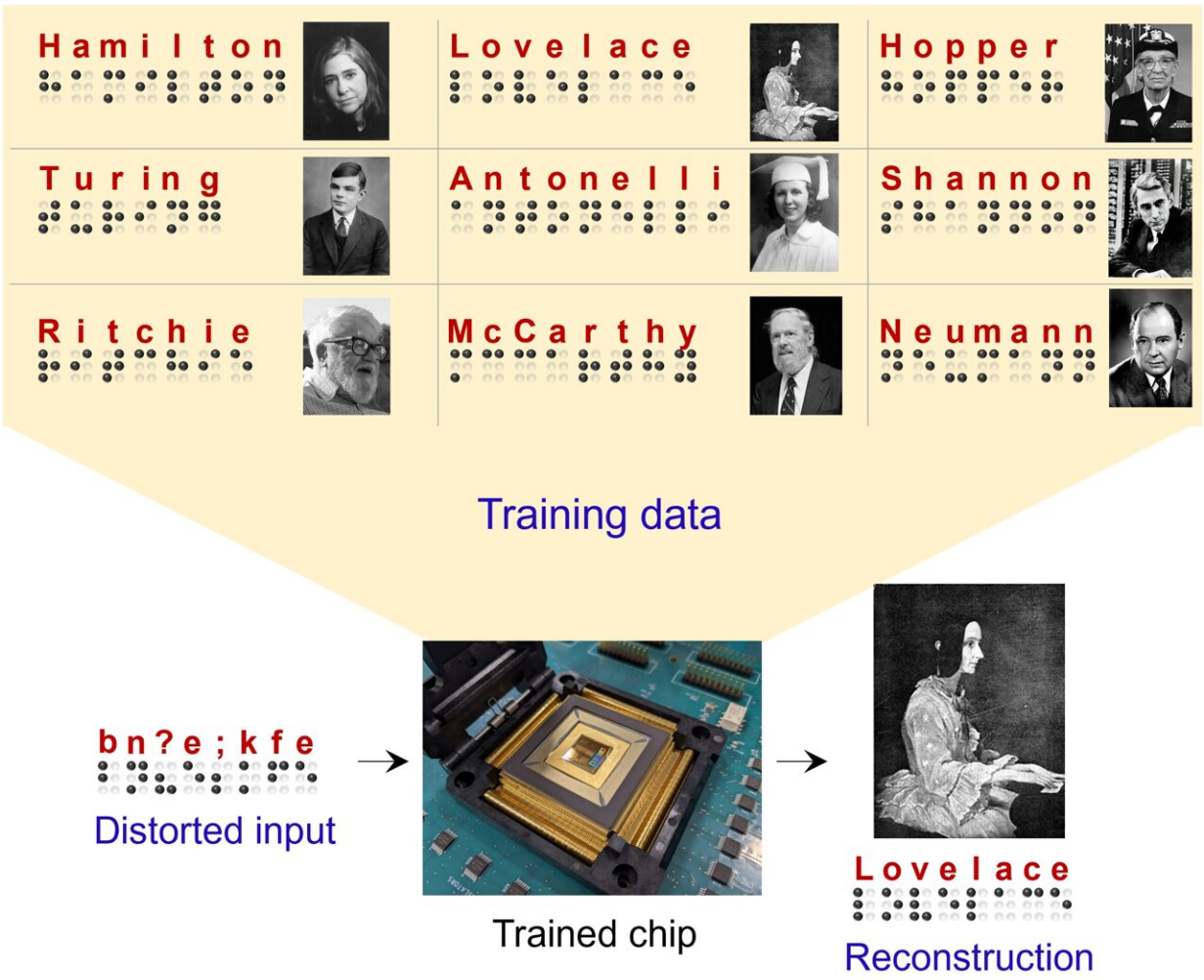
A memristor crossbar-based learning system for scalable and energy-efficient AI

December 14 2022, by Ingrid Fadelli



A chip consisting of memristor crossbars was trained using a local on-chip learning algorithm. The team demonstrated that their approach could accurately reconstruct Braille representations of nine famous computer scientists from highly distorted inputs. Credit: Yi et al.

Deep-learning models have proven to be highly valuable tools for making predictions and solving real-world tasks that involve the analysis of data. Despite their advantages, before they are deployed in real software and devices such as cell phones, these models require extensive training in physical data centers, which can be both time and energy consuming.

Researchers at Texas A&M University, Rain Neuromorphics and Sandia National Laboratories have recently devised a new system for training deep learning models more efficiently and on a larger scale. This system, introduced in a paper published in *Nature Electronics*, relies on the use of new training algorithms and memristor crossbar hardware, that can carry out multiple operations at once.

"Most people associate AI with health monitoring in smart watches, face recognition in smart phones, etc., but most of AI, in terms of energy spent, entails the training of AI models to perform these tasks," Suhas Kumar, the senior author of the study, told TechXplore.

"Training happens in warehouse-sized data centers, which is very expensive both economically and in terms of carbon footprint. Only fully trained models are then downloaded onto our low-power devices."

Essentially, Kumar and his colleagues set out to devise an approach that could reduce the carbon footprint and financial costs associated with the training of AI models, thus making their large-scale implementation easier and more sustainable. To do this, they had to overcome two key limitations of current AI training practices.

The first of these challenges is associated with the use of inefficient hardware systems based on graphics processing units (GPUs), which are

not inherently design to run and train deep learning models. The second entails the use of ineffective and math-heavy software tools, specifically utilizing the so-called backpropagation algorithm.

"Our objective was to use new hardware and new algorithms," Kumar explained. "We leveraged our previous 15 years of work on memristor-based hardware (a highly parallel alternative to GPUs), and recent advances in brain-like efficient algorithms (a non-backpropagation local learning technique). Though advances in hardware and software existed previously, we codesigned them to work with each other, which enabled very power efficient AI training."

The training of deep neural networks entails continuously adapting its configuration, comprised of so-called "weights," to ensure that it can identify patterns in data with increasing accuracy. This process of adaptation requires numerous multiplications, which conventional digital processors struggle to perform efficiently, as they will need to fetch weight-related information from a separate memory unit.

"Nearly all training today is performed using the backpropagation algorithm, which employs significant data movement and solving math equations, and is thus suited to digital processors," Suin Yi, lead author of the study, told TechXplore.

"As a hardware solution, analog memristor crossbars, which emerged within the last decade, enable embedding the synaptic weight at the same place where the computing occurs, thereby minimizing data movement. However, traditional backpropagation algorithms, which are suited for high-precision digital hardware, are not compatible with memristor crossbars due to their hardware noise, errors and limited precision."

As conventional backpropagation algorithms were poorly suited to the system they envisioned, Kumar, Yi and their colleagues developed a new

co-optimized learning algorithm that exploits the hardware parallelism of memristor crossbars. This algorithm, inspired by the differences in neuronal activity observed in neuroscience studies, is tolerant to errors and replicates the brain's ability to learn even from sparse, poorly defined and "noisy" information.

"Our algorithm-hardware system studies the differences in how the synthetic neurons in a neural network behave differently under two different conditions: one where it is allowed to produce any output in a free fashion, and another where we force the output to be the target pattern we want to identify," Yi explained.

"By studying the difference between the system's responses, we can predict the weights needed to make the system arrive at the correct answer without having to force it. In other words, we avoid the complex math equations backpropagation, making the process more noise resilient, and enabling local training, which is how the brain learns new tasks."

The brain-inspired and analog-hardware-compatible algorithm developed as part of this study could thus ultimately enable the energy-efficient implementation of AI in edge devices with small batteries, thus eliminating the need for large cloud servers that consume vast amounts electrical power. This could ultimately help to make the large-scale training of deep learning algorithms more affordable and sustainable.

"The algorithm we use to train our neural network combines some of the best aspects of deep learning and neuroscience to create a system that can learn very efficiently and with low-precision devices," Jack Kendall, another author of the paper, told TechXplore.

"This has many implications. The first is that, using our approach, AI models that are currently too large to be deployed can be made to fit in

cellphones, smartwatches, and other untethered devices. Another is that these networks can now learn on-the-fly, while they're deployed, for instance to account for changing environments, or to keep user data local (avoiding sending it to the cloud for training)."

In initial evaluations, Kumar, Yi, Kendall and their colleague Stanley Williams showed that their approach can reduce the power consumption associated with AI training by up to 100,000 times when compared to even the best GPUs on the market today. In the future, it could enable the transfer of massive data centers onto users' personal devices, reducing the carbon footprint associated with AI training, and promoting the development of more artificial neural networks that support or simplify daily human activities.

"We next plan to study how these systems scale to much larger networks and more difficult tasks," Kendall added. "We also plan to study a variety of brain-inspired learning algorithms for training deep neural networks and find out which of these have perform better in different networks, and with different hardware resource constraints. We believe this will not only help us understand how to best perform learning in resource constrained environments, but it may also help us understand how biological brains are able to learn with such incredible efficiency."

More information: Su-in Yi et al, Activity-difference training of deep neural networks using memristor crossbars, *Nature Electronics* (2022). DOI: 10.1038/s41928-022-00869-w