Juyi Lin

lin.juy@northeastern.edu | (+1) 617-212-8191 | Northeastern University, Boston, US
GitHub | LinkedIn

EDUCATION

PhD student at Northeastern University, Boston, US Sept 2024 - May 2029(Expect)
Major in Computer Engineering, GPA: 3.83/4, Advisor: Prof. Yanzhi Wang

King Abdullah University of Science and Technology Sept.2022 - May.2024
Master of Science in Computer Science, GPA: 3.71/4, Advisor: Marco Canini

Zhejiang University Sept. 2018 - June 2022
Bachelor of Engineering in Electronic Engineering, GPA: 3.75/4

2020&2021's Second-Class Scholarship(Top 8%)

EXPERIENCES

Infrastructure Internship
Beijing VirtAI Tech March 2022 - July 2022

- Integrated PaddlePaddle distribution library and Deep Graph Library(DGL) into the performance measurement system.
- Conducted an in-depth performance analysis of PaddlePaddle distributed training.

Research Assistant
The University of Hong Kong July 2021 - Sept. 2021

- Optimized resource allocation and device placement for clusters.
- Designed a hierarchical scheduler, which employed Deep Reinforcement Learning (DRL) to enhance decision-making.
- Built a communication cost model for different distributed ML architectures.

PROJECTS

Vision Language Action(VLA) Robotics Model
NEU & Funded by EmbodyX Jan 2025 – Present

- Proposed **VOTE**, an efficient fine-tuning framework for parallel action prediction in VLA models, reducing computational overhead and accelerating inference. Paper link: VOTE. Code Link: Github.
- Proposed an ensemble voting strategy for the action sampling, improving model performance and enhances generalization across diverse tasks.
- Improved the average success rates of OpenVLA by over 20% across four LIBERO task suites, surpassed 7% average success rate of the state-of-the-art VLA model in SimplerEnv WidowX Robot, and accelerated action generation throughput by 39× on edge device NVIDIA Jetson Orin.

GNN neighbor sample acceleration
UMass Amherst July 2022 - August 2023

- Collaborated with Prof. Hui Guan and Prof. Marco Serafini. Designed mini-batch splitting algorithms, pruned redundant computation graphs, and extracted intermediate embeddings from cache.
- Proposed a graph pruning and embedding cache reuse strategy that reduced GPU memory usage by 47% without accuracy loss.

PUBLIC SERVICES

- AAAI-26 Program Committee, EuroSys'23 Shadow PC Reviewer
- Artifact Evaluation: MLSys'23, EuroSys'24/25, OSDI'24 & ATC'24
- Open Source Promotion Plan 2023

SKILLS

Programming Languages: Python, C++, Golang, Bash, Matlab, C
Tools: Git, LaTeX, Docker
Libraries/Frameworks: PyTorch, LoRA, Transformers lib, Diffusers, Deep Graph Library, PyTorch Geometric, Wandb