Sparse ReRAM Engine: Joint Exploration of Activation and Weight Sparsity in Compressed Neural Networks

Corresponding author

Tzu-Hsien Yang, Hsiang-Yun Cheng, Chia-Lin Yang, I-Ching Tseng, Han-Wen Hu, Hung-Sheng Chang, Hsiang-Pang Li, National Taiwan University, Academic Sinica, Macronix International Co., Ltd.

Keywords

neural network, sparsity, ReRAM, accelerator architecture

Summary

*Challenge*

1. Existing architecture studies on ReRAM-based NN accelerators assume that an entire crossbar array can be activated in a single cycle. This is an over-idealized assumption. However, due to inference accuracy considerations, matrix-vector computation must be conducted in a smaller granularity in practice.
2. The coupled crossbar structure makes it difficult to efficiently exploit sparsity in ReRAM-based DNN accelerators.

*Contribution*

1. They study the design challenges of sparsity exploration on ReRAM-based DNN accelerators, and show that a practical OU-based design enables new opportunities to effectively exploit DNN sparsity.
2. Propose a sparse ReRAM engine (SRE) to jointly exploit weight and activation sparsity, with only minimal indexing overhead.
3. Compare SRE with the over-idealized ReRAM architecture, which overlooks the inference accuracy loss caused by the accumulated effect of per-cell current deviation.

*Innovation Points*

This is the first sparsity exploration work that targets a practical hardware design for ReRAM-based accelerators instead of an over-idealized architecture.

*Result*

Evaluation results show that, for neural networks that use the structural pruning algorithm to regularize weight sparsity during training, SRE provides up to 42.3x performance speedups (average 13.1x) and up to 95.4% energy savings (average 85.3%) over a baseline that does not exploit sparsity.