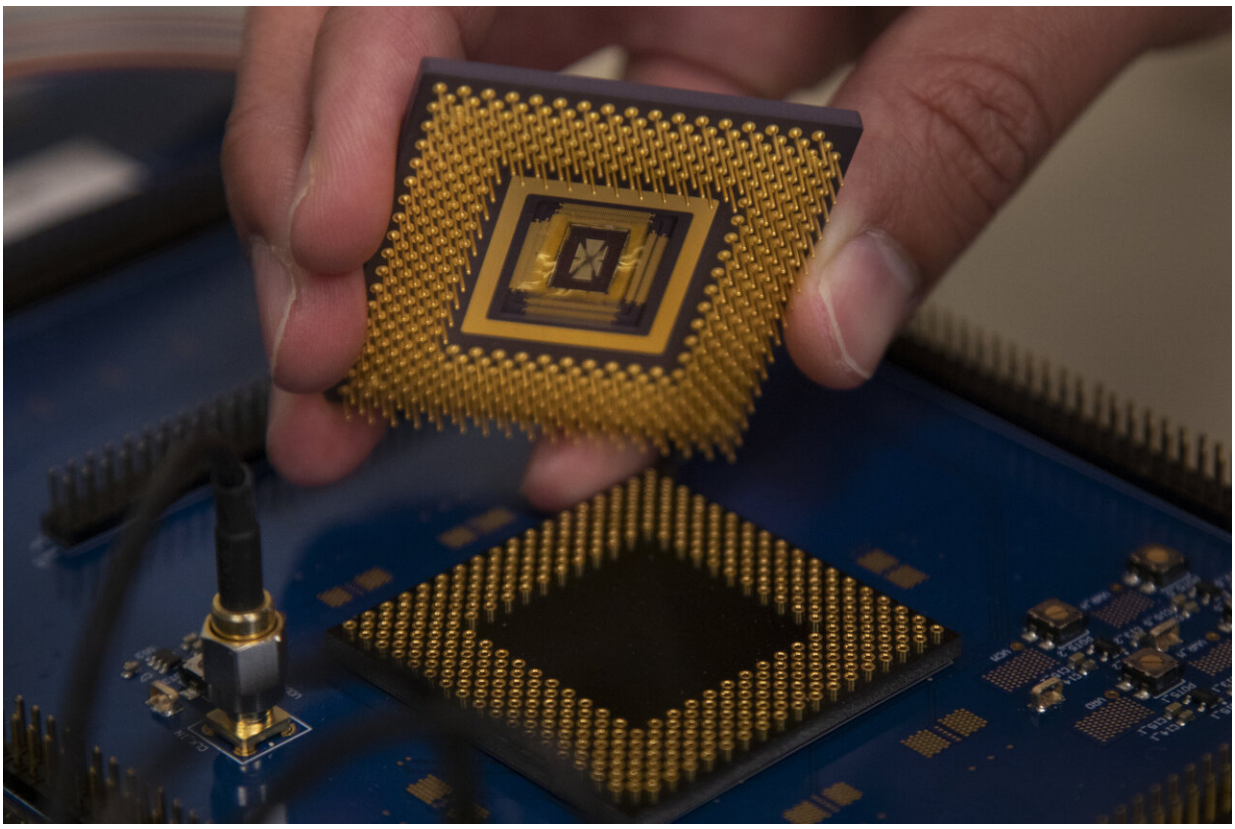
First programmable memristor computer aims to bring AI processing down from the cloud

July 17 2019, by Nicole Casal Moore



The memristor array chip plugs into the custom computer chip, forming the first programmable memristor computer. The team demonstrated that it could run three standard types of machine learning algorithms. Credit: Robert Coelius, Michigan Engineering

The first programmable memristor computer—not just a memristor array operated through an external computer—has been developed at the University of Michigan.

It could lead to the processing of artificial intelligence directly on small, energy-constrained devices such as smartphones and sensors. A smartphone AI processor would mean that voice commands would no longer have to be sent to the cloud for interpretation, speeding up response time.

"Everyone wants to put an AI processor on smartphones, but you don't want your cell phone battery to drain very quickly," said Wei Lu, U-M professor of electrical and computer engineering and senior author of the study in *Nature Electronics*.

In medical devices, the ability to run AI algorithms without the cloud would enable better security and privacy.

Why memristors are good for machine learning

The key to making this possible could be an advanced computer component called the memristor. This circuit element, an electrical resistor with a memory, has a variable resistance that can serve as a form of information storage. Because memristors store and process information in the same location, they can get around the biggest bottleneck for computing speed and power: the connection between memory and processor.

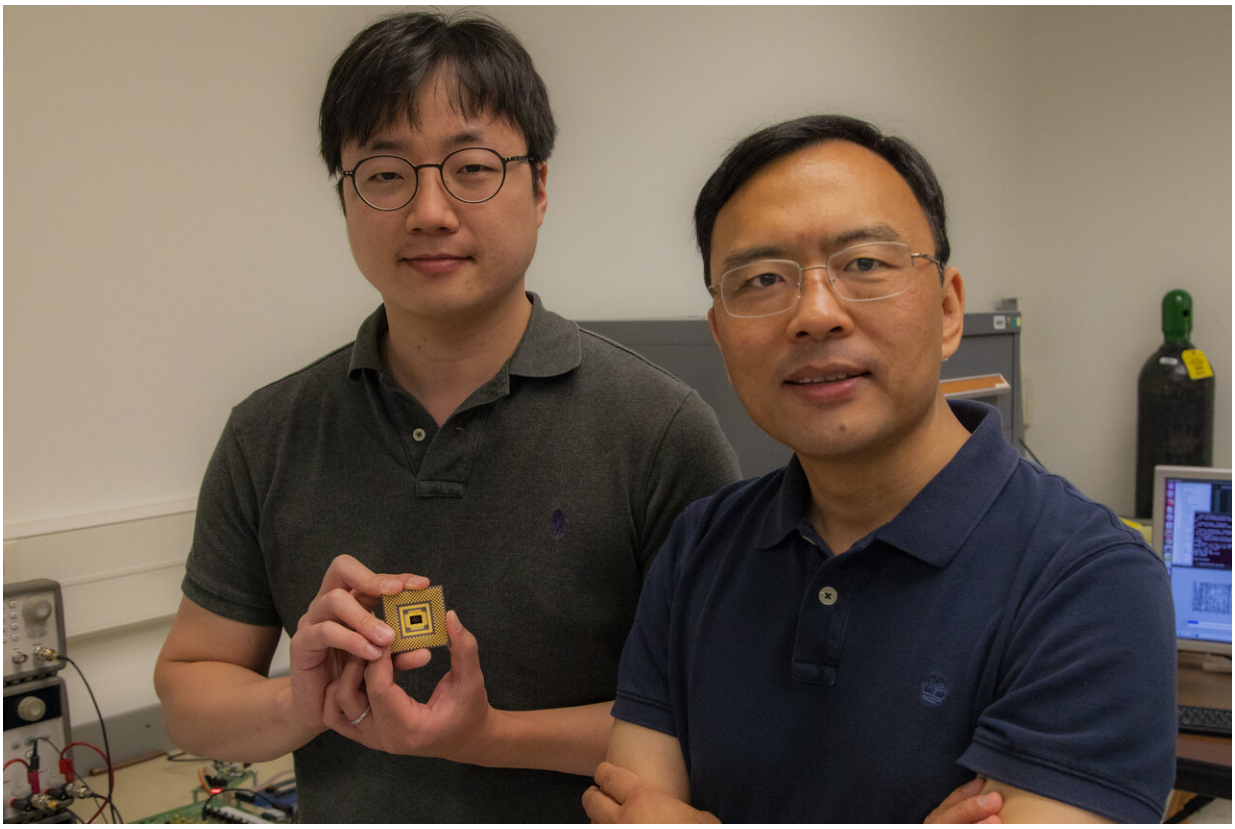
This is especially important for machine-learning algorithms that deal with lots of data to do things like identify objects in photos and videos—or predict which hospital patients are at higher risk of infection. Already, programmers prefer to run these algorithms on graphical processing units rather than a computer's main processor, the central

processing unit.

"GPUs and very customized and optimized digital circuits are considered to be about 10-100 times better than CPUs in terms of power and throughput." Lu said. "Memristor AI processors could be another 10-100 times better."

GPUs perform better at machine learning tasks because they have thousands of small cores for running calculations all at once, as opposed to the string of calculations waiting their turn on one of the few powerful cores in a CPU.

A memristor array takes this even further. Each memristor is able to do its own calculation, allowing thousands of operations within a core to be performed at once. In this experimental-scale computer, there were more than 5,800 memristors. A commercial design could include millions of them.



Wei Lu stands with first author Seung Hwan Lee, an electrical engineering PhD student, who holds the memristor array. Credit: Robert Coelius, Michigan Engineering

Memristor arrays are especially suited to machine learning problems. The reason for this is the way that machine learning algorithms turn data into vectors—essentially, lists of data points. In predicting a patient's risk of infection in a hospital, for instance, this vector might list numerical representations of a patient's risk factors.

Then, machine learning algorithms compare these "input" vectors with "feature" vectors stored in memory. These feature vectors represent certain traits of the data (such as the presence of an underlying disease). If matched, the system knows that the input data has that trait. The

vectors are stored in matrices, which are like the spreadsheets of mathematics, and these matrices can be mapped directly onto the memristor arrays.

What's more, as data is fed through the array, the bulk of the mathematical processing occurs through the natural resistances in the memristors, eliminating the need to move feature vectors in and out of the memory to perform the computations. This makes the arrays highly efficient at complicated matrix calculations. Earlier studies demonstrated the potential of memristor arrays for speeding up machine learning, but they needed external computing elements to function.

Building a programmable memristor computer

To build the first programmable memristor computer, Lu's team worked with associate professor Zhengya Zhang and professor Michael Flynn, both of electrical and computer engineering at U-M, to design a chip that could integrate the memristor array with all the other elements needed to program and run it. Those components included a conventional digital processor and communication channels, as well as digital/analog converters to serve as interpreters between the analog memristor array and the rest of the computer.

Lu's team then integrated the memristor array directly on the chip at U-M's Lurie Nanofabrication Facility. They also developed software to map machine learning algorithms onto the matrix-like structure of the memristor array.

The team demonstrated the device with three bread-and-butter machine learning algorithms:

- Perceptron, which is used to classify information. They were able to identify imperfect Greek letters with 100 percent accuracy

- Sparse coding, which compresses and categorizes data, particularly images. The computer was able to find the most efficient way to reconstruct images in a set and identified patterns with 100 percent accuracy
- Two-layer neural network, designed to find patterns in complex data. This two-layer network found commonalities and differentiating factors in breast cancer screening data and then classified each case as malignant or benign with 94.6 percent accuracy.

There are challenges in scaling up for commercial use—memristors can't yet be made as identical as they need to be and the information stored in the array isn't entirely reliable because it runs on analog's continuum rather than the digital either/or. These are future directions of Lu's group.

Lu plans to commercialize this technology. The study is titled, "A fully integrated reprogrammable memristor–CMOS system for efficient multiply–accumulate operations."

More information: Fuxi Cai et al. A fully integrated reprogrammable memristor–CMOS system for efficient multiply–accumulate operations, *Nature Electronics* (2019). DOI: 10.1038/s41928-019-0270-x

Provided by University of Michigan