Zhao SF, Wang F, Liu B *et al*. LayCO: Achieving least lossy accuracy for most efficient RRAM-based deep neural network accelerator via layer-centric co-optimization. JOURNAL OF COMPUTER SCIENCE AND TECHNOLOGY 38(2): 328−347 Mar. 2023. DOI: 10.1007/s11390-023-2545-y

LayCO: Achieving Least Lossy Accuracy for Most Efficient RRAM-Based Deep Neural Network Accelerator via Layer-Centric Co-Optimization

Shao-Feng Zhao[1;4] (赵少锋), Student Member, CCF, Fang Wang[1;2] (王芳), Member, CCF, Bo Liu[3] (刘博), Dan Feng[1;2] (冯丹), Member, CCF, ACM, IEEE, and Yang Liu[4] (刘洋)

1 Wuhan National Laboratory for Optoelectronics, Key Laboratory of Information Storage System, Engineering Research Center of data storage systems and Technology (School of Computer Science & Technology, Huazhong University of Science & Technology), Ministry of Education of China
2 Research Institute of Huazhong University of science and technology in Shenzhen
3 School of Computer and Artificial Intelligence, Network Management Center, Zhengzhou University
4 Cloud Computing and Big Data Institute, Cyberspace Administration Center, Henan University of Economics and Law

Research Context

- Resistive random access memory (RRAM)
  - enables to operate massively parallel dot products and accumulations
- RRAM-based accelerator
  - effective approach to bridge the gap between Internet of Things devices’ constrained resources and DNNs’ tremendous cost
- Analog RRAM buffer
  - due to the huge overhead of A/D conversions and digital accumulations
  - offers potential solutions to A/D conversion issues

Research Objectives

- Our research aims to address:
  - the energy consumption in resource-constrained environments
  - the critical concerns over endurance
  - strictly provides an inference accuracy guarantee

Research Method

- A co-optimizing strategy
  - combining both voltage regulation and bit-width compaction
  - in a layer-centric fashion
- A data mapping and wear-aware data swapping method
  - designate RRAM partition to DNN
  - control the write balance of whole RRAM arrays

Research Results

- LayCO outperforms state-of-the-art designs in:
  - energy efficiency (27× over TIMELY-like configuration)
  - lifetime prolongation (308× over RAQ)
  - area reduction (6× over RAQ)
  - strictly ensuring a target DNN accuracy (accuracy loss less than 1%)

Research Conclusions

- Three key contributions on LayCO:
  - voltage regulating, bit-width tightening, and data mapping and swapping method
- Enlighten further research
  - Low-voltage approximate memory for error-tolerant deep learning workloads