Does AI have a hardware problem?

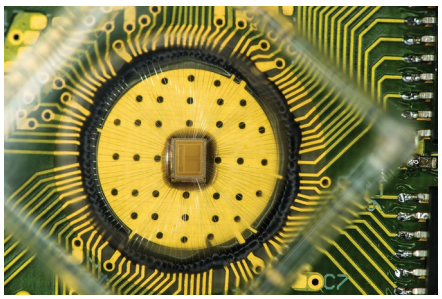
As deep neural networks continue to improve and grow, innovations in hardware will be required in order to meet the increasing computational demands.

Deep learning has been at the forefront of recent developments in artificial intelligence (AI). It involves a set of machine learning algorithms that are inspired by biological neural networks and can teach machines to find patterns in large amounts of data. These deep neural networks have led to significant improvements in areas such as speech and object recognition, and have been the basis of computer programs that exhibit superhuman capabilities in specific tasks.

The most high-profile demonstration of the current capabilities of such methods is probably that of AlphaGo[1]. Developed by researchers at DeepMind in London, the program defeated the world champion of the game Go, Lee Sedol, in a five-game match back in March 2016; the score was 4 games to 1. And at this point, AlphaGo's only competition is from better versions of itself. In October 2017, the DeepMind team reported an updated program — AlphaGo Zero[2] — that uses reinforcement learning and trains solely by playing games against itself; AlphaGo had relied on unsupervised learning from millions of human expert moves. Pitted against the AlphaGo program that had defeated Sedol, AlphaGo Zero won 100 games to 0.

Deep neural networks involve multiple layers of 'neurons' connected through digital 'synapses'. They are trained on large datasets, accompanied by the answers to the desired tasks, during which the strength, or weight, of the connections between the neurons is adjusted until the top-level outputs are correct. A trained neural network, run with the weights determined during the training phase, can then be applied to new data in a step known as inference.

The recent success of deep neural networks has been driven by advances in algorithms and network architectures, but also, notably, through the growing availability of vast amounts of data and the continuing development of ever more powerful computers. And at this point, the computational demands of deep neural networks with state-of-the-art accuracy are considerable[3]. As Yiyu Shi and colleagues illustrate in a Perspective in this issue of *Nature Electronics*, this presents an emerging problem for deep neural networks and, in particular, their potential implementation on mobile and embedded devices, such as smart sensors or wearable devices, where area and power resources are limited.



A two-dimensional array of phase-change memory devices. Credit: courtesy of IBM Research.

The researchers — who are based at the University of Notre Dame, the University of California, Los Angeles and Huazhong University of Science and Technology — examine data on the accuracy and size of deep neural networks, and the capacity of different hardware platforms. They show that gaps exist between the scaling of deep neural networks for edge inference (where inference is performed locally on embedded platforms) and the scaling of complementary metal–oxide–semiconductor (CMOS) technology — and these gaps are growing. As deep neural networks have become more accurate, their size (layers, parameters and number of operations) has increased dramatically. But, as Shi and colleagues show, the performance of the typical hardware platforms — graphics processing units (GPUs), field-programmable gate arrays (FPGAs) and application-specific integrated circuits (ASICs) — cannot keep pace with the increasing size of leading deep neural network designs. Similarly, the energy efficiency of the memory required by the hardware platforms to accommodate the networks cannot keep pace with the increasing size of the networks.

As Shi and colleagues note, "CMOS scaling does not offer much help in meeting the increasingly demanding requirements for computation density and energy efficiency, so innovations in architecture, circuit and device are required instead." They thus go on to examine different architecture and algorithm innovations that could, jointly, help to bridge these gaps.

One such approach is to try to move away from conventional von Neumann computing systems in which memory and processing units are physically separated. Nanoscale resistive memory devices (memristive devices) can, for example, be used for both processing and memory. However, device variability remains an issue, which limits the accuracy with which computations can be performed. Elsewhere in this issue, Manuel Le Gallo and colleagues at IBM Research – Zurich and ETH Zurich show that this problem could potentially be circumvented by combining in-memory processing using resistive memory devices with conventional digital processing. Here the in-memory processor unit — specifically, an array of phase-change memory devices (pictured) — carries out the bulk of a computational task, and the conventional processing unit iteratively improves the accuracy of the solution. Le Gallo and colleagues illustrate the capabilities of the approach, which they term mixed-precision in-memory computing, by solving systems of linear equations. But the approach has already also been applied to the training of deep neural networks[4].

The potential of developing devices and chips that are specifically suited to AI applications has also awakened interest in chip start-ups. Earlier this year, *The New York Times* reported that there are currently at least 45 start-ups working on such chips, and that venture capitalists last year invested over US$1.5 billion in chip start-ups, which was almost double the investment of two years ago[5].

The possible benefits of this technology are considerable, and researchers from across academia and industry are responding to the hardware challenges — and opportunities — machine learning and AI present. ❐

Published online: 17 April 2018
https://doi.org/10.1038/s41928-018-0068-2

References
1. Silver, D. et al. *Nature* **529**, 484–489 (2016).
2. Silver, D. et al. *Nature* **550**, 354–359 (2017).
3. Sze, V., Chen, Y.-H., Yang, T.-J. & Emer, J. S. *Proc. IEEE* **105**, 2295–2329 (2017).
4. Nandakumar, S. R. et al. Preprint at https://arxiv.org/abs/1712.01192 (2017).
5. Metz, C. Big bets on A.I. open a new frontier for chip start-ups, too. *The New York Times* (14 January 2018); https://go.nature.com/2Gz7HHz