Time-Domain Neuron for RRAM-Based Computing-In-Memory Hardware with Quantization-Aware Training

Jeongmin Lee [(1)] and Byung-Geun Lee [(1)]

[(1)] Gwangju Institute of Science and Technology

RRAM-based computing-in-memory (CIM) hardware has emerged as a promising architecture for energy-efficient deep neural network (DNN) inference by performing matrix-vector multiplication (MVM) in the analog domain using crossbar arrays. However, the limited number of conductance states in RRAM devices constrains achievable weight precision, necessitating algorithm-hardware co-design based on quantized neural networks (QNNs). In particular, maintaining floating-point model accuracy under sub-4-bit weights and activations requires a design approach based on quantization-aware training (QAT) [1].

In conventional CIM architectures, the MVM output current is digitized by an analog-to-digital converter (ADC) before subsequent operations are performed, introducing an additional quantization stage absent during network training. The step size and full scale of the ADC are fixed independently of the parameters learned through QAT, leading to a fundamental algorithm-hardware mismatch [2].

To address this issue, we propose a time-domain neuron (TDN) circuit based on learned step size quantization (LSQ) adapted for CIM hardware, where LSQ-trained parameters are directly mapped to the current source of a voltage-to-time converter. This enables the step size and full scale in the time domain to match the learned values, significantly reducing the algorithm-hardware mismatch. The proposed TDN integrates dequantization and ReLU activation for the layer output, and re-quantization for the next layer into a single circuit, generating outputs that can be directly fed forward without additional post-processing.

To validate the proposed approach, we design a DNN accelerator consisting of three convolutional layers fully mapped onto 32×32 RRAM arrays and two fully connected layers implemented in software. Behavioral simulations show that the proposed TDN achieves 98.6% accuracy with 3-bit weights and 4-bit activations on the MNIST dataset, comparable to floating-point baseline performance. Moreover, accuracy degradation remains below 1% when post-layout simulation parameters, temperature variations, and device mismatch are incorporated, demonstrating the robustness of the proposed circuit.

[1] S. K. Esser et al., ICLR, 2020.

[2] B. Zhang et al., ICML, 2024.