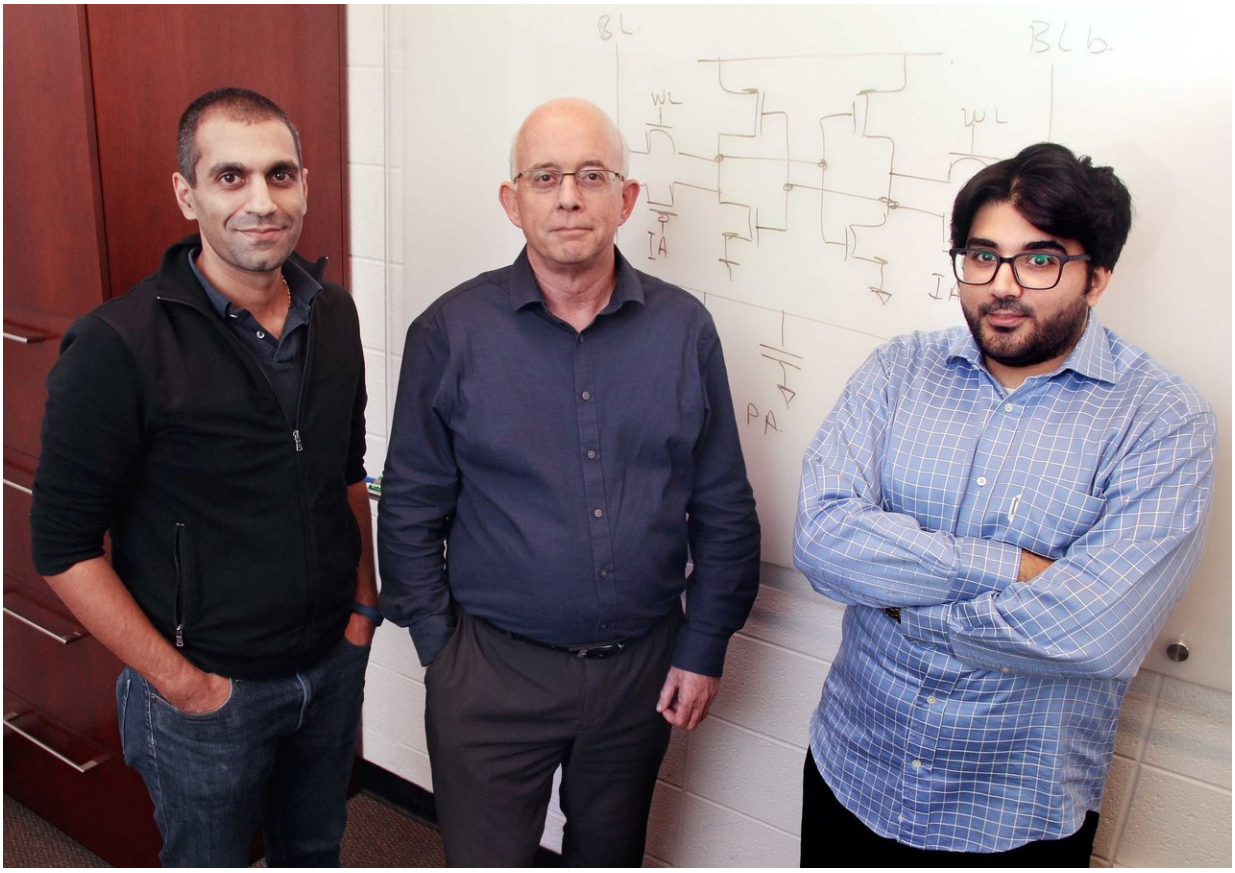
Chip ramps up artificial intelligence systems' performance

September 24 2018, by Adam Hadhazy



Professors Naveen Verma and Peter Ramadge, and Hossein Valavi, a graduate student, have fabricated a chip that markedly boosts the performance and efficiency of neural networks—computer algorithms modeled on the workings of the human brain. Photos by Frank Wojciechowski. Credit: Princeton University

Princeton researchers, in collaboration with Analog Devices Inc., have fabricated a chip that markedly boosts the performance and efficiency of neural networks—computer algorithms modeled on the workings of the human brain.

In a series of tests, the Princeton chip performed tens to hundreds of times better than other advanced, neural-network chips.

The researchers believe that with further development, the chip could help advance image recognition and numerous other neural-network applications, including artificial intelligence systems in autonomous vehicles and robots.

"This kind of improved performance could let mobile devices do intensive tasks, like recognizing their owner's face, without taking up too much time or eating into the device's battery life," said the paper's lead author Hossein Valavi, a graduate student in the lab of co-author Naveen Verma, an associate professor of electrical engineering at Princeton.

Other authors of the study, which published in *IEEE Symposium on VLSI Circuits*, in June, are Peter Ramadge, the Gordon Y.S. Wu Professor of Engineering and director of the Center for Statistics and Machine Learning, and Eric Nestler of Analog Devices Inc, a Massachusetts-based semiconductor company.

Artificial neural networks are complexes of interconnected units—akin to neurons in the human brain—that can be trained to make valuable decisions from data given in many different, possibly naturally-occurring, but structurally-complex forms. A key component of neural-network systems is accelerator chips, which boost computational performance, to enable large and powerful neural networks. But the accelerator chips themselves can suffer from bottlenecking due to the heavy dataflows coursing through their components.

The researchers took a fresh approach to eliminating much of this snarling traffic. The accelerator chip they fabricated works with the technique, called in-memory computing, which substantially reduces the energy and time used to fetch information by performing computations on data in place where it is stored, rather than moving it to a different location.

The technique can also make chips susceptible to signal-to-noise problems, because it crams lots of information into signals. The result is increased efficiency – but it also means the information processed can be corrupted by all sorts of practical error sources such as fluctuations in voltages and currents.

"Computation signal-to-noise-ratio has been the main barrier for achieving all the benefits in-memory computing can offer," said Valavi.

The researchers addressed this performance problem by opting for a type of computing that uses capacitors, rather than transistors, to perform computation. Capacitors, which are devices that store electrical charge, offer several advantages. They can be manufactured with an extremely high degree of precision in modern micro-chip technologies, which is important in circuit design, and they are not affected greatly by changes in voltage or temperature. Capacitors also take up relatively little space - Princeton's in-memory computing chip places them on top of the memory cells, so they don't take up space beyond the cells. This further reduces the chip's data communication costs by placing capacitors inside memory components. This setup slims down the amount of area the electrical signals conveying data must cross, thereby delivering high processing speeds and lower energy.

"We end up with very precise circuits and these capacitors don't take up any extra area on the chip," said Verma.

The Princeton team put their system through its paces on several standard benchmark tests. These included identifying numbers scrawled by human hands, a task complicated by our tremendous range of handwriting styles, from punctilious to kindergarten-sloppy. A similar task involved parsing street-view house numbers, which likewise vary wildly in shape, form, picture clarity, orientation, and so on. In a third test, the chip-augmented neural network went about recognizing everyday objects such as cats, dogs, birds, cars, airplanes, ships, and so on.

The researchers tested their design against others currently available. In one, they measured the number of computational operations the chip could perform in one second. In real-life, this kind of throughput evaluation equates to how long someone has to wait before a piece of hardware, such as a cell phone, spits out a final answer. The Princeton chip performed 9.4 trillion binary operations per second.

The test results are encouraging but the researchers said the chip will need further work before it can be incorporated into electronic devices. Its architecture will need to be made programmable and compatible with other bits of hardware, including central processing units, the control centers of computers. After that, the software infrastructure must be built out so artificial intelligence designers can create new apps that leverage the chip's potentially breakthrough performance.

Naresh Shanbhag, a professor of electrical and computer engineering at the University of Illinois Urbana-Champaign who was not involved in the Princeton study, believes this potential is eminently realizable. "The technical challenges [the chip] faces in a commercial setting are eminently surmountable via standard engineering best practices," Shanbhag said.

Shanbhag further commented on the chip's applications. "This work

opens up new application domains for artificial intelligence systems," he said, specifying "energy- and latency-constrained computing platforms, such as autonomous vehicles and robots, as well as various sensor-rich Internet-of-Things devices."

The researchers look forward to taking the in-memory computing chip to a higher level of technological readiness.

"The next step is to take this very high efficiency and high computational throughput and make it accessible to a broad range of applications," said Verma. "The chip's major drawback is that it uses a very disruptive architecture. That needs to be reconciled with the massive amount of infrastructure and design methodology we have and use today, in practice."

More information: A Mixed-Signal Binarized Convolutional-Neural-Network Accelerator Integrating Dense Weight Storage and Multiplication for Reduced Data Movement. www.princeton.edu/~nverma/Verm … tlerVerma_VLSI18.pdf

Provided by Princeton University

Citation: Chip ramps up artificial intelligence systems' performance (2018, September 24) retrieved 10 April 2024 from https://phys.org/news/2018-09-chip-ramps-artificial-intelligence.html