SAND2019-2330C

Designing and Modeling Analog Neural Network Training Accelerators

Sapan Agarwal[1], Robin B. Jacobs-Gedrim[2], Christopher Bennett[2], Alex Hsia[2], Michael S. Van Heukelom[2], David Hughart[2], Elliot Fuller[1], Yiyang Li[1], A. Alec Talin[1], Matthew J. Marinella[2]
[1]Sandia National Laboratories, Livermore, CA
[2]Sandia National Laboratories, Albuquerque, NM
E-mail: sagarwa@sandia.gov

Abstract

Analog crossbars have the potential to reduce the energy and latency required to train a neural network by three orders of magnitude when compared to an optimized digital ASIC. The crossbar simulator, CrossSim, can be used to model device nonidealities and determine what device properties are needed to create an accurate neural network accelerator. Experimentally measured device statistics are used to simulate neural network training accuracy and compare different classes of devices including TaOx ReRAM, $Li_{1-x}Co_xO_2$ devices, and conventional floating gate SONOS memories. A technique called "Periodic Carry" can overcomes device non-idealities by using a positional number system while maintaining the benefit of parallel analog matrix operations.

Introduction

Training a neural network is dominated by data movement. Even in an optimized digital ASIC accelerator data must constantly be moved between local caches and the computational unit. Analog crossbars can eliminate most of this data movement and can potentially reduce the energy and latency required to train a neural network by three orders of magnitude when compared to an optimized digital ASIC[2]. They accelerate three key operations that are the bulk of the computation in a neural network: vector matrix multiplication (VMM), matrix vector multiplication (MVM), and outer product rank 1 updates (OPU) [3]. For each operation, the computations are performed in a single parallel step in memory. Thus, for an NxN array, the CV^2 and I×V energy scale as the array size, $O(N^2)$ [3]. This is O(N) better than trying to read or write a digital memory. Each row of an NxN digital memory array must be accessed sequentially, resulting in N columns of length O(N) being charged N times, requiring $O(N^3)$ energy to read a digital memory.

Unfortunately, analog devices are noisy and suffer from several non-idealities including read noise, write noise and write nonlinearity. Analog arrays suffer from parasitic voltage drops. Furthermore, analog systems tend to have limited bit precision on the inputs and outputs to a crossbar, with the fewer bits used, the faster and more energy efficient an analog system is. All of these issues will impact the final classification accuracy of a neural network. To compensate for these issues and take advantage of large gains in energy and latency enabled by analog systems, neural algorithms will need to be designed specifically to overcome the hardware limitations. This will require new co-design tools where the impact of device level properties on algorithmic performance can be assessed so that analog device development can be driven by algorithmic requirements. Consequently, we have developed a new open source simulation tool called CrossSim [4] to quantify the impact of device level properties on algorithmic performance.

Required Device Properties

In [1] we use CrossSim to evaluate the impact of read noise, write noise and write nonlinearity on training a two layer neural network for MNIST (recognizing handwritten digits). Training or classifying with analog devices that have a read noise standard deviation (σ) up to 5% of the total conductance range does not significantly degrade the accuracy (~1%). Neural network training requires a smaller write noise with σ<0.4% of the weight range. Nevertheless, this can still be 3X larger than a typical update as training a neural network requires small updates on the order of 0.1% of the weight range. This will vary slightly depending on the dataset and the neural network architecture. Even a slightly asymmetric write nonlinearity substantially degrades classification accuracy, as shown in Fig 1. To compute in a large energy efficient crossbar, resistive memories must also have a high on-state resistance. Given that scaled wires at a 10nm half pitch can only handle 10 μA before electromigration occurs, reading 100 devices in parallel already sets a limit of 100 nA per-device current draw. Assuming a 1V read, this suggests a minimum on-state resistance of 10 MΩ. Parasitic voltage drops also become an issue for higher currents or larger arrays.

Evaluating Experimental Devices

To understand how different types of analog devices perform in a training accelerator, we compare three different analog devices: a TaOx ReRAM [5], a battery inspired $Li_{1-x}Co_xO_2$ device[6], and a conventional floating gate SONOS (Silicon-oxygen-nitrogen-oxygen-silicon) memory[7]. The respective structures are illustrated in Fig 2. The analog write noise statistics and write nonlinearity are measured and directly used by CrossSim to simulate the accuracy of a neural network training accelerator built on those devices as illustrated in Fig 3. Both $Li_{1-x}Co_xO_2$ and SONOS can train to high accuracy, while the TaOx device is limited to an accuracy of ~80%.

Periodic Carry

In order to compensate for the remaining device non-idealities a technique called periodic carry [8] can be used. Multiple devices can be used to represent a weight with a positional number system, such as base 2 or base 10, exponentially increases the number of levels with the number of devices. However, this is not compatible with a parallel write as carries need to be performed between digits. This can be overcome by allowing devices to store extra levels and periodically (every 100-1000 updates) reading the device and performing any necessary carries. This allows noisy, nonlinear TaO_x devices that previously trained to 80% accuracy on MNIST, to achieve 97% accuracy, only 1% away from the ideal numeric accuracy of 98%. In addition, both the SONOS and $Li_{1-x}Co_xO_2$ devices can achieve ideal accuracy using periodic carry.

Architecture Comparison

To understand the potential advantages of accelerators built on different devices, we compare kernel level energy, latency and area for 4 accelerator architectures[2, 7]: digital SRAM, digital ReRAM, analog ReRAM and analog SONOS. The energy and latency advantages strongly depend on the bit precision of the accelerator.

Analog ReRAM has the greatest possible advantages over a digital SRAM based ASIC of 11X in area, 430X in energy and 34X in latency. ReRAM based memories are also starting to be integrated in commercial foundries but are too noisy and nonlinear to train to high accuracies. Single device per weight accelerators are limited in accuracy to around 80%, necessitating the use of techniques like periodic carry to help compensate for poor device properties. As a nearer term option, SONOS devices are currently available in commercial foundries but typically require long μs to ms write pulses and high voltages around 10V to program. Nevertheless, SONOS

Sandia National Laboratories is a multimission laboratory managed and operated by National Technology & Engineering Solutions of Sandia, LLC, a wholly owned subsidiary of Honeywell International Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.

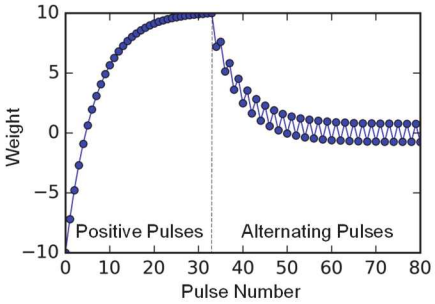

Fig. 1: (from [4]) Applying identical alternating positive and negative pulses causes the weight to decay towards a center value when it should remain constant. When the weight is near the maximum, a positive pulse does not change the weight much, but a negative pulse significantly decreases it. The opposite holds for weights near the minimum weight.

still has area, energy and latency advantages of 4X, 120X and 2X respectively over a digital ASIC. $Li_{1-x}Co_xO_2$ devices also have the potential to be highly accurate and efficient but additional research is needed in fundamental device physics and process integration before their full potential is realized.

ACKNOWLEDGEMENTS

This work was supported by the Department of the Defense, Defense Threat Reduction Agency, under Grant HDTRA1-17-1-0038, the Department of Energy (DOE) Advanced Manufacturing Office and the Laboratory Directed Research and Development program at Sandia National Laboratories, a multimission laboratory managed and operated by National Technology and Engineering Solutions of Sandia, LLC., a wholly owned subsidiary of Honeywell International, Inc., for the U.S. DOE's National Nuclear Security Administration under contract DE-NA-0003525. This paper describes objective technical results and analysis. Any subjective opinions do not necessarily represent the views of the U.S. DOE or the US Govt.

REFERENCES

[1] S. Agarwal *et al.*, "Resistive memory device requirements for a neural algorithm accelerator," in *2016 International Joint Conference on Neural Networks (IJCNN)*, 2016, pp. 929-938.

[2] M. J. Marinella *et al.*, "Multiscale Co-Design Analysis of Energy, Latency, Area, and Accuracy of a ReRAM Analog Neural Training Accelerator," *IEEE Journal on Emerging and Selected Topics in Circuits and Systems,* 2018.

[3] S. Agarwal *et al.*, "Energy Scaling Advantages of Resistive Memory Crossbar Based Computation and its Application to Sparse Coding," *Frontiers in Neuroscience,* vol. 9, p. 484, 2016, Art. no. 484.

[4] S. Agarwal *et al.* (2017). *CrossSim*. Available: http://cross-sim.sandia.gov

[5] R. B. Jacobs-Gedrim *et al.*, "Impact of Linearity and Write Noise of Analog Resistive Memory Devices in a Neural Algorithm Accelerator," presented at the IEEE International Conference on Rebooting Computing (ICRC) Washington, DC, November 2017.

[6] E. J. Fuller *et al.*, "Li-Ion Synaptic Transistor for Low Power Analog Computing," *Advanced Materials,* vol. 29, no. 4, p. 1604310, 2017.

[7] S. Agarwal *et al.*, "Using Floating Gate Memory to Train Ideal Accuracy Neural Networks," *IEEE Journal of Exploratory Solid-State Computational Devices and Circuits,* 2019.

[8] S. Agarwal *et al.*, "Achieving ideal accuracies in analog neuromorphic computing using periodic carry," in *VLSI Technology, 2017 Symposium on*, 2017, pp. T174-T175: IEEE.

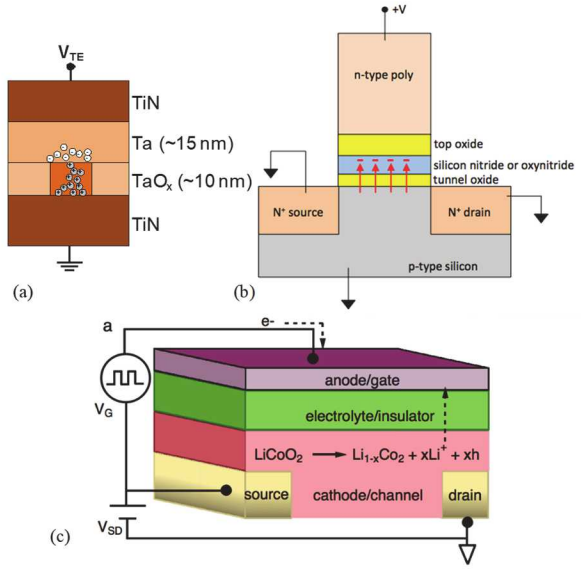

Fig 2: we compare three different analog devices: (a) TaOx ReRAM [5], (b) conventional floating gate SONOS (Silicon-oxygen-nitrogen-oygen-silicon) memory[7] and (c) battery inspired $Li_{1-x}CoO_2$ devices[6].

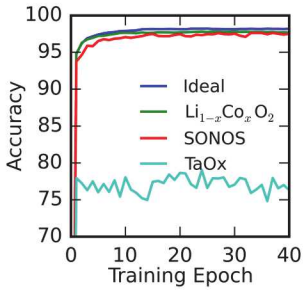

Fig 3: Neural accelerators based on SONOS or $Li_{1-x}Co_xO_2$ devices can reach near ideal accuracies while accelerators based on TaOx can reach around 80% accuracy on MNIST.

TABLE I
AREA COMPARISONS

	8 bit	**4 bit**	**2 bit**
Digital SRAM (μm^2)	836,000	814,000	800,000
Digital ReRAM (μm^2)	137,000	114,000	101,000
Analog ReRAM (μm^2)	75,000	46,000	41,000
Analog SONOS (μm^2)	195,000	166,000	161,000

TABLE II
ENERGY AND LATENCY COMPARISONS

	VMM			MVM			OPU			Total		
	8 bit	4 bit	2 bit	8 bit	4 bit	2 bit	8 bit	4 bit	2 bit	8 bit	4 bit	2 bit
Energy – Digital SRAM (nJ)	2850	2237	1848	4855	4241	3852	4300	3673	3274	12,000	10,150	8974
Energy – Digital ReRAM (nJ)	2139	1502	1098	2139	1502	1098	3246	2572	2143	7525	5577	4339
Energy – Analog ReRAM (nJ)	12.8	1.00	0.44	12.8	1.00	0.44	2.2	1.00	0.46	27.9	2.66	1.35
Energy – Analog SONOS (nJ)	14.4	2.25	1.5	14.4	2.25	1.5	71.5	30.9	10.6	100	35.4	13.6
Latency – Digital SRAM (µs)	4	4	4	32	32	32	8	8	8	44	44	44
Latency – Digital ReRAM (µs)	176	176	176	176	176	176	340	340	340	692	692	692
Latency – Analog ReRAM (µs)	0.384	0.024	0.011	0.384	0.024	0.011	0.512	0.032	0.032	1.28	0.080	0.054
Latency – Analog SONOS (µs)	0.402	0.032	0.014	0.402	0.032	0.014	20	20	20	20.80	20.06	20.02