FloatPIM: in-memory acceleration of deep neural network training with high precision

Corresponding author

Mohsen Imani, University of California, San Diego

Keywords

DNN accelerator for training and inference, processing in memory architecture,

Summary

*Challenge*

Existing PIM architectures do not support **high precision computation**, e.g., in floating point precision, which is essential for training accurate CNN models. In addition, most of the existing PIM approaches **require analog/mixed-signal circuits**, which do not scale, exploiting insufficiently reliable multi-bit Non-Volatile Memory (NVM).

*Contribution*

1. Propose FloatPIM, a fully-digital scalable PIM architecture that accelerates CNN in both training and testing phase.
   1. FloatPIM natively supports floating-point representation, thus enabling accurate CNN training.
   2. .FloatPIM also enables fast communication between neighboring memory blocks to reduce internal data movement of the PIM architecture
2. We evaluate the efficiency of FloatPIM on ImageNet dataset using popular large-scale neural networks.

*Result*

Our evaluation shows that FloatPIM supporting floating point precision can achieve up to **5.1% higher classification accuracy** as compared to existing PIM architectures with limited fixed-point precision. FloatPIM **training** is on **average 303.2× and 48.6× (4.3× and 15.8×) faster and more energy efficient** as compared to GTX 1080 GPU (PipeLayer [1] PIM accelerator). For testing, FloatPIM also provides 324.8× and 297.9× (6.3× and 21.6×) speedup and energy efficiency as compared to GPU (ISAAC [2] PIM accelerator) respectively.