A Simulation Study on the System Performance of Neural Networks using Embedded Nonvolatile Memory

IIS, University of Tokyo. °Paul Davin Johansen, Masaharu Kobayashi
Email: johansenp@nano.iis.u-tokyo.ac.jp

I. Introduction

Hardware neural networks (HNNs) can provide faster training/inference latency and reduced power consumption than neural networks implemented via software [1]. While neural networks have been studied extensively, as of yet, there has been little research on how device configuration and variability can affect performance metrics for neural networks on a system level. In this simulation study, several system performance metrics such as accuracy, latency, and energy consumption for HNNs were analyzed.

II. Simulation Methods

NeuroSim simulation software was used to simulate both analog and digital emerging non-volatile memory (eNVM) at various bit precision and conductance state values for key figures of merit such as accuracy, latency, and energy [2]. The simulation was performed using the default NeuroSim 3-layer neural network consisting of a 20x20 node input, 100 node hidden layer, and 10 node output with feedforward and back propagation. This arrangement was used for processing the well-known MNIST dataset over just a 10-epoch period until the accuracy had converged sufficiently.

III. Results and Discussion

In Fig. 1 and 2, the accuracy of each neural network was simulated for analog eNVM and digital eNVM by adjusting the number of conductance states and bit precision, respectively. We found that analog eNVM requires more than 100 states while digital eNVM requires at least 2 bits to achieve >80% accuracy. At this minimum number of conductance states and bit precision, the latency and energy for read and write operations were estimated and summarized in Fig. 3. This data shows that while the analog eNVM device depicts a possible lower energy consumption and improved read latency over digital eNVM with the given device parameters, there is a considerable total write latency (training time). Therefore, if the neural network requires frequent training, analog eNVM is not suitable for such applications. If inference is performed after only training once, then analog eNVM is suitable. HNNs do already provide advantages in terms of power consumption over traditional neural network simulated through software from an architectural standpoint [1]. But it is important to select an appropriate device configuration for the desired application.

IV. Conclusion

The impact of device configuration (analog and digital eNVM) on the metrics of HNN was investigated through a NeuroSim simulation study. Analog eNVM has a benefit of lower energy operation than its digital counterpart. However, the long write latency of analog eNVM may inhibit its usage in applications which require frequent training. Analog eNVM will be suitable for inference-only applications when compared to digital eNVM. The impact of device variability for eNVM will be illustrated in further detail in the upcoming presentation.

V. References

[1] J. Misra and I. Saha, "Artificial neural networks in hardware: A survey of two decades of progress", *Neurocomputing*, vol. 74, no. 1-3, pp. 239-255, 2010.

[2] P. Chen, X. Peng and S. Yu, "NeuroSim: A Circuit-Level Macro Model for Benchmarking Neuro-Inspired Architectures in Online Learning", *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 37, no. 12, pp. 3067-3080, 2018.

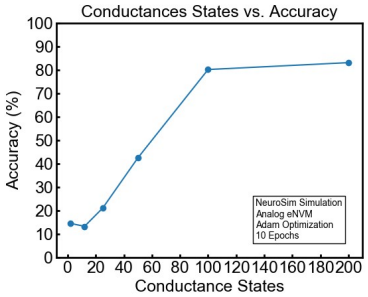

Fig. 1. Simulated recognition accuracy versus conductance states of analog eNVM devices. 100 conductance states are required for >80% accuracy.

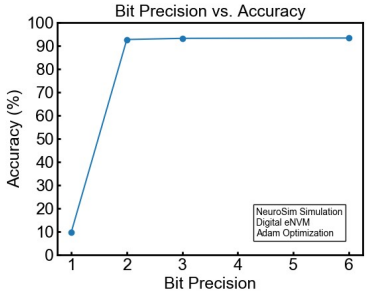

Fig. 2. Simulated recognition accuracy versus bit precision of digital eNVM devices. 2 bits are required for >80% accuracy.

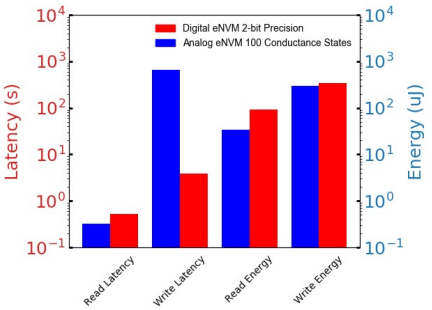

Fig. 3. A comparison of read/write latency/energy for select 2-bit digital eNVM (red) and 100 conductance state analog eNVM (blue).