects of deploying machine learning in daily life: Might they, by relying on data encoding human judgements, systematize discrimination and bias? They summarize their argument as follows: "... existing legal, regulatory, and related systems for detecting discrimination were originally built for a world of human decision makers, unaided by algorithms. Without changes to these systems, the introduction of algorithms will not help with the challenge of detecting discrimination and could potentially make the whole problem worse." The authors finish on an optimistic note: "Algorithms by their nature require a far greater level of specificity than is usually involved with human decision making, which in some sense is the ultimate 'black box.' With the right legal and regulatory systems in place, algorithms can serve as something akin to a Geiger counter that makes it easier to detect—and hence prevent—discrimination" (12).

These articles expose many surprises, paradoxes, and challenges. They remind us that there are many academic research opportunities emerging from this rapidly developing field. Mentioning only a few: Deep learning might be deployed more broadly in science itself, thereby accelerating the progress of existing fields; theorists might develop better understanding of the conundrums and paradoxes posed by this decade's deep-learning revolution; and scientists might understand better how industry-driven innovations in machine learning are affecting societal-level systems. Such opportunities will be challenging to pursue, not least because they demand new resources and talent. We hope that this special issue stimulates vigorous new scientific efforts pursuing such opportunities, leading perhaps to further discussions on deep learning in the pages of future editions of PNAS.

Acknowledgments

The authors thank Dame Jillian Sackler for her many years of sponsorship of National Academy of Sciences Sackler Colloquia. They also thank many Washington, DC-area residents, including representatives of the DoD, NIH, and NSF for participating. Many graduate students came to Washington, DC from around the United States to participate actively. The National Academy of Sciences and PNAS staff have also been very helpful at every stage.

1 The White House, Artificial intelligence for the American people. https://www.whitehouse.gov/briefings-statements/artificial-intelligence-american-people/. Accessed 3 November 2020.
2 T. J. Sejnowski, The unreasonable effectiveness of deep learning in artificial intelligence. *Proc. Natl. Acad. Sci. U.S.A.* **117**, 30033–30038 (2020).
3 E. P. Wigner, The unreasonable effectiveness of mathematics in the physical sciences. *Commun. Pur. Appl. Anal.* **13**, 1–14 (1960).
4 A. Halevy, P. Norvig, F. Pereira, The unreasonable effectiveness of data. *IEEE Intell. Syst.* **24**, 8–12 (2009).
5 T. Poggio, A. Banburski, Q. Liao, Theoretical issues in deep networks. *Proc. Natl. Acad. Sci. U.S.A.* **117**, 30039–30045 (2020).
6 C. D. Manning, K. Clark, J. Hewitt, U. Khandelwal, O. Levy, Emergent linguistic structure in artificial neural networks trained by self-supervision. *Proc. Natl. Acad. Sci. U.S.A.* **117**, 30046–30054 (2020).

7 K. Cranmer, J. Brehmer, G. Louppe, The frontier of simulation-based inference. *Proc. Natl. Acad. Sci. U.S.A.* **117**, 30055–30062 (2020).
8 P. L. Bartlett, P. M. Long, G. Lugosi, A. Tsigler, Benign overfitting in linear regression. *Proc. Natl. Acad. Sci. U.S.A.* **117**, 30063–30070 (2020).
9 D. Bau *et al.*, Understanding the role of individual units in a deep neural network. *Proc. Natl. Acad. Sci. U.S.A.* **117**, 30071–30078 (2020).
10 A. Barreto, S. Hou, D. Borsa, D. Silver, D. Precup, Fast reinforcement learning with generalized policy updates. *Proc. Natl. Acad. Sci. U.S.A.* **117**, 30079–30087 (2020).
11 V. Antun, F. Renna, C. Poon, B. Adcock, A. C. Hansen, On instabilities of deep learning in image reconstruction and the potential costs of AI. *Proc. Natl. Acad. Sci. U.S.A.* **117**, 30088–30095 (2020).
12 J. Kleinberg, J. Ludwig, S. Mullainathan, C. R. Sunstein, Algorithms as discrimination detectors. *Proc. Natl. Acad. Sci. U.S.A.* **117**, 30096–30100 (2020).