RRAM Based Learning Acceleration

Yu Wang, Lixue Xia, Ming Cheng, Tianqi Tang, Boxun Li, Huazhong Yang
Dept. of E.E., Tsinghua National Laboratory for Information Science and Technology (TNList),
Tsinghua University, Beijing, China
e-mail: yu-wang@mail.tsinghua.edu.cn

1. INTRODUCTION

Deep Learning (DL) is becoming popular in a wide range of domains. Many emerging applications, ranging from image and speech recognition to natural language processing and information retrieval, rely heavily on deep learning techniques, especially the Neural Networks (NNs). NNs have led to great advances in recognition accuracy compared with other traditional methods in recent years. NN-based methods demand much more computation and memory resource, and therefore a number of NN accelerators have been proposed on CMOS-based platforms, such as FPGA and GPU [1]. However, it becomes more and more difficult to obtain substantial power efficiency and gains directly through the scaling down of traditional CMOS technique. Meanwhile, the large data amount in DL applications also meets an ever-increasing "memory wall" challenge because of the efficiency of von Neumann architecture. Consequently, there is a growing research interest of exploring emerging nano-devices and new computing architectures to further improve power efficiency [2].

The emerging metal-oxide resistive random-access memory (RRAM) device provides promising solutions to boost the energy efficiency [3] of NNs. The RRAM crossbar structure is able to perform analog matrix-vector multiplication. In this way, the computation is processed just in the memory without high-cost data transportation, which breaks the memory wall bottleneck and achieve high energy efficiency.

2. RELATED WORK

2.1 Challenge And Its Method

However, there are some challenges limiting the efficiency of RRAM-based neural computing system and some methods were proposed to ease and solve them.

2.1.1 Algorithm Mapping

Except the matrix-vector multiplication, there are several peripheral functions in NNs, such as the non-linear neural function in Deep Neural Networks (DNNs) and the pooling function in Convolutional Neural Networks (CNNs). These functions are hard to be implemented in the RRAM crossbar. Moreover, the size of the RRAM crossbar is limited by the IR-drop problem, which means multiple crossbars need to be connected to perform the computation of large matrix. Therefore, complete circuit design is required to practically mapping different NNs into RRAM-based circuits. We propose several RRAM-based solutions for different NN algorithms. First, to verify the efficiency of the RRAM-based structure, a programmable the RRAM-based approximate computing unit (RRAM-ACU) is introduced to accelerate numeric computation and a scalable approximate computing framework is proposed on top of the RRAM-ACU [3]. The results show that the RRAM-ACU achieves 10.264̃91.02× speedup and power efficiency of 24.595̃67.98 GFLOPS/W. Second, we builded a SNN-based energy efficient system for real time classification with RRAM devices [4, 5]. Simulation results illustrate that the system achieves 91.2% accuracy on the MNIST dataset with an ultra-low power consumption of 3.5mW. Third, for the popular CNN algorithm, we implement the main function, namely the Convolution kernels, also in RRAM crossbars and propose the RRAM-based CNN accelerator structure [4, 5]. The results show that RRAM-based design can obtain more than 40× energy efficiency gains compared with the best FPGA and GPU implementations. Forth, we introduce a mixed-signal training acceleration framework, which realizes the self-training of RRAM-based neural network [6].

2.1.2 Interface

As analog device, RRAM-based structure processes the computation with analog signals. Therefore, the Analog-to-Digital Converters (ADCs) and Digital-to-Analog Converters (DACs) are required to work as the interfaces between RRAM-crossbar and peripheral digital modules. However, the high-precision interfaces cost much more energy and area than RRAM crossbars, which becomes a new bottleneck of RRAM-based design. For small scale network, a MEI structure is proposed to directly learn the relationship between the binary 0/1 arrays and eliminate the interfaces [7]. For large scale networks, we propose an energy efficient SEI structure for RRAM-based CNN that reduces the ADC cost for merging results of multiple crossbars [8]. Both these two designs can save more than 80% area and energy consumptions compared with original ADC/DAC-based structure.

2.1.3 Non-ideal Factors of Devices

There are several kinds of non-ideal factors impacting the availability and efficiency of RRAM-based system, such as the nonlinear V-I characteristic, fabrication variations, defects, endurance problem, resistance drifts, etc. For static factors, we analyze the impact of both device level and circuit level non-ideal factors, including the nonlinear current-voltage relationship of RRAM devices, the variation of device fabrication and write operation, and the interconnect resistance as well as other crossbar array parameters [9]. On top of that, we propose a technological exploration flow for device parameter configuration to overcome the impact of non-ideal factors [10]. The proposed technological exploration flow is able to achieve accuracy improvement and energy saving simultaneously. For dynamic drifts, we propose an inline calibration mechanism to guarantee the computation quality [11], which achieves a calibration efficiency of 91.18%.

2.1.4 EDA Tools

Traditional simulators and Electronic Design Automation (EDA) tools cannot support the simulation and optimization of emerging RRAM-based structure. We develop the first behavior-level simulation platform for RRAM-based neural computing system named MNSIM [12]. MNSIM proposes a general hierarchical structure for RRAM-based neural computing system, and proposes behavior-level models to accelerate the simulation. Experimental results show that MNSIM achieves more than 7000 times speed-up compared with SPICE and obtains reasonable accura-

CASES '16, October 01-07, 2016, Pittsburgh, PA, USA
© 2016 ACM. ISBN 978-1-4503-4482-1/16/10. . . $15.00
DOI: http://dx.doi.org/10.1145/2968455.2981124

cy. Yiran Chen's group proposes AutoNCS ÍC an EDA framework that can automate the NCS designs that combine memristor crossbars and discrete synapse modules. Based on the previous research on RRAM-based Non-volatile Memory (NVM), Yuan Xie's group proposes a reconfigurable processing-in-memory architecture for neural computing applications and improves NN performance by 1800 times compared with the state-of-the-art neural processing unit on large neural networks with only 6.4% area overhead on RRAM chips [12]. Yu Cao's group focuses on the device-level optimization. Based on the strong knowledge and test result of practical RRAM devices, they discuss the design challenges on scaling up the array size due to non-ideal device properties and array parasitics, and propose circuit-level mitigation strategies to minimize the learning accuracy loss in a large array [13].

2.2 RRAM-based Reinforcement Learning

Supervised learning in neural network has obtained huge success in many fields, such as computer vision, pattern recognition and speech recognition, etc. However, supervised learning requires examples provided by a knowledgable external supervisor. For interactive problem, such as game theory, it is impractical to obtain examples of desired behavior that are both correct and representative of all situations. Reinforcement learning has demonstrated its powerful capacity in interactive problem, such as game theory and multi-agent systems. Neural network based on reinforcement learning helps it perform much better than human experts in some games,such as AlphaGo. It is designed based on neural network of reinforcement learning, which beated the world champion of Go and astonished human. However, the huge power consumption of Alphago seriouly hinders its practical application, which consumes more than 1202 CPUs and 176 GPUs when competing with Lee Sedol [14]. In recent years, the emerging device technologies offer great potentials of efficient hardware implementation of reinforcement learning and enable totally different computational paradigms. The metal-oxide resistive random access memory (RRAM) device (or the memristor) is one of these promising devices. Most importantly, RRAM and its crossbar structure can realize the matrix-vector multiplication with ultra-high efficient energy, which transfers time complexity from $O(n^2)$ to $O(1)$.

The high energy efficiency of the RRAM crossbar provides a potential to boost the energy efficiency of reinforcement learning. However, there are some challenges to realize reinforcement learning on RRAM. First, in the original reinforcement learning algorithm, there exists a copy operation between two neural networks [15]. It means that we need two crossbars to realize conductances copy between them. But accurately writing cell of the crossbar to a target conductance is difficult because of the stochastic characteristic of RRAM. For the challenge of copy operation, we propose a new architecture based on RRAM to overcome it without any modification of the original algorithm. Besides, for the training phase of neural network, calculation of the weight variation needs lots of high precision registers and computing units, such as the multiplier. In order to reduce the cost of training phase in RRAM, the stochastic gradient algorithm is modified into a style that is more adaptive for RRAM without loss of convergence.

3. REFERENCES

[1] J. Qiu, J. Wang, S. Yao, K. Guo, B. Li, E. Zhou, J. Yu, T. Tang, N. Xu, S. Song *et al.*, "Going deeper with embedded fpga platform for convolutional neural network," in *Proceedings of the 2016 ACM/SIGDA International Symposium on Field-Programmable Gate Arrays*. ACM, 2016, pp. 26–35.

[2] Y. Wang, B. Li, R. Luo, Y. Chen, N. Xu, and H. Yang, "Energy efficient neural networks for big data analytics," in *Proc. Design, Automation and Test in Europe Conference and Exhibition (DATE)*. IEEE, 2014, pp. 1–2.

[3] B. Li *et al.*, "Memristor-based approximated computation," in *ISLPED*, 2013, pp. 242–247.

[4] Y. Wang, T. Tang, L. Xia, B. Li, P. Gu, H. Yang, H. Li, and Y. Xie, "Energy efficient RRAM spiking neural network for real time classification," in *Proc. the 25th edition on Great Lakes Symposium on VLSI*. ACM, 2015, pp. 189–194.

[5] T. Tang, R. Luo, B. Li, H. Li, Y. Wang, and H. Yang, "Energy efficient spiking neural network design with rram devices," in *Proc. 14th International Symposium on Integrated Circuits (ISIC)*. IEEE, 2014, pp. 268–271.

[6] B. Li, Y. Wang, Y. Wang, Y. Chen, and H. Yang, "Training itself: Mixed-signal training acceleration for memristor-based neural network," in *Proc. 19th Asia and South Pacific Design Automation Conference (ASP-DAC)*. IEEE, 2014, pp. 361–366.

[7] B. Li, L. Xia, P. Gu, Y. Wang, and H. Yang, "Merging the interface: Power, area and accuracy co-optimization for rram crossbar-based mixed-signal computing system," in *Proceedings of the 52nd Annual Design Automation Conference*. ACM, 2015, p. 13.

[8] L. Xia, T. Tang, W. Huangfu, M. Cheng, X. Yin, B. Li, Y. Wang, and H. Yang, "Switched by input: Power efficient structure for rrambased convolutional neural network," *2016 53nd ACM/EDAC/IEEE Design Automation Conference (DAC)*.

[9] P. Gu, B. Li, T. Tang, S. Yu, Y. Cao, Y. Wang, and H. Yang, "Technological exploration of rram crossbar array for matrix-vector multiplication," *The 20th Asia and South Pacific Design Automation Conference*.

[10] L. Xia, P. Gu, B. Li, T. Tang, X. Yin, W. Huangfu, S. Yu, Y. Cao, Y. Wang, and H. Yang, "Technological exploration of rram crossbar array for matrix-vector multiplication," *Journal of Computer Science and Technology*.

[11] B. Li, Y. Wang, Y. Chen, H. H. Li, and H. Yang, "Ice: inline calibration for memristor crossbar-based computing engine," *Proceedings of the conference on Design, Automation & Test in Europe*, 2014.

[12] L. Xia, B. Li, T. Tang, P. Gu12, X. Yin, W. Huangfu, P.-Y. Chen, S. Yu, Y. Cao, Y. Wang *et al.*, "Mnsim: A simulation platform for memristor-based neuromorphic computing system," *Proceedings of the conference on Design, Automation & Test in Europe*, 2016.

[13] S. Yu, P.-Y. Chen, Y. Cao, L. Xia, Y. Wang, and H. Wu, "Scaling-up resistive synaptic arrays for neuro-inspired architecture: Challenges and prospect," *2015 IEEE International Electron Devices Meeting (IEDM)*.

[14] D. Silver, A. Huang, C. J. Maddison, A. Guez, L. Sifre, G. Van Den Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot *et al.*, "Mastering the game of go with deep neural networks and tree search," *Nature*, vol. 529, no. 7587, pp. 484–489, 2016.

[15] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski *et al.*, "Human-level control through deep reinforcement learning," *Nature*, vol. 518, no. 7540, pp. 529–533, 2015.