(Invited) Novel Technologies for Artificial Intelligence: prospects and challenges

Stefano Ambrogio, Pritish Narayanan, Hsinyu Tsai, Charles Mackin, An Chen,
Bob M. Shelby, Geoffrey W. Burr
IBM Research-Almaden, 650 Harry Road, 95120 San Jose, USA

ABSTRACT

Artificial Intelligence (AI) exploiting bio-inspired algorithms such as Spike-Timing-Dependent-Plasticity (STDP) and back-propagation algorithms as found in Deep Neural Networks (DNNs) is able to perform accurate classification of large amounts of data. However, to further proceed in the development of AI, novel hardware technologies supporting fast calculation should be developed. Recently, many algorithms have been efficiently mapped into arrays of Non-Volatile Memories, such as Phase-Change Memory (PCM) or Resistive Memory (RRAM) [1,2].

In this invited talk, we provide a summary of recent progress in hardware acceleration of AI algorithms, such as the training of Fully Connected (FC) DNNs based on large arrays of PCMs. In such schemes, crossbar arrays of weights encoded as conductances and shown to provide orders of magnitude increases in speed and energy efficiency with respect to current state of the art CPUs and GPUs [2].

In addition to speed and power consumption, the desired chip for FC DNNs training should provide equivalent accuracy to software training. We recently demonstrated a novel weight scheme based on PCM and CMOS circuitry able to obtain software-equivalent training accuracy on MNIST and other small and medium size datasets. Results were obtained with mixed hardware-software experiments in which CMOS circuitry was simulated and PCM behavior was measured experimentally using actual device arrays [3].

After this, we provide some design guidelines for the implementation of a multicore chip able to perform training of DNNs. This is obtained using many NVM arrays connected through routing circuitry to efficiently distribute internal signals, external inputs such as images and labels and, finally, to provide trained weights to the output [4].

REFERENCES

[1] G. W. Burr et al., “Experimental demonstration and tolerancing of a large-scale neural network (165,000 synapses) using phase-change memory as the synaptic weight element”, IEEE Trans. Elec. Dev, 62(11), pp. 3498 (2015).

[2] G. W. Burr et al., “Large-scale neural networks implemented with non-volatile memory as the synaptic weight element: Comparative performance analysis (accuracy, speed, and power)”, IEDM Tech. Digest, 4.4 (2015).

[3] S. Ambrogio et al., “Equivalent-Accuracy Accelerated Neural Network Training using Analog Memory”, Nature 558 (7708), 60 (2018).

[4] P. Narayanan et al., “Toward on-chip acceleration of the backpropagation algorithm using nonvolatile memory”, IBM J. Res. Dev., 61 (4), 1-11 (2017).