Energy-friendly chip can perform powerful artificial-intelligence tasks

February 3 2016, by Larry Hardesty

MIT researchers have designed a new chip to implement neural networks. It is 10 times as efficient as a mobile GPU, so it could enable mobile devices to run powerful artificial-intelligence algorithms locally, rather than uploading data to the Internet for processing. Credit: MIT News

In recent years, some of the most exciting advances in artificial

intelligence have come courtesy of convolutional neural networks, large virtual networks of simple information-processing units, which are loosely modeled on the anatomy of the human brain.

Neural networks are typically implemented using graphics processing units (GPUs), special-purpose graphics chips found in all computing devices with screens. A mobile GPU, of the type found in a cell phone, might have almost 200 cores, or processing units, making it well suited to simulating a network of distributed processors.

At the International Solid State Circuits Conference in San Francisco this week, MIT researchers presented a new chip designed specifically to implement neural networks. It is 10 times as efficient as a mobile GPU, so it could enable mobile devices to run powerful artificial-intelligence algorithms locally, rather than uploading data to the Internet for processing.

Neural nets were widely studied in the early days of artificial-intelligence research, but by the 1970s, they'd fallen out of favor. In the past decade, however, they've enjoyed a revival, under the name "deep learning."

"Deep learning is useful for many applications, such as object recognition, speech, face detection," says Vivienne Sze, an assistant professor of electrical engineering at MIT whose group developed the new chip. "Right now, the networks are pretty complex and are mostly run on high-power GPUs. You can imagine that if you can bring that functionality to your cell phone or embedded devices, you could still operate even if you don't have a Wi-Fi connection. You might also want to process locally for privacy reasons. Processing it on your phone also avoids any transmission latency, so that you can react much faster for certain applications."

The new chip, which the researchers dubbed "Eyeriss," could also help usher in the "Internet of things"—the idea that vehicles, appliances, civil-engineering structures, manufacturing equipment, and even livestock would have sensors that report information directly to networked servers, aiding with maintenance and task coordination. With powerful artificial-intelligence algorithms on board, networked devices could make important decisions locally, entrusting only their conclusions, rather than raw personal data, to the Internet. And, of course, onboard neural networks would be useful to battery-powered autonomous robots.

Division of labor

A neural network is typically organized into layers, and each layer contains a large number of processing nodes. Data come in and are divided up among the nodes in the bottom layer. Each node manipulates the data it receives and passes the results on to nodes in the next layer, which manipulate the data they receive and pass on the results, and so on. The output of the final layer yields the solution to some computational problem.

In a convolutional neural net, many nodes in each layer process the same data in different ways. The networks can thus swell to enormous proportions. Although they outperform more conventional algorithms on many visual-processing tasks, they require much greater computational resources.

The particular manipulations performed by each node in a neural net are the result of a training process, in which the network tries to find correlations between raw data and labels applied to it by human annotators. With a chip like the one developed by the MIT researchers, a trained network could simply be exported to a mobile device.

This application imposes design constraints on the researchers. On one

hand, the way to lower the chip's power consumption and increase its efficiency is to make each processing unit as simple as possible; on the other hand, the chip has to be flexible enough to implement different types of networks tailored to different tasks.

Sze and her colleagues—Yu-Hsin Chen, a graduate student in electrical engineering and computer science and first author on the conference paper; Joel Emer, a professor of the practice in MIT's Department of Electrical Engineering and Computer Science, and a senior distinguished research scientist at the chip manufacturer NVidia, and, with Sze, one of the project's two principal investigators; and Tushar Krishna, who was a postdoc with the Singapore-MIT Alliance for Research and Technology when the work was done and is now an assistant professor of computer and electrical engineering at Georgia Tech—settled on a chip with 168 cores, roughly as many as a mobile GPU has.

Act locally

The key to Eyeriss's efficiency is to minimize the frequency with which cores need to exchange data with distant memory banks, an operation that consumes a good deal of time and energy. Whereas many of the cores in a GPU share a single, large memory bank, each of the Eyeriss cores has its own memory. Moreover, the chip has a circuit that compresses data before sending it to individual cores.

Each core is also able to communicate directly with its immediate neighbors, so that if they need to share data, they don't have to route it through main memory. This is essential in a convolutional neural network, in which so many nodes are processing the same data.

The final key to the chip's efficiency is special-purpose circuitry that allocates tasks across cores. In its local memory, a core needs to store not only the data manipulated by the nodes it's simulating but data describing

the nodes themselves. The allocation circuit can be reconfigured for different types of networks, automatically distributing both types of data across cores in a way that maximizes the amount of work that each of them can do before fetching more data from main memory.

At the conference, the MIT researchers used Eyeriss to implement a neural network that performs an image-recognition task, the first time that a state-of-the-art neural network has been demonstrated on a custom chip.

This story is republished courtesy of MIT News (web.mit.edu/newsoffice/), a popular site that covers news about MIT research, innovation and teaching.

Provided by Massachusetts Institute of Technology