Ziming Liu (刘子铭)

E-mail: liuziming@comp.nus.edu.sg
Homepage: https://maruyamaaya.github.io/
Twitter/X: @lzm_mlsys
National University of Singapore / Peking University

Education

National University of Singapore, School of Computing

- Ph.D. in Computer Science **Jan. 2023 – Present**

National University of Singapore, School of Computing

- Master's degree in computer science (Artificial Intelligence) **Aug. 2021 – Jan. 2023**

Peking University, School of Electronics Engineering and Computer Science

- B.S. in Computer Science and Technology **Sep. 2016 – Jul. 2020**

Industry Experience

Microsoft Research Asia. **May. 2024 – current**

Research Intern, System Group

HPC-AI Tech. **May. 2022 – Dec. 2022**

Research Intern

ByteDance Inc. **Aug. 2020 – Jul. 2021**
Machine Learning Engineer, Lark

Research Interests

Machine Learning System and High Performance Computing.

Including distributed model training (parallelism schemes) / inference and serving systems.

Research Experience

Hanayo:HarnessingWave-like Pipeline Parallelism for Enhanced Large Model Training Efficiency
Advisor: Presidential Young Prof. You Yang ***Dec. 2022 – Apr.2023***

*Objective:We develop a new pipeline parallel technique to solve the problem the bubbles in existing pipeline model training techniques and achieve SOTA results in multiple tasks. (**Python**)*

- This paper has been accepted by SC '23(The International Conference for High Performance Computing, Networking, Storage, and Analysis).
- Design methods to help reduce the bubble rate and improve communication-computation overlap.
- Write the codes and carry out the experiments. Design experiments that can prove we outperform the existing techniques.
- Write the main parts of the paper.

EnergonAI: An Inference System for 10-100 Billion Parameter Transformer Models

Advisor: Presidential Young Prof. You Yang **Apr. 2022 – Dec. 2022**

*Objective: With the large Transformer models trending, we develop a new inference system that support multiple parallelism(Tensor, Data, Pipeline and so on) and use various techniques to speed up the process. (**Python**)*

- Design and implement checkpoint saving and loading system that supports various parallel schemes.

- Design and implement dynamic batch warping algorithm to speed up the process of multiple inference requests.
- Improve the implement of models like GPT and Bert to fit in with our parallel schemes.

Publication

Hanayo: Harnessing Wave-like Pipeline Parallelism for Enhanced Large Model Training Efficiency
Ziming Liu, Shenggan Cheng, Haotian Zhou, and Yang You
SC ’23, *In Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis*, 2023

HeteGen: Efficient Heterogeneous Parallel Inference for Large Language Models on Resource-Constrained Devices
Xuanlei Zhao, Bin Jia, Haotian Zhou, **Ziming Liu**, Shenggan Cheng, and Yang You
MLSys 2024, *In Proceedings of Machine Learning and Systems，2024*

Preprints

EnergonAI: An Inference System for 10-100 Billion Parameter Transformer Models
Jiangsu Du, **Ziming Liu**, Jiarui Fang, Shenggui Li, and Yongbin Li, Yutong Lu, Yang You
Arxiv: 2301.08658 , 2022

ATP: Adaptive Tensor Parallelism for Foundation Models
Shenggan Cheng, **Ziming Liu**, Jiangsu Du, and Yang You
Arxiv: 2209.02341, 2023

DSP: Dynamic Sequence Parallelism for Multi-Dimensional Transformers
Xuanlei Zhao, Shenggan Cheng, Zangwei Zheng, Zheming Yang, **Ziming Liu**, and Yang You
2024
Arxiv: 2403.10266, 2024

Skills

Languages: Python, C, C++, Latex
Frameworks: Pytorch, Huggingface, Numpy