10°
0°
350°
40°
50°
60°
70°
80°
100°
120°
130°
140°
170°
180°
190°
230°
240°
260°
280°
290°
300°
310°
320°
"

The Deep (Learning) Transformation of Mobile and Embedded Computing

Nicholas D. Lane, University of Oxford

Pete Warden, Google

Mobile and embedded devices increasingly rely on deep neural networks to understand the world—a feat that would have overwhelmed their system resources only a few years ago. Further integration of machine learning and embedded/mobile systems will require additional breakthroughs of efficient learning algorithms that can function under fluctuating resource constraints, giving rise to a field that straddles computer architecture, software systems, and artificial intelligence.

In the past three years, deep learning techniques have quietly transformed how mobile and embedded devices interpret and react to the world.[1] Discriminative tasks performed by mobile phones and smartwatches—such as the recognition of faces, words, and objects—increasingly rely on low-resource, high-efficiency versions of deep learning models that previously needed cloud-scale resources to function. Through a variety of recently developed methods, it is now possible to use on-device-only resources, or a blend of on-device and cloud-compute resources, to bring the accuracy and robustness of deep neural networks to devices in our homes, offices, cars, and pockets.

These advances are gradually erasing the gap in machine learning quality that previously existed for constrained compute platforms and replacing it with (arguably) human-level, or better, performance for key cognitive and perceptual tasks (for example, machine translation, image understanding, and speech synthesis). This in turn is transforming what is possible in embedded/mobile systems even at a consumer-level of quality. Examples of this include Microsoft's Seeing AI, a mobile vision application that can see and describe accurately the environment to sight-impaired people; or Babylon from Babylon Health, a digital assistant capable of medical diagnosis and advice at a level comparable to calling hospital non-emergency phone help services. Both of these mobile applications are used on daily basis by thousands (in the case of Babylon), and millions (in the case of Seeing AI) of people and are powered to varying degrees by efficient, mobile-based, deep learning models.

Incredibly, these dramatic changes are just the beginning. Deep learning, and machine learning in general, continue their expansion from what is today largely classification and perceptual tasks, to roles that make an

impact across the complete stack of mobile and embedded computing. Deep learning techniques are already at the heart of control algorithms for autonomous systems (ranging from home robots, to drones and cars). More broadly, we are witnessing the discovery of how even very mature system components and algorithms can have their performance and function radically improved by the integration of deep neural networks. For example, even in fundamental data structures like B+ trees,[2] but also in the areas of video codecs, network protocols, and data compression and encryption.[3]

Seamlessly blending machine learning into the design and operation of mobile/embedded systems is an approach that will continue to gain momentum. In fact, within this transformation, deep learning is poised to have an even more pivotal role in the evolution of smart devices (such as phones, watches, and embedded sensors) than what we have already witnessed. It is through this combination of learning algorithms and the rethinking of mobile and embedded systems design—along with further research investment into the use of deep neural networks in support of the sensor systems' inference and reasoning needs—that will give rise to the next generation of smart devices necessary to drive future ambitious sensor-driven applications.

For these reasons, we must continue to advance our knowledge of how to best reduce, control, and shape the system's resource demands (including energy, compute, and memory) of deep models and algorithms. This is a foundational building block, as without progress the resource constraints of mobile and embedded devices will limit our ability to utilize deep learning and other machine learning techniques. Although it is becoming increasingly routine for deep learning to appear within this class of computing, the process of migrating a brand new deep learning innovation to a constrained platform remains a black art requiring many person-hours of highly trained experts. These deep networks are composed of hundreds of layers of interconnected nodes, and a single recognition task can demand the evaluation of hundreds of millions of model parameters. Representations of such models, and the associated inference algorithms can easily introduce extreme levels of resource overhead. To completely address the barriers that currently exist between learning algorithms and embedded/mobile devices will likely require a complete redesign of deep model representations and algorithms; our recent progress on this front, while extremely promising, only represents the first steps towards making such progress.

Solving the challenges of machine learning efficiency will require advances across a range of interdependent domains, including hardware, systems, and the learning algorithms themselves. Promising advances are already beginning to appear in each of these traditionally separate fields.[4-6] This progress is further accelerated by the maturing of large-scale, often commercially supported, deep learning tools, libraries, runtimes, and frameworks that are starting to directly address the specific needs of constrained devices (examples include TensorFlow, Caffe2, SNPE, Compute Library, and TensorRT from Google, Facebook, Qualcomm, ARM, and Nvidia respectively). More importantly, this collection of software is beginning to offer building blocks that directly support much-needed fundamental research in this area by simplifying key steps like: enabling deep models to be easily tested and profiled on Android devices, matrix operation libraries highly tuned for specific processor architectures and even opening access to typically inaccessible non-CPU processors like the DSP and GPU present on mobile and embedded device SoCs.

[DEEP LEARNING IS POISED TO HAVE AN EVEN MORE PIVOTAL ROLE IN THE EVOLUTION OF SMART DEVICES THAN WHAT WE HAVE ALREADY WITNESSED.]

IN THIS ISSUE

We believe this special issue offers a representative snapshot of the current breadth of deep learning research for mobile and embedded devices. These six cross-cutting articles examine core issues of algorithm efficiency, hardware specialization, sensor processing, activity recognition, and applications.

Beginning with "Exploiting Typical Values to Accelerate Deep Learning,"

Andreas Moshovos, Jorge Albericio, Patrick Judd, Alberto Delmás Lascorz, Sayeh Sharify, Zissis Poulos, Tayler Hetherington, Tor Aamodt, and Natalie Enright Jerger present innovative ideas in hardware optimization for deep learning workloads that exploit core inefficiencies of existing deep model representation. Research of this type is changing the processor architecture design of the SoCs that are appearing in mobile and embedded systems. Complementing this perspective, in "Deep Learning for the Internet of Things," Shuochao Yao, Yiran Zhao, Aston Zhang, Shaohan Hu, Huajie Shao, Chao Zhang, Lu Su, and Tarek Abdelzaher detail the design of deep networks specifically for processing data captured by embedded sensors (including microphones, accelerometers, magnetometers), along with a set of software-based methods for scaling down their resource needs to fit the constraints of such devices. These two articles offer examples of software- and hardware-grounded efficiency methods that aim to lower the system resource needs of deep models to acceptable levels.

The overarching concern of privacy in deploying this technology is considered in "Private and Scalable Personal Data Analytics Using Hybrid Edge-to-Cloud Deep Learning," by Seyed Ali Osia, Ali Shahin Shamsabadi, Ali Taheri, Hamid R. Rabiee, and Hamed Haddadi. The authors examine how models partitioned between the device and cloud, a common design pattern, can be architected to provide guarantees to users as to what is, and is not, inferred from captured data.

The issue concludes with three articles that describe application-level advances extending from deep learning usage. In "Deep Learning for Human Activity Recognition in Mobile Computing," Thomas Plötz and Yu Guan offer an overview and specific examples of how the models of human behavior and context for mobile and embedded devices have dramatically changed by embracing the principles and algorithms of neural networks. In "Breathing-Based Authentication on Resource-Constrained IoT Devices Using Recurrent Neural Networks," Jagmohan Chauhan, Suranga Seneviratne, Yining Hu, Archan Misra, Aruna Seneviratne, and Youngki Lee describe a novel application for mobile user authentication that relies on deep learning methods; and finally, in "Finding Small-Bowel Lesions: Challenges in Endoscopy-Image-Based Learning Systems," Jungmo Ahn, Huynh Nguyen Loc, Rajesh Krishna Balan, Youngki Lee, and JeongGil Ko outline the potential for new medical instruments that use machine learning, with a focus on automated assessment made possible by convolutional neural networks.

THE ROAD AHEAD

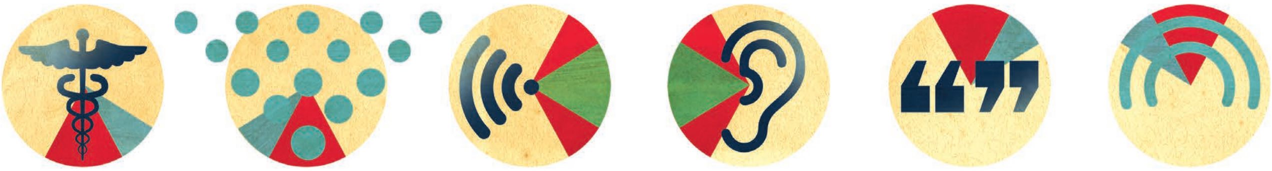
Integrating and adopting deep learning within all aspects of constrained classes of computing—things such as phones, watches, drones, robots, and sensors—is increasingly pervasive and common. Fueling this revolutionary shift are the academic and industrial researchers who can bring the previously disjoint worlds of machine learning and embedded/mobile computing together through advances in processor architecture, mobile systems, and learning algorithms. The papers in this special issue represent just a small sampling of ongoing research in this emerging field.

Speculating about this area's future, we think in the near term it is likely we will see continued improvements in the ability of devices to understand and reason about even the most complex environments as advances from deep learning migrate into sensor inference. This will be complemented by a shift from inference-only use of deep models, to one in which training and adaption of learning models

ABOUT THE AUTHORS

NICHOLAS D. LANE is an associate professor at the University of Oxford. His interests include efficient deep learning under mobile and embedded system constraints, and more broadly topics at the intersection of machine learning and software systems. Lane received his PhD from Dartmouth College. Contact him at nicholas.lane@cs.ox.ac.uk.

PETE WARDEN is the tech lead of the TensorFlow Mobile team, and was formerly the CTO of Jetpac, which was acquired by Google for its deep learning technology optimized to run on mobile and embedded devices. He is the author of a recent O'Reilly ebook *Building Mobile Applications with TensorFlow*. Contact him at petewarden@google.com.



occur directly on-device. More profoundly, whereas current integration of deep learning is almost exclusively restricted to classification tasks, there will be a broader trend of keeping deep learning to enable control and decision tasks. We believe a persistent and longer-term trend will be for learning algorithms, dominated by deep learning techniques, to replace (and augment) the application and system component logic within embedded and mobile systems. This will broadly change the internal functioning of these devices including the operating system, wireless and networking stacks, and sensor processing pipelines. Such changes will bring about leaps in efficiency and functionality and could give embedded and mobile devices the ability to learn and adapt dynamically to real-world situations and human behavior, a capability that has long been sought but proved difficult to robustly achieve.

REFERENCES

1. N.D. Lane et al., "Squeezing Deep Learning into Mobile and Embedded Devices," *IEEE Pervasive Computing*, vol. 16, no. 3, 2017. pp. 82–88.
2. T. Kraska et al., "The Case for Learned Index Structures," Arxiv.org, 11 Dec. 2017; https://arxiv.org/abs/1712.01208.
3. G. Toderici et al., "Variable Rate Image Compression with Recurrent Neural Networks," *Proc. Int'l Conf. Learning Representations* (ICLR), 2016; https://arxiv.org/abs/1511.06085.
4. R. LiKamWa et al., "RedEye: Analog ConvNet Image Sensor Architecture for Continuous Mobile Vision," *Proc. 43rd ACM/IEEE Int'l Symp. Computer Architecture* (ISCA 16), 2016, pp. 255–266.
5. S. Han et al., "Deep Compression: Compressing Deep Neural Networks, with Pruning, Trained Quantization and Huffman Coding," *Proc. Int'l Conf. Learning Representations* (ICLR 16), 2016; https://arxiv.org/pdf/1510.00149.pdf.
6. G. Huang et al., "Densely Connected Convolutional Networks," *Proc. IEEE Conf. Computer Vision and Pattern Recognition* (CVPR 17), 2017, pp. 2261–2269.