A computing in-memory system based on stacked 3D resistive memories

September 1 2022, by Ingrid Fadelli

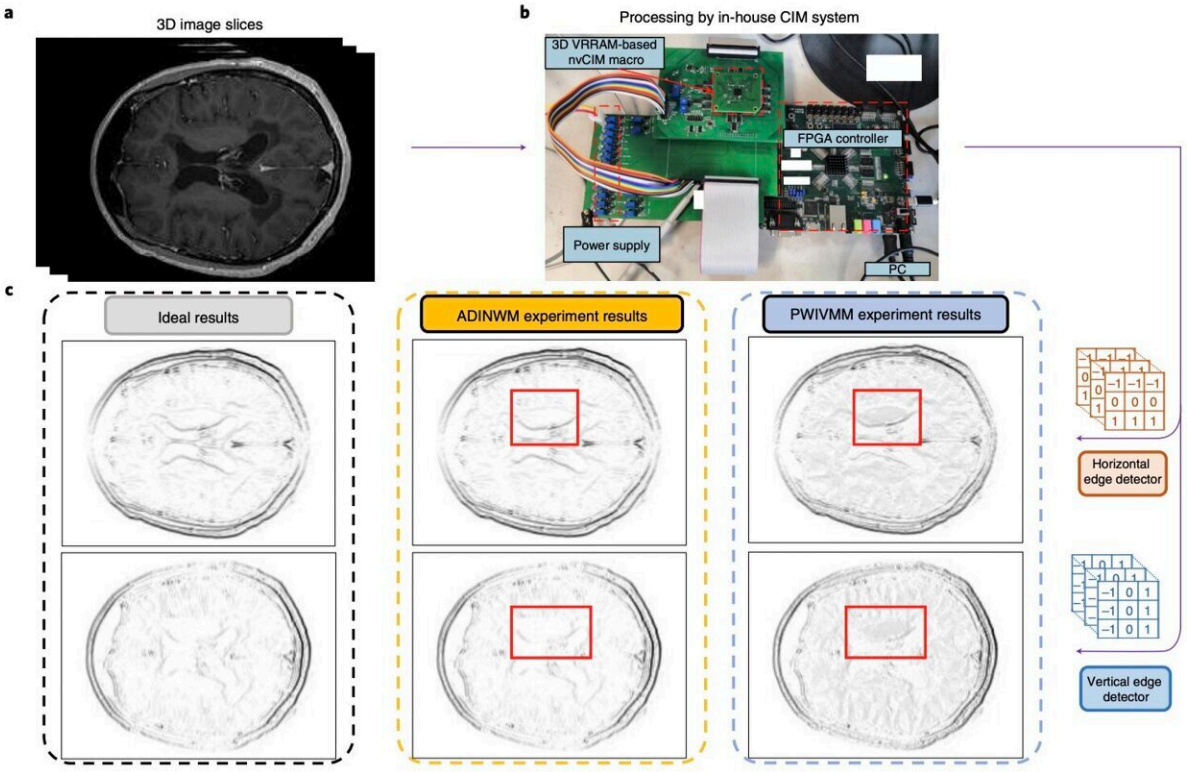

Figure summarizing the evaluation and performance of the researchers' computing-in-memory macro. Credit: Huo et al (*Nature Electronics*, 2022).

Machine learning architectures based on convolutional neural networks (CNNs) have proved to be highly valuable for a wide range of applications, ranging from computer vision to the analysis of images and

the processing or generation of human language. To tackle more advanced tasks, however, these architectures are becoming increasingly complex and computationally demanding.

In recent years, many electronics engineers worldwide have thus been trying to develop devices that can support the storage and computationally load of complex CNN-based architectures. This includes denser memory devices that can support large amounts of weights (i.e., the trainable and non-trainable parameters considered by the different layers of CNNs).

Researchers at the Chinese Academy of Sciences, Beijing Institute of Technology, and other Universities in China have recently developed a new computing-in-memory system that could help to run more complex CNN-based models more effectively. Their memory component, introduced in a paper published in *Nature Electronics*, is based on non-volatile computing-in-memory macros made of 3D memristor arrays.

"Scaling such systems to 3D arrays could provide higher parallelism, capacity and density for the necessary vector-matrix multiplication operations," Qiang Huo and his colleagues wrote in their paper. "However, scaling to three dimensions is challenging due to manufacturing and device variability issues. We report a two-kilobit non-volatile computing-in-memory macro that is based on a three-dimensional vertical resistive random-access memory fabricated using a 55 nm complementary metal-oxide-semiconductor process."

Resistive random-access memories, or RRAMs, are non-volatile (i.e., retaining data even after breaks in power supply) storage devices based on memristors. Memristors are electronic components that can limit or regulate the flow of electrical current in circuits, while recording the amount of charge that previously flowed through them.

RRAMs essentially work by varying the resistance across a memristor. While past studies have demonstrated the great potential of these memory devices, conventional versions of these devices are separate from computer engines, which limits their possible applications.

Computing-in-memory RRAM devices were designed to overcome this limitation, by embedding the computations inside the memory. This can greatly reduce the transfer of data between memories and processors, ultimately enhancing the overall system's energy-efficiency.

The computing-in-memory device created by Huo and his colleagues is a 3D RRAM with vertically stacked layers and peripheral circuits. The device's circuits were fabricated using 55 nm CMOS technology, the technology underpinning most integrated circuits on the market today.

The researchers evaluated their device by using it to carry out complex operations and to run a model for detecting edges in MRI brain scans. The team trained their models using two existing MRI datasets for training image recognition tools, known as the MNIST and CIFAR-10 datasets.

"Our macro can perform 3D vector-matrix multiplication operations with an energy efficiency of 8.32 tera-operations per second per watt when the input, weight and output data are 8,9 and 22 bits, respectively, and the bit density is 58.2 bit μm^{-2}," the researchers wrote in their paper. "We show that the macro offers more accurate brain MRI edge detection and improved inference accuracy on the CIFAR-10 dataset than conventional methods."

In initial tests, the computing-in-memory vertical RRAM system created by Huo and his colleagues achieved remarkable results, outperforming conventional RRAM approaches. In the future, it could thus prove to be highly valuable for running complex CNN-based models more energy-

efficiently, while also enabling better accuracies and performances.

More information: Qiang Huo et al, A computing-in-memory macro based on three-dimensional resistive random-access memory, *Nature Electronics* (2022). DOI: 10.1038/s41928-022-00795-x