MaxNVM: Maximizing DNN Storage Density and Inference Efficiency with Sparse Encoding and Error Mitigation

Lillian Pentecost, Marco Donato, Brandon Reagen,
Udit Gupta, Siming Ma, Gu-Yeon Wei, David Brooks
Harvard University
Cambridge, MA

December 9, 2019

Abstract

Deeply embedded applications require low-power, low-cost hardware that fits within stringent area constraints. Deep learning has many potential uses in these domains, but introduces significant inefficiencies stemming from off-chip DRAM accesses of model weights. Ideally, models would fit entirely on-chip. However, even with compression, memory requirements for state-of-the-art models make on-chip inference impractical. Due to increased density, emerging eNVMs are one promising solution.

MaxNVM [8] is a principled co-design of sparse encodings, protective logic, and fault-prone MLC eNVM technologies (i.e., RRAM and CTT) to enable highly-efficient DNN inference. We find bit reduction techniques (e.g., clustering and sparse compression) *increase* weight vulnerability to faults. This limits the capabilities of MLC eNVM. To circumvent this limitation, we improve storage density with minimal overhead using protective logic. Tradeoffs between density and reliability result in a rich design space. We show that by balancing these techniques, the weights of large DNNs are able to reasonably fit on-chip. Compared to a naive, single-level-cell eNVM solution, our highly-optimized MLC memories reduce weight area by up to 29×. We compare our technique against NVDLA, a state-of-the-art industry-grade CNN accelerator, and demonstrate up to 3.2× reduced power and up to 7.5× reduced energy per ResNet50 inference depending on input frame rate.

1 Introduction

DNNs are in use everywhere from self-driving cars to wireless sensor nodes and implanted medical devices [5, 3, 1, 7, 4]. For state-of-the-art DNN hardware accelerators, fetching weights from DRAM is a main performance and energy bottleneck. Ideally, DNNs weights would be stored entirely on-chip, but the capacity requirements are unrealistic for SRAM storage.

Emerging embedded non-volatile memory (eNVM) technologies are one promising solution for eliminating DRAM inefficiencies. eNVMs provide high-capacity, low read-latency storage and can be significantly denser than SRAM via aggressive Multi-level Cell (MLC) designs.

MaxNVM demonstrates that MLC eNVMs can be used for highly-efficient DNN inference through rigorous co-design. For example, in considering fault-prone MLC eNVMs and sparse-encoded weights, we find a tension between the two: sparse encoding increases fault vulnerability, limiting the efficacy of MLCs. To reduce overall memory footprint, we first sparse encode weights to save raw bits, then set the levels-per-cell to the highest configuration without accuracy loss. To further increase storage density, we use protective logic. We consider IndexSynchronization, a proposed fault mitigation technique, and ECC. With judicious use, the total number of required memory cells to store DNN weights decreases by up to 22% with our proposed technique, and ECC overhead is never more than 1% of total DNN storage. Optimal MLC designs provide up to 29× area reduction relative to SLC eNVM. Additionally, this work proposes and evaluates eNVM-based memory systems for NVDLA [10], an industry-grade CNN accelerator. Using our co-design approach, DNN weights can fit on-chip, eliminating the need for DRAM. Compared to the baseline NVDLA implementation, MLC eNVMs enable entirely on-chip ResNet50 inference in about 2mm^2.

2 Evaluation Methodology

Our proposed, principled co-design incorporates optimizations and techniques at algorithmic and architectural levels. After developing a fault model based on technology-specific device characteristics and SPICE models of sensing circuitry, we use a previously-validated fault injection framework to quantify the impact of faults on DNN accuracy [9]. We leverage well-known tools to model the energy, performance, and area of our proposals [2, 6]. The interaction of these methods as they contribute to the final evaluation is summarized in Figure 1.

3 Fault Tolerance of Sparse Encodings for DNN Inference

Model optimizations and sparse storage schemes significantly impact DNN fault tolerance. We quantify the impact of different encoding strategies on DNN classification error, which guides us in incorporating error correction and mitigation techniques in order to maximize the effectiveness of MLC eNVM storage for DNN inference.

3.1 Index Synchronization

We propose a novel, light-weight error mitigation technique for bitmask-based sparse encoding methods which

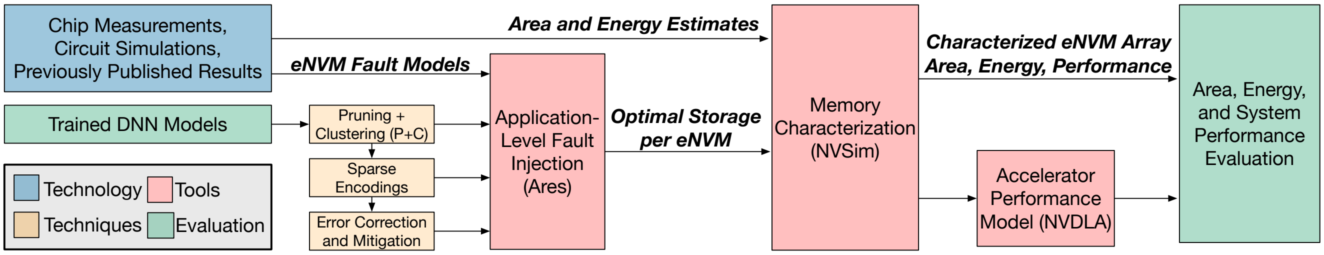
Chip Measurements, Circuit Simulations, Previously Published Results
Area and Energy Estimates
eNVM Fault Models
Trained DNN Models
Pruning + Clustering (P+C)
Sparse Encodings
Error Correction and Mitigation
Technology
Tools
Techniques
Evaluation
Application-Level Fault Injection (Ares)
Optimal Storage per eNVM
Memory Characterization (NVSim)
Characterized eNVM Array Area, Energy, Performance
Accelerator Performance Model (NVDLA)
Area, Energy, and System Performance Evaluation

Figure 1: Summary of the tools, optimizations, and intermediate results used in final system evaluations.

we call Index Synchronization (IdxSync). The intuition behind this technique is to leverage the inherent fault tolerance of the DNN weight values in contrast with the vulnerability of sparse encoding metadata (e.g., a bitmask representing which data values in the original weight matrix are non-zero).

4 Benefits of Non-Volatility

We propose a completely self-contained inference accelerator that stores all of the weights in on-chip eNVM and does not require external DRAM, while the baseline NVDLA relies on LPDDR4 DRAM for all weight storage. Depending on how frequently inferences occur (i.e., required frame rate for an image processing task), eNVMs have an inherent relative benefit by virtue of not needing to reload weight values when powered on or, alternatively, keeping DRAM powered to avoid this cost. Thus, optimized MLC eNVM solutions are particularly compelling for applications with lower frame-rate requirements, and these benefits would be exaggerated for systems with less frequent wake-ups.

5 Potential Impact

In addition to demonstrating compelling potential benefits of MLC eNVMs for DNN inference, MaxNVM exposes and explores a rigorous co-design of eNVM fault characteristics with architectural and algorithmic choices in order to maximize storage density and efficiency.

References

[1] G. Desoli, N. Chawla, T. Boesch, S. p. Singh, E. Guidetti, F. De Ambroggi, T. Majo, P. Zambotti, M. Ayodhyawasi, H. Singh, and N. Aggarwal. 14.1 a 2.9tops/w deep convolutional neural network soc in fd-soi 28nm for intelligent embedded systems. In *2017 IEEE International Solid-State Circuits Conference (ISSCC)*, pages 238–239, Feb 2017.

[2] X. Dong, C. Xu, Y. Xie, and N. P. Jouppi. Nvsim: A circuit-level performance, energy, and area model for emerging nonvolatile memory. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, 31(7):994–1007, July 2012.

[3] Nicholas D. Lane, Sourav Bhattacharya, Petko Georgiev, Claudio Forlivesi, and Fahim Kawsar. An early resource characterization of deep learning on wearables, smartphones and internet-of-things devices. In *Proceedings of the 2015 International Workshop on Internet of Things Towards Applications*, IoT-App '15, pages 7–12, New York, NY, USA, 2015. ACM.

[4] K. H. Lee, S. Y. Kung, and N. Verma. Improving kernel-energy trade-offs for machine learning in implantable and wearable biomedical applications. In *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1597–1600, May 2011.

[5] NASA. Autonomous car facts 2016. keynote: Autonomous car a new driver for resilient computing and design-for-test. 2016.

[6] NVIDIA. Nvidia deep learning accelerator (nvdla): a free and open architecture that promotes a standard way to design deep learning inference accelerators. 2017.

[7] S. Park, S. Choi, J. Lee, M. Kim, J. Park, and H. J. Yoo. 14.1 a 126.1mw real-time natural ui/ux processor with embedded deep-learning core for low-power smart glasses. In *2016 IEEE International Solid-State Circuits Conference (ISSCC)*, pages 254–255, Jan 2016.

[8] L. Pentecost, M. Donato, B. Reagen, U. Gupta, S. Ma, G. Wei, and D. Brooks. Maxnvm: Maximizing dnn storage density and inference efficiency with sparse encoding and error mitigation https://dl.acm.org/citation.cfm?id=3358258. *52nd IEEE/ACM International Symposium on Microarchitecture (MICRO 2019)*, October 2019.

[9] Brandon Reagen, Lillian Pentecost, Udit Gupta, Paul Whatmough, Sae Kyu Lee, Niamh Mulholland, David Brooks, and Gu-Yeon Wei. Ares: A framework for quantifying the resilience of deep neural networks. In *2018 The 55th Annual Design Automation Conference (DAC)*, June 2018.

[10] F Sijstermans. The nvidia deep learning accelerator. In *Hot Chips*, 2018.