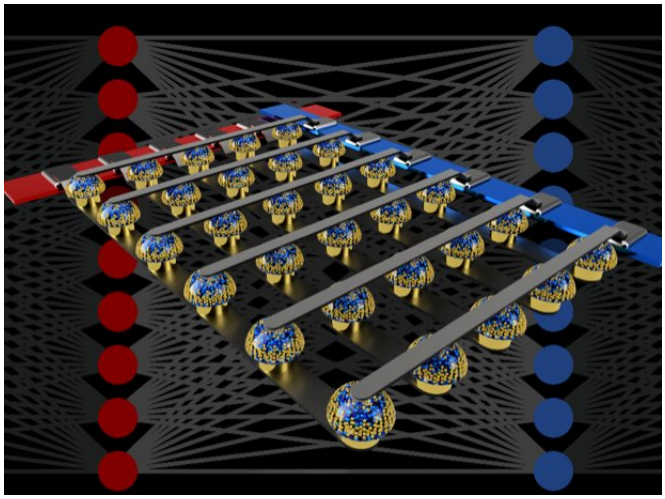
The future of AI needs hardware accelerators based on analog memory devices

14 June 2018, by Stefano Ambrogio



Crossbar arrays of non-volatile memories can accelerate the training of fully connected neural networks by performing computation at the location of the data. Credit: IBM

Imagine personalized Artificial Intelligence (AI), where your smartphone becomes more like an intelligent assistant – recognizing your voice even in a noisy room, understanding the context of different social situations or presenting only the information that's truly relevant to you, plucked out of the flood of data that arrives every day. Such capabilities might soon be within our reach – but getting there will require fast, powerful, energy-efficient AI hardware accelerators.

In a recent paper published in *Nature*, our IBM Research AI team demonstrated deep neural network (DNN) training with large arrays of analog memory devices at the same accuracy as a Graphical Processing Unit (GPU)-based system. We believe this is a major step on the path to the kind of hardware accelerators necessary for the next AI breakthroughs. Why? Because delivering the Future of AI will require vastly expanding the scale of AI calculations.

DNNs must get larger and faster, both in the cloud and at the edge – and this means energy-efficiency must improve dramatically. While better GPUs or other digital accelerators can help to some extent, such systems unavoidably spend a lot of time and energy moving data from memory to processing and back. We can improve both speed and energy-efficiency by performing AI calculations in the analog domain with right at the location of the data – but this only makes sense to do if the resulting neural networks are just as smart as those implemented with conventional digital hardware.

Analog techniques, involving continuously variable signals rather than binary 0s and 1s, have inherent limits on their precision—which is why modern computers are generally digital computers. However, AI researchers have begun to realize that their DNN models still work well even when digital precision is reduced to levels that would be far too low for almost any other computer application. Thus, for DNNs, it's possible that maybe analog computation could also work.

However, until now, no one had conclusively proven that such analog approaches could do the same job as today's software running on conventional digital hardware. That is, can DNNs really be trained to equivalently high accuracies with these techniques? There is little point to being faster or more energy-efficient in training a DNN if the resulting classification accuracies are always going to be unacceptably low.

In our paper, we describe how analog non-volatile memories (NVM) can efficiently accelerate the "backpropagation" algorithm at the heart of many

recent AI advances. These memories allow the "multiply-accumulate" operations used throughout these algorithms to be parallelized in the analog domain, at the location of weight data, using underlying physics. Instead of large circuits to multiply and add digital numbers together, we simply pass a small current through a resistor into a wire, and then connect many such wires together to let the currents build up. This lets us perform many calculations at the same time, rather than one after the other. And instead of shipping digital data on long journeys between digital memory chips and processing chips, we can perform all the computation inside the analog memory chip.

However, due to various imperfections inherent to today's analog memory devices, previous demonstrations of DNN training performed directly on large arrays of real NVM devices failed to achieve classification accuracies that matched those of software-trained networks.

By combining long-term storage in phase-change memory (PCM) devices, near-linear update of conventional Complementary Metal-Oxide Semiconductor (CMOS) capacitors and novel techniques for cancelling out device-to-device variability, we finessed these imperfections and achieved software-equivalent DNN accuracies on a variety of different networks. These experiments used a mixed hardware-software approach, combining software simulations of system elements that are easy to model accurately (such as CMOS devices) together with full hardware implementation of the PCM devices. It was essential to use real analog memory devices for every weight in our neural networks, because modeling approaches for such novel devices frequently fail to capture the full range of device-to-device variability they can exhibit.

Using this approach, we verified that full chips should indeed offer equivalent accuracy, and thus do the same job as a digital accelerator – but faster and at lower power. Given these encouraging results, we've already started exploring the design of prototype hardware accelerator chips, as part of an IBM Research Frontiers Institute project.

From these early design efforts we were able to provide, as part of our Nature paper, initial estimates for the potential of such NVM-based chips for training fully-connected layers, in terms of the computational energy efficiency (28,065 GOP/sec/W) and throughput-per-area (3.6 TOP/sec/mm2). These values exceed the specifications of today's GPUs by two orders of magnitude. Furthermore, fully-connected layers are a type of neural network layer for which actual GPU performance frequently falls well below the rated specifications.

This paper indicates that our NVM-based approach can deliver software-equivalent training accuracies as well as orders of magnitude improvement in acceleration and energy efficiency despite the imperfections of existing analog memory devices. The next steps will be to demonstrate this same software equivalence on larger networks calling for large, fully-connected layers – such as the recurrently-connected Long Short Term Memory (LSTM) and Gated Recurrent Unit (GRU) networks behind recent advances in machine translation, captioning and text analytics – and to design, implement and refine these analog techniques on prototype NVM-based hardware accelerators. New and better forms of analog memory, optimized for this application, could help further improve both areal density and energy efficiency.

More information: Stefano Ambrogio et al. Equivalent-accuracy accelerated neural-network training using analogue memory, *Nature* (2018). DOI: 10.1038/s41586-018-0180-5

Provided by IBM