03

Integrating sparsity and quantization for model compression

Towards energy-efficient AI models

Abstract

The increasing performance of AI models has led to growing computational cost and energy consumption, motivating efficiency techniques such as sparsity and quantization. However, conventional approaches often rely on their combined use, which can restrict deployment to digital hardware that efficiently supports sparse computation. In this work, we present a training framework that does not require learned sparsity, while remaining compatible with learned quantization. This enables neural networks to be flexibly deployed across a wider range of hardware platforms, including energy-efficient analog devices where sparsity is difficult to exploit. By decoupling model efficiency from specific structural constraints, our approach broadens the applicability of model compression. Ultimately, this work aims to support the development of energy-efficient, reconfigurable AI systems that can operate across diverse computational substrates, from digital to emerging analog hardware.

Neural Network (NN) Efficiency

■ NNs require large computation and memory
➡ **Pruning** and **quantization** are key for efficiency

Precise and Dense NN — All connections, full-precision ➡ computation and memory costs are high — e.g., 32bit x 9 weights

Pruning — **Eliminate unnecessary connections** ➡ smaller cost — e.g., 32bit x 5 weights

Quantization — **Reduce numerical precision** ➡smaller cost — e.g., 3bit x 9 weights

Prun.＋Quant. — Simplify structure and values ➡smallest cost — e.g., 3bit x 5 weights

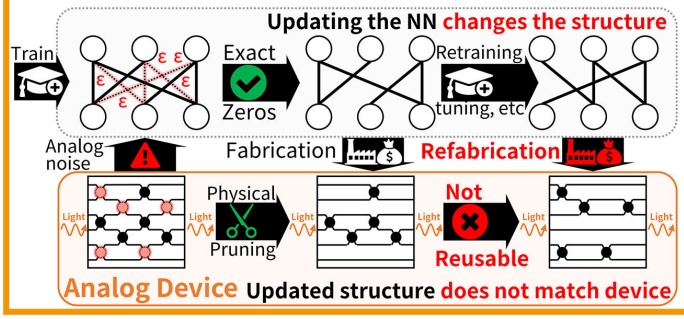
Sparse Computation on Analog Devices

■Pruning assumes **exact zeros in digital devices**, but fails in **analog devices** (e.g., optical) due to **noise**

Digital Computing — Pruned connections **represented with exact zeros** ➡ accurate

Analog Computing — Pruned connections **reactivated by noise** ➡ accuracy collapse

■ **Conventional approach: physical pruning at fabrication**
✓ Remove connections from topology → exact zeros
✓ No need to store zeros → higher compression
✓ Reduced hardware → lower area, noise, power
✗ Less reconfigurability → limited practicality



Our Research: Towards Reusable Analog AI

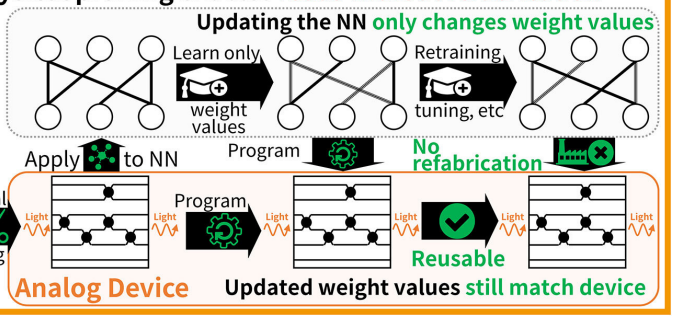
■ What to learn differs across hardware: digital prefers connections; analog prefers signs and magnitudes

Digital Computing
Weight values: **Fixed** — Predetermined (0bit)
AND
Structure: **Learned** — e.g., 1bit x 9 weights
Flip
Efficient and sparse NN
Sparsity **represented** with **exact zeros**

Analog Computing — Learn • Fix targets — flip
Weight values: **Learned** — e.g., 1bit x 5weights
learn signs
AND/OR
learn magnitudes — e.g., 3bit x 5 weights
AND
Structure: **Fixed** — Predetermined (0bit)
Efficient and sparse NN
Sparsity **implemented** with **physical connections**

Proposed: Learn weight signs and magnitudes over a fixed topology
★Quantization-aware training (QAT) under structural constraints
✓ Retains benefits of conventional method
✓ Structure unchanged → reusable
✓ No structure learning → can be determined by hardware topology optimization

■ Learning only **signs and magnitudes after applying physical pruning** enables reuse without refabrication



References

[1] Á. López García-Arias, Y. Okoshi, H. Otsuka, D. Chijiwa, Y. Fujiwara, S. Takeuchi, M. Motomura, "The Trichromatic Strong Lottery Ticket Hypothesis: Neural Compression With Three Primary Supermasks," *Workshop on Machine Learning and Compression, Conference on Neural Information Processing Systems (NeurIPS)*, 2024.

[2] Á. López García-Arias, Y. Okoshi, H. Otsuka, D. Chijiwa, Y. Fujiwara, S. Takeuchi, M. Motomura, "The Trichromatic Strong Lottery Ticket Hypothesis: A Unifying View of Supermask-Based Learning," *IEEE International Joint Conference on Neural Networks (IJCNN)*, 2026.

Contact

Ángel López García-Arias, Recognition Research Group, Media Information Laboratory