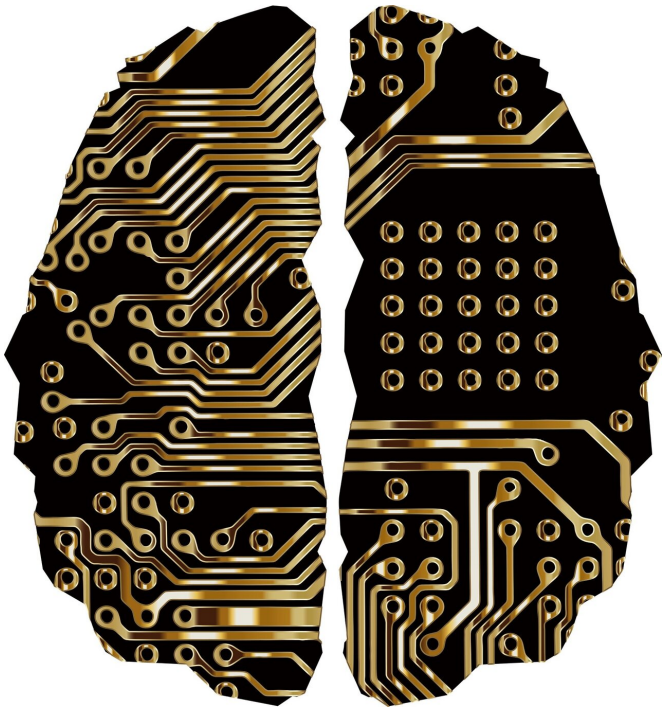
# Engineers offer smart, timely ideas for AI bottlenecks

11 June 2020, by Mike Williams



Credit: Pixabay/CC0 Public Domain

Rice University researchers have demonstrated methods for both designing innovative data-centric computing hardware and co-designing hardware with machine-learning algorithms that together can improve energy efficiency by as much as two orders of magnitude.

Advances in machine learning, the form of artificial intelligence behind self-driving cars and many other high-tech applications, have ushered in a new era of computing—the data-centric era—and are forcing engineers to rethink aspects of computing architecture that have gone mostly unchallenged for 75 years.

"The problem is that for large-scale deep neural networks, which are state-of-the-art for machine learning today, more than 90% of the electricity needed to run the entire system is consumed in moving data between the memory and processor," said Yingyan Lin, an assistant professor of electrical and computer engineering.

Lin and collaborators proposed two complementary methods for optimizing data-centric processing, both of which were presented June 3 at the International Symposium on Computer Architecture (ISCA), one of the premier conferences for new ideas and research in computer architecture.

The drive for data-centric architecture is related to a problem called the von Neumann bottleneck, an inefficiency that stems from the separation of memory and processing in the computing architecture that has reigned supreme since mathematician John von Neumann invented it in 1945. By separating memory from programs and data, von Neumann architecture allows a single computer to be incredibly versatile; depending upon which stored program is loaded from its memory, a computer can be used to make a video call, prepare a spreadsheet or simulate the weather on Mars.

But separating memory from processing also means that even simple operations, like adding 2 plus 2, require the computer's processor to access the memory multiple times. This memory bottleneck is made worse by massive operations in deep neural networks, systems that learn to make humanlike decisions by "studying" large numbers of previous examples. The larger the network, the more difficult the task it can master, and the more examples the network is shown, the better it performs. Deep neural network training can require banks of specialized processors that run around the clock for more than a week. Performing tasks based on the learned networks—a process known as inference—on a smartphone can drain its battery in less than an hour.

"It has been commonly recognized that for the data-centric algorithms of the machine-learning era, we need innovative data-centric hardware architecture," said Lin, the director of Rice's Efficient and Intelligent Computing (EIC) Lab. "But what is the optimal hardware architecture for machine learning?

"There are no one-for-all answers, as different applications require machine-learning algorithms that might differ a lot in terms of algorithm structure and complexity, while having different task accuracy and resource consumption—like energy cost, latency and throughput—tradeoff requirements," she said. "Many researchers are working on this, and big companies like Intel, IBM and Google all have their own designs."

One of the presentations from Lin's group at ISCA 2020 offered results on TIMELY, an innovative architecture she and her students developed for "processing in-memory" (PIM), a non-von Neumann approach that brings processing into memory arrays. A promising PIM platform is "resistive random access memory" (ReRAM), a nonvolatile memory similar to flash. While other ReRAM PIM accelerator architectures have been proposed, Lin said experiments run on more than 10 deep neural network models found TIMELY was 18 times more energy efficient and delivered more than 30 times the computational density of the most competitive state-of-the-art ReRAM PIM accelerator.

TIMELY, which stands for "Time-domain, In-Memory Execution, LocalitY," achieves its performance by eliminating major contributors to inefficiency that arise from both frequent access to the main memory for handling intermediate input and output and the interface between local and main memories.

In the main memory, data is stored digitally, but it must be converted to analog when it is brought into the local memory for processing in-memory. In prior ReRAM PIM accelerators, the resulting values are converted from analog to digital and sent back to the main memory. If they are called from the main memory to local ReRAM for subsequent operations, they are converted to analog yet again, and so on.

TIMELY avoids paying overhead for both unnecessary accesses to the main memory and interfacing data conversions by using analog-format buffers within the local memory. In this way, TIMELY mostly keeps the required data within local memory arrays, greatly enhancing efficiency.

The group's second proposal at ISCA 2020 was for SmartExchange, a design that marries algorithmic and accelerator hardware innovations to save energy.

"It can cost about 200 times more energy to access the main memory—the DRAM—than to perform a computation, so the key idea for SmartExchange is enforcing structures within the algorithm that allow us to trade higher-cost memory for much-lower-cost computation," Lin said.

"For example, let's say our algorithm has 1,000 parameters," she added. "In a conventional approach, we will store all the 1,000 in DRAM and access as needed for computation. With SmartExchange, we search to find some structure within this 1,000. We then need to only store 10, because if we know the relationship between these 10 and the remaining 990, we can compute any of the 990 rather than calling them up from DRAM.

"We call these 10 the 'basis' subset, and the idea is to store these locally, close to the processor to avoid or aggressively reduce having to pay costs for accessing DRAM," she said.

The researchers used the SmartExchange algorithm and their custom hardware accelerator to experiment on seven benchmark deep neural network models and three benchmark datasets. They found the combination reduced latency by as much as 19 times compared to state-of-the-art deep neural network accelerators.

Provided by Rice University

APA citation: Engineers offer smart, timely ideas for AI bottlenecks (2020, June 11) retrieved 21 October 2021 from https://techxplore.com/news/2020-06-smart-ideas-ai-bottlenecks.html