BEICHEN HUANG

huangb21@mcmaster.ca +1 647-896-3986 beichenhuang.github.io

EDUCATION

McMaster University, Canada *Sept. 2019 - now*

- Bachelor of Engineering: Mechatronics Engineering Cumulated GPA: *11.0 / 12.0*

RESEARCH INTEREST

Developing algorithms to enhance efficiency and accuracy of large-scale models, and designing cost-effective machine learning systems, with a focus on applying optimization theory for data-efficient tuning, model compression, and efficient inference.

PUBLICATION

- **MiLo: Effective Quantized MoE Inference with Mixture of Low-Rank Compensators** *Submitted to MLSys 2025*
  Beichen Huang*, Yueming Yuan*, Zelei Shao*, Minjia Zhang
- **Multidimensional Fractional Programming for Normalized Cuts** *NeurIPS 2024*
  Yannan Chen*, **Beichen Huang***, Licheng Zhao, Kaiming Shen
- **Aerial-IRS-Assisted Load Balancing In Downlink Networks** *ICASSP 2024*
  Shuyi Ren, **Beichen Huang**, Xiaoyang Li, Kaiming Shen

RESEARCH EXPERIENCE

Efficient Low-Bit Quantization and Inference System for MoE March 2024 - Present
Supervisor: Prof. Minjia Zhang *University of Illinois Urbana-Champaign*

- Conducted a comprehensive analysis and comparison of low-bit quantization techniques, factoring in MoE model structures and weight distributions, leading to **an adaptive and effective quantization strategy** tailored for complex MoE architectures.
- Proposed a novel calibration-free quantization algorithm, with jointly optimized low-rank compensators. Achieved **time-efficient** and nearly **loss-less 3-bit quantization** with minor memory overhead.
- Recovered **87% performance recovery** with **22% compression ratio**, and significantly outperformed prevailing quantization methods across 7 benchmark tests. Implemented a highly efficient 3-bit GeMM CUDA kernel, delivering a **1.3× speedup**.

Advancements in Fractional Programming for Clustering and Wireless Communication Sept. 2023 - May 2024
Supervisor: Prof. Kaiming Shen *The Chinese University of Hong Kong (ShenZhen)*

- Extended the **Quadratic Transform algorithm** to multi-dimensional cases, which effectively solved generalized fractional programming problems and addressed complex challenges in machine learning.
- Optimized the **NP-complete clustering problem** by analyzing from fractional programming perspective and directly applying Quadratic Transform, and achieved state-of-the-art clustering performance across 8+ benchmark datasets.
- Developed an **innovative wireless communication model** leveraging Aerial Intelligent Reflective Surface, and optimized load balancing within the model using an adaptive particle swarm optimization algorithm, significantly enhancing system efficiency.

WORK EXPERIENCE

Software Engineer Intern *Magna Electronics, May 2022 - May 2023*

- Designed, developed, and debugged for image processing algorithm with ground truth and debugging information visualization function for the autonomous driving system. Diligently managed the project repository on GitHub.
- Effectively maintained the C++ Advanced Driver-Assistance System program, mainly focused on solving the defects of the Human Machine Interface and the data pipeline in response to customer feedback.

Teaching Assisstant *McMaster University, Dec. 2021 - May 2022*

- Actively engaged in 10 lab and tutorial sessions related to embedded programming, designed and taught material to inspire students to have a clear understanding of the software for embedded systems.
- Taught and solved questions and requests from over 40 students, and received a high rating at the end of the term.

SKILLS

Programming: Pytorch, MATLAB, C, C++, ARM Assembly, Simulink, Keil, Git, LaTeX, R
Software & Tool: PyChram, MATLAB, Colab, VS Code, Autodesk Inventor, Altium Designer, NI Multisim