Extended Abstracts of the 2017 International Conference on Solid State Devices and Materials, Sendai, 2017, pp587-588

M-2-03 (Invited)

Low Power Deep Neural Network Hardware Based on Memristive Crossbar Circuits

Irina Kataeva and Shigeki Ohtsuka

Advanced Research and Innovation Center, DENSO CORPORATION
500-1 Minamiyama, Komenoki-cho, Nisshin-shi, 470-0011 Japan
Phone: +81-561-75-1885 E-mail: irina_kataeva@denso.co.jp

Abstract

We discuss our motivation behind development of memristive crossbar based Deep Neural Network hardware for automotive applications and briefly review experimental demonstrations of perceptron circuits and our progress in development of efficient training algorithms.

1. Introduction

The field of artificial intelligence is experiencing a new renascence with Deep Neural Networks algorithms consistently outperforming other machine learning approached by a wide margin [1]-[4]. Though offering unprecedented performance DNNs are very computationally expensive and require powerful and power-hungry parallel hardware such as GPUs. This approach is suitable for cloud-based applications, but it is not acceptable for automotive applications in which life-critical safety systems, e.g. object detection and recognition, have to be implemented in-car and operate robustly in real time without failure due to communication speed or computing resource availability. Furthermore, the hardware have to be low power to guarantee fuel and/or battery efficiency and minimize environmental impact of CO_2 exhaust. Thus low power and high speed hardware is crucial for automotive applications.

A lot of effort is currently being put into development of mobile GPUs and digital accelerators for DNNs [5]-[10]. An alternative approach is to develop specialized hardware that utilizes neural network's potential for low-power and high-speed information processing [11], [12]. The majority of such efforts rely on conventional technology, such as CMOS circuits to implement artificial neural networks [13]-[19].

Emerging memory device technologies [20], while not yet mature for large scale implementations, could offer further improvements in performance in the future [21]. One of the most promising proposals utilizing emerging memory technologies is hybrid CMOS-crossbar circuits [12] with integrated two- or three-terminal resistive switching (memristive) devices [22], [23], which can be implemented using phase change memories, magnetic tunnel junctions, ferroelectric memories, solid state electrolyte, or metal oxide resistive switching devices [24]-[29].

Our focus is on the latter in particular, as metal oxide devices such as TiO_{2-x} offer analog memory functionality with up to 100 memory states [30] and small footprint due to no use of selector devices [31]. Furthermore, our estimates show that CMOS-nanocrossbar circuits based on metal oxide memristors can offer up to two orders of magnitude improvement in computational efficiency over conventional CMOS technology (Fig. 1).

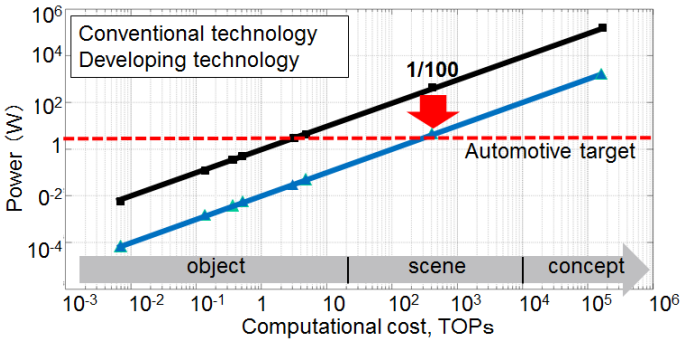

Fig. 1 Deep Neural Networks computational cost vs. power required for real-time implementation.

2. Artificial neural networks based on memristive crossbar circuits

Typical CMOS-crossbar implementation of neural networks is realized by integrating memristive devices into crossbar circuits to implement analog weights (synapses) and combining it with CMOS circuitry that implements neuron functionality and other peripheral functions (Fig. 2).

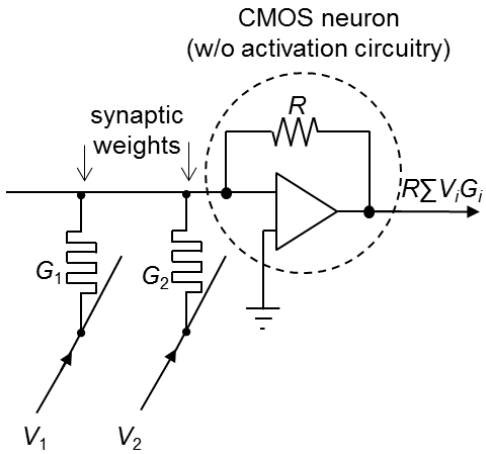

Fig. 2 Electrical circuit of a CMOS-memristive crossbar implementation of a neural network.

Inputs and neurons outputs are coded by voltages V, while synaptic weights – by memristive conductance G. The resulting currents $I = GV$ are injected into the common wire and summed up according to Kirchhoff's law before being converted to voltage using operational amplifier. As a result, memristive crossbars enable compact analog implementation of vector-matrix multiplication, core computation of artificial neural networks and Deep Neural Networks, directly in-memory.

3. Experimental results

Single layer perceptron was the first experimental demonstration of a memristive crossbar based neural network [31]. A 12x12 memristive crossbar was used to implement synaptic weights and perform analog vector-matrix multiplication with all neuron functionality emulated using measurements set-up.

The work was later expanded to experimentally demonstrate a multilayer perceptron with one hidden layer using two 20x20 memristive crossbars with neuron circuits integrated on a printed circuit board [32].

4. Training neural networks based on memristive crossbar circuits

Training of the neural networks based on memristive crossbar circuits comes with challenges specific to physics of memristive devices and crossbar topology. The challenges include the need to program selected memristive devices, which exhibit highly non-linear behavior [30], in practical amount of time with certain precision and without disturbing the memory states of other devices on the same crossbar lines. We have considered two hardware implementations of a backpropagation algorithm, ex-situ and in-situ.

In ex-situ training, the weights are trained in a software network and then imported into crossbar conductances by programming individual devices using, for example, feedback tuning algorithm [33]. The advantage is that any training algorithm can be implemented in software and tuning analog memory requires minimal peripheral circuitry. The disadvantage is that a relatively crude precision of analog memory has to be taken into account during training. Most importantly, fabrication defects such as stuck-on-close and stuck-on-open devices can make it impossible to import desired weight values resulting in DNN performance degradation [34], [35].

The alternative is in-situ training approach that relies on implementing conductance adjustments directly in hardware. We have adapted backpropagation to memristive crossbar circuits as a short series of pulses with variable amplitude and duration to train conductances, two pulses per crossbar line and four pulses for the whole crossbar for batch and stochastic training, respectively [35].

We have benchmarked both ex-situ and in-situ training algorithms on a variety of DNN architectures and data sets (MNIST, GTSRB, CIFAR-10) and demonstrated the scalability of memristive crossbar circuits and recognition performance comparable to that of conventional software-only implementations of DNNs [34], [35].

3. Conclusions

We have briefly reviewed the need for low power hardware for Deep Neural Networks for automotive applications, the motivation behind our development of memristive crossbar circuits and various approaches. We have also introduced the experimental demonstrations of single layer and multilayer perceptrons and training algorithms we have developed.

References

[1] A. Krizhevsky et al., *Advances in Neural Information Processing Systems (2012)* 1097.
[2] V. Mnih et al. Nature, **518, 7540** (2015) 529.
[3] A. Hannun at al., arXiv:1412.5567, 2014.
[4] C. Szegedy et al., arXiv:1409.4842, 2014.
[5] http://www.nvidia.com/object/tegra.html.
[6] http://www.nvidia.com/object/drive-px.html.
[7] I. Sato et al., GTC2014.
[8] S. Han et al., *Proceedings of the 43rd International Symposium on Computer Architecture*. (2016)
[9] Y. Chen et al., *Proceedings of the 47th Annual IEEE/ACM International Symposium on Microarchitecture* (2014).
[10] Y. Chen et al., IEEE Journal of Solid-State Circuits **52 1** (2017) 127.
[11] C. Mead, Analog VLSI and Neural Systems (1989).
[12] K. K. Likharev, Sci. Adv. Mater. **3** (2011) 322.
[13] S. K. Kim at al., *IEEE Conf. Field Programmable Logic and Applications (2009)* 367.
[14] J. Kim at al., J. Solid-State Circuits **45 1** (2010) 32.
[15] S. Chakradhar at al., SIGARCH Comput. Archit. News **38** (2010) 247.
[16] B. V. Benjamin at al., Proceedings of the IEEE (2014) 699.
[17] P. Merolla at al., *IEEE Custom Integr. Circuits Conf.* (2011) 1.
[18] S. Ramakrishnan and J. Hasler IEEE Trans. Very Large Scale Integr. Syst., **22** (2014) 353.
[19] J. Lu at al., IEEE Int. Solid-State Circuits Conf. Dig. Tech. Papers (2014).
[20] ITRS 2013 ed. available online at http://www.itrs.net/.
[21] D. Strukov, Nature **476** (2011).
[22] R. Williams, IEEE Spectr. **45** (2008) 28.
[23] D. Strukov and H. Kohlstedt, MRS Bulletin (2012) **37**.
[24] D. Kuzum at al., Nano Letters, **12 5** (2012) 2179.
[25] N. Locatelli et al., Nature materials **13 1** (2014) 11.
[26] Y. Kaneko at al., IEEE Transactions on Electron Devices **61 8** (2014) 2827.
[27] F. Alibart et al., Nature Comm. **4** (2013) 2072.
[28] S. H. Jo et al., Nano Letters **10 4** (2010) 1297.
[29] T. Ohno at al., Nature materials **10 8** (2011) 591.
[30] F. Merrikh Bayat at al., Applied Physics A (2015) 1.
[31] M. Prezioso et al., Nature **521 7550** (2015) 61.
[32] F. Merrikh-Bayat et al., in print ICCAD'2017 (2017).
[33] F. Alibart et al. Nanotechnology **23 7** (2012) 075201.
[34] M. Prezioso et al., *IEEE International Electron Devices Meeting (2015)*.
[35] I. Kataeva et al., *IEEE International Joint Conference on Neural Networks (2015)*.