BEICHEN HUANG

huangb21@mcmaster.ca +1 647-896-3986
https://github.com/BeichenHuang https://beichenhuang.github.io/

EDUCATION

McMaster University, Canada *Sept. 2019 - now*

- Bachelor of Engineering: Mechatronics Engineering Cumulated GPA: *3.9 / 4.0*

RESEARCH INTEREST

Build efficient systems to make LLM affordable, develop high-performance algorithms to improve LLM efficiency and accuracy; with a focus on the application of optimization theory in algorithm, model quantization and compression.

RESEARCH EXPERIENCE

Research Assistant

Supervisor: Prof. Minjia Zhang *University of Illinois Urbana-Champaign, March 2024 - Present*

- Focused on developing an affordable and efficient LLM in the **Mixture-of-Expert (MoE)** structure, combining the mathematical **optimization** method with **Post Training Quantization** to compress the model.
- Conducted an in-depth evaluation of trending quantization methods, including both data-free methods and calibration methods, across 6 benchmarks; and successfully advancing the quantization from 4-bit to **3-bit**.
- Innovatively brought the **low rank matrix method** to compensate the error of quantized weight for MoE models, and conducted experiments on a variety of rank strategies. Resulted in significant performance improvement with negligible additional memory overhead.

Research Assistant

Supervisor: Prof. Kaiming Shen *The Chinese University of Hong Kong (ShenZhen), Sept 2023 - May 2024*

- Enhanced the Quadratic Transform algorithm to multi-dimensional cases, in order to address general **fractional programming** problems; and successfully apply the advanced method in solving complex challenges in machine learning and wireless communication.
- Addressed the **clustering** problem using fractional programming. Employed the Quadratic Transform for direct optimization of the **discrete NP-complete problem**, resulting in SOTA clustering performances on over 8 datasets.
- Formulated an innovative **wireless communication model** incorporating Aerial Intellectual Reflective Surface (AIRS). Optimizing load balancing within the model using adaptive particle swarm method, enhancing system efficiency.

PUBLICATION

- **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

WORKING 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 60 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