ZHEYU YAN

Software-Hardware Co-Design for Neural Networks Acceleration

@ zyan2@nd.edu ☎ 574-210-6016 Notre Dame, IN sites.nd.edu/zheyu-yan/ zyyan

PROFESSIONAL SUMMARY

- Independent researcher in neural architecture search (NAS)-based software-hardware co-design.
- Proficient in compute-in-memory using non-volatile memory devices.

EXPERIENCES & PROJECTS

ASIC Engineer Intern

June 2022 – Sept. 2022 Amazon Lab 126

- Developed a model compression framework for Transformer models using knowledge distillation and quantization.
- Applied the compression framework to multiple Transformer models targeting multiple tasks including image classification, sentence matching and machine translations.
- Achieved up to 120x compression rate with moderate accuracy drop.

Device Variation Analysis for DNN Accelerators

July 2018 – Present University of Notre Dame

- Developed a special scheme to protect non-volatile compute-in-memory (nvCiM) DNN accelerators from device variations. Achieved 20x speedup while preserving the DNN accuracy by using write-verify to protect only a small portion of the weights using the knowledge from second-order derivative [1].
- Proposed and mathematically proved the efficacy of an attack method targeting nvCiM DNN accelerators. The proposed method reduces DNN accuracy from close to 100% to 0% with less than 1% change in weight value [2].

Memory Efficient Differentiable NAS

May 2020 – May 2021 University of Notre Dame

- Demonstrated that conventional differentiable neural architecture search (DNAS) is not scalable with the increase of search spaces.
- Proposed RADARS, a scalable reinforcement learning (RL) aided DNAS framework that can explore large search spaces in a fast and memory-efficient manner. RADARS iteratively applies RL to prune undesired architecture candidates and identifies a promising subspace to carry out DNAS. Achieved similar search results with 100x memory usage reduction [3].

Software-Hardware Co-Design for DNN Accelerators

Nov. 2019 – Jan. 2020 University of Notre Dame

- Collaborated in proposing a co-design framework for heterogenous application-specific integrated circuit (ASIC) DNN accelerators using NAS. This ASIC-based design runs multiple tasks simultaneously using multiple accelerators. Achieved 17.77%, 2.49×, and 2.32× reductions in latency, energy, and area, respectively [4].
- Collaborated in proposing a cross-layer co-design framework for nvCiM DNN accelerators to find Pareto optimal solution for accuracy, robustness, latency, and energy consumption by co-exploring device, circuit, and architecture level design parameters via NAS. Achieved 3.17× higher energy efficiency than the state-of-the-art [5].
- Developed a python integration layer for open-source hardware performance evaluation tools (NeuroSIM and MAESTRO) that abstracts search results as hardware parameters and parses evaluation results as NAS-friendly data.

EDUCATION

Ph.D. Student in Computer Science

University of Notre Dame

Aug. 2019 – May 2024 (Expected)

B.S. in Electrical Engineering

Zhejiang University

Aug. 2015 – July 2019

SKILLS & AWARDS

Python C Assembly Verilog PyTorch Matlab Git Vivado

DAC Young Student Fellow

China National Scholarship

LEADERSHIP & SERVICE

Academic Director/Organizer

Model UN Association of ZJU

Sept. 2015 – Nov. 2018

Teaching Assistant

University of Notre Dame

Aug. 2019 – May 2020

SELECTED PUBLICATIONS

[1] **Z. Yan**, X. S. Hu, and Y. Shi, "Swim: Selective write-verify for computing-in-memory neural accelerators," *2022 59th ACM/IEEE Design Automation Conference (DAC)*, 2022.

[2] **Z. Yan**, X. S. Hu, and Y. Shi, "Computing in memory neural network accelerators for safety-critical systems: Can small device variations be disastrous?" *2022 International Conference on Computer-Aided Design (ICCAD)*, 2022.

[3] **Z. Yan**, W. Jiang, X. S. Hu, and Y. Shi, "Radars: Memory efficient reinforcement learning aided differentiable neural architecture search," *ASP-DAC*, 2022.

[4] L. Yang, **Z. Yan**, M. Li, *et al.*, "Co-exploration of neural architectures and heterogeneous asic accelerator designs targeting multiple tasks," in *2020 57th ACM/IEEE Design Automation Conference (DAC)*, IEEE, 2020.

[5] W. Jiang, Q. Lou, **Z. Yan**, *et al.*, "Device-circuit-architecture co-exploration for computing-in-memory neural accelerators," *IEEE Transactions on Computers*, 2020.