https://doi.org/10.1038/s41928-023-01014-x

AI hardware has an energy problem

New energy-efficient electronic hardware will be required to sustain the development of machine learning and artificial intelligence.

ChatGPT — the artificial intelligence (AI) chatbot developed by OpenAI — was first released on 30 November 2022. Its results can be both remarkable and terrible, depending on what you ask it to do. But its release — together with other generative AI tools such as Midjourney, which generates digital images from natural language descriptions — has led to rapid discussions about the implications of such technology for science, education and society. (The current editorial polices of the Nature Portfolio journals regarding AI can be found here, which clarify that ChatGPT cannot be an author on a research paper and any use in a manuscript should be properly documented in the Methods section.)

The improving capabilities of these machine learning techniques can be linked to the rapid increase in the scale of the artificial neural networks they rely on. GPT-3 — the large language model that was originally the basis for ChatGPT — has, for instance, 175 billion parameters[1]. But the size of the networks places considerable demands on the underlying electronic hardware, particularly their energy consumption. As Alexander Conklin and Suhas Kumar explain in a Comment article in this issue of *Nature Electronics*: "Up until 2019, the computing capacity required to train the largest AI models doubled around every 3.4 months [...] Hypothetically, and even when accounting for continual hardware improvements, this growth rate would mean the energy required to train a leading AI model would exceed global yearly energy expenditure by 2030".

Training these AI models is also expensive: GPT-3 probably cost around US$12 million to train[2]. As a result, the growth in leading AI models has likely slowed over the last few years, and we are potentially entering an era of economics-limited computing. Conklin and Kumar go on to show that substantial improvements in computing energy efficiency will be required to solve major computing problems — such as planetary-scale weather modelling, real-time, brain-scale modelling and human evolutionary simulation — in the twenty-first century. And thus call for the development of novel strategies in computing technology, energy production and commercial computing budgets to address this.

In terms of investing in new approaches to computing technology, Conklin and Kumar — who are based at Rain AI in San Francisco and Sandia National Laboratories — highlight the potential of memristive devices (or memristors) to deliver energy-efficient AI hardware. Such devices can be used for both information processing and memory. And they can be used to build memristive crossbar arrays that provide large parallelism for matrix multiplication operations, a key computation in most AI models.

Memristive devices are typically based on metal oxides (such as titanium dioxide) or phase-change materials (such as germanium antimony tellurium alloys). But they can be created from other systems. In a Review article also in this issue, Joshua Yang and colleagues explore the potential of memristive devices based on van der Waals materials.

The researchers — who are based at the University of Southern California and the National Chung Hsing University — examine how imperfections in these layered two-dimensional materials, together with inherent physicochemical properties, can create a range of switching mechanisms. Three key switching mechanisms are highlighted — those based on electrochemical metallization, those based on valence-change mechanisms and those based on phase-change mechanisms — each of which is associated with a specific group of imperfections. The different mechanisms can then be exploited in different applications, and Yang and colleagues explore their potential for use in memory devices, radiofrequency switches and neuromorphic electronics.

Work on these van der Waals memristors is only at an early stage, and the development of any computer technology requires advances in computer architecture, as well as computational devices. But when it comes to developing hardware for machine learning, is there a disconnect between the device community and the computer architecture community? This is what Nathaniel Tye, Stephan Hofmann and Phillip Stanley-Marbell at the University of Cambridge argue in a Perspective article elsewhere in this issue.

Focusing on machine learning accelerators, the researchers explore this disconnect and how it restricts progress. They then suggest that directly mapping computational problems in machine learning to materials and device properties provides a route forward. And highlight cases in which materials and devices have been successfully applied as solutions to computational problems, including the use of memristive crossbar arrays for matrix multiplication. The team also propose metrics to facilitate comparison between different solutions to machine learning tasks. The ability to accurately benchmark the performance of different AI hardware — whether based on traditional silicon systems or novel material devices — should help close the gap between communities, a step that will be required if we are to develop the energy-efficient electronic hardware the field needs.

Published online: 26 July 2023

References

1. Brown, T. B. et al. Preprint at https://doi.org/10.48550/arXiv.2005.14165 (2020).
2. Wiggers, K. OpenAI's massive GPT-3 model is impressive, but size isn't everything. *VentureBeat* https://go.nature.com/3NjpWWc (1 June 2020).