Neural networks everywhere | MIT News

Most recent advances in artificial-intelligence systems such as speech- or face-recognition programs have come courtesy of neural networks, densely interconnected meshes of simple information processors that learn to perform tasks by analyzing huge sets of training data.

But neural nets are large, and their computations are energy intensive, so they're not very practical for handheld devices. Most smartphone apps that rely on neural nets simply upload data to internet servers, which process it and send the results back to the phone.

Now, MIT researchers have developed a special-purpose chip that increases the speed of neural-network computations by three to seven times over its predecessors, while reducing power consumption 94 to 95 percent. That could make it practical to run neural networks locally on smartphones or even to embed them in household appliances.

"The general processor model is that there is a memory in some part of the chip, and there is a processor in another part of the chip, and you move the data back and forth between them when you do these computations," says Avishek Biswas, an MIT graduate student in electrical engineering and computer science, who led the new chip's development.

"Since these machine-learning algorithms need so many computations, this transferring back and forth of data is the dominant portion of the energy consumption. But the computation these algorithms do can be simplified to one specific operation, called the dot product. Our approach was, can we implement this dot-product functionality inside the memory so that you don't need to transfer this data back and forth?"

Biswas and his thesis advisor, Anantha Chandrakasan, dean of MIT's School of Engineering and the Vannevar Bush Professor of Electrical Engineering and Computer Science, describe the new chip in a paper that Biswas is presenting this week at the International Solid State Circuits Conference.

Back to analog

Neural networks are typically arranged into layers. A single processing node in one layer of the network will generally receive data from several nodes in the layer below and pass data to several nodes in the layer above. Each connection between nodes has its own "weight," which indicates how large a role the output of one node will play in the computation performed by the next. Training the network is a matter of setting those weights.

A node receiving data from multiple nodes in the layer below will multiply each input by the weight of the corresponding connection and sum the results. That operation — the summation of multiplications — is the definition of a dot product. If the dot product exceeds some threshold value, the node will transmit it to nodes in the next layer, over connections with their own weights.

A neural net is an abstraction: The "nodes" are just weights stored in a computer's memory. Calculating a dot product usually involves fetching a weight from memory, fetching the associated data item, multiplying the two, storing the result somewhere, and then repeating the operation for every input to a node. Given that a neural net will have thousands or even millions of nodes, that's a lot of data to move around.

But that sequence of operations is just a digital approximation of what happens in the brain, where signals traveling along multiple neurons meet at a "synapse," or a gap between bundles of neurons. The neurons' firing rates and the electrochemical signals that cross the synapse correspond to the data values and weights. The MIT researchers' new chip improves efficiency by replicating the brain more faithfully.

In the chip, a node's input values are converted into electrical voltages and then multiplied by the appropriate weights. Summing the products is simply a matter of combining the voltages. Only the combined voltages are converted back into a digital representation and stored for further processing.

The chip can thus calculate dot products for multiple nodes — 16 at a time, in the prototype — in a single step, instead of shuttling between a processor and memory for every computation.

All or nothing

One of the keys to the system is that all the weights are either 1 or -1. That means that they can be implemented within the memory itself as simple switches that either close a circuit or leave it open. Recent theoretical work suggests that neural nets trained with only two weights should lose little accuracy — somewhere between 1 and 2 percent.

Biswas and Chandrakasan's research bears that prediction out. In experiments, they ran the full implementation of a neural network on a conventional computer and the binary-weight equivalent on their chip. Their chip's results were generally within 2 to 3 percent of the conventional network's.

"This is a promising real-world demonstration of SRAM-based in-memory analog computing for deep-learning applications," says Dario Gil, vice president of artificial intelligence at IBM. "The results show impressive specifications for the energy-efficient implementation of convolution operations with memory arrays. It certainly will open the possibility to employ more complex convolutional neural networks for image and video classifications in IoT [the internet of things] in the future."