Yizhen Yao

Shanghai, China
yzhenyao.cs@gmail.com linkedin.com/in/yizhen-yao-b6bb28317/ preordinary.github.io/

Education

M.Sc. in Computer Technology **09/2022 – 03/2025**
Department of Computer Science and Engineering, **Shanghai Jiao Tong University** Shanghai, China
Supervised by Prof. Zhenzhe Zheng and Dr. Pengfei Zuo
GPA 3.84/4.00 **Ranking** 17/129 **IELTS** 7.0

B.Eng. in Computer Science **09/2018 – 06/2022**
Department of Computer Science and Engineering, **Shanghai Jiao Tong University** Shanghai, China
GPA 87.42/100

Research Interest: Machine Learning System; LLM System and Algorithm; Distributed Computing

Academic Experience

Multi-level Checkpoint Cache for LLM Training 03/2024 – 09/2024
- *Supervised by Dr. Pengfei Zuo*
- *First authorship; Planned to submit to EuroSys'25*
- *Keywords: LLM Training; In-Memory Checkpoint; Fast Failure Recovery*

- Analyzed the memory, computation, and communication bottlenecks in current parallelized LLM training schemes and the limitations of conventional checkpoint mechanisms.
- Developed a novel multi-level checkpoint mechanism where the LLM model weights and training states are saved to memory and disk at different frequencies depending on failure risk.
- Used Megatron-Deepspeed to implement prototype, drastically reducing checkpoint overhead and fault recovery efficiency.

Sublayer Skipping for Accelerating LLM Inference 09/2023 – 02/2023
- *Supervised by Dr. Pengfei Zuo*
- *First authorship; Submitted to AAAI, Under Review*
- *Keywords: LLM Inference; Inference Acceleration; Layer-wise Skipping*

- Developed a layer-wise skipping strategy to reduce the high computational cost and latency in large language model (LLM) inference.
- Performed comprehensive analysis on the importance of Attention and Feed Forward sublayers in transformer layers across a variety of models to devise the skipping algorithm.
- Contributed a training-free, auto-adaptive, sublayer-wise skipping method for both the prefilling and decoding phases of LLMs, demonstrating favorable inference performance over the baselines on various benchmarks and datasets.

Distributed Large Model Agent Chain 06/2024 – 09/2024
- *Collaborated with Dr. Bin Gao*
- *Co-author; Planned to submit*
- *Keywords: LLM Agent; Distributed Computing*

- In view of the current situation where a single large model agent cannot handle complex tasks, it is necessary to build an agent chain. How to efficiently deploy multiple agent chains in a GPU cluster is a potential problem.
- An agent chain scheduling algorithm is designed to optimize the agent's resource overhead and task completion time at the same time.
- Experiments in actual agent tasks show that our scheduling algorithm is superior to other traditional algorithms.

Domain Generalization in Federated Learning 03/2023 – 08/2023
- *Supervised by Prof. Zhenzhe Zheng*
- *First authorship; Submitted to TMC; Under Review*
- *Keywords: Mobile Computing; Federated Learning; Domain Generalization*

- Introduced a disentanglement approach to Federated Domain Generalization (FedDG), where the main objective is to generalize into unseen domains under the context of Federated Learning.
- Used a global model to extract domain-invariant features and a local model to extract domain-specific style features.
- Utilized contrastive learning for separate learning of domain-invariant and domain-specific features, designed one reconstruction loss functions for preserving information among features.
- Conducted tests and experiments on various benchmarks, demonstrating outstanding performance and even surpass most centralized DG methods.

Working and Teaching Experience

Huawei Cloud Computing Technology Co., Ltd. Chengdu, China
Cloud Storage Innovation Lab Intern 09/2023 – 08/2024

- Implemented the above In-Memory Checkpoint mechanism into the company's deep learning training framework Mindspore.
- Conducted LLM training and checkpoint testing through bash scripts and parallel processing.
- Collaborated in a team of forty people, participating in building training environment and clusters, designing application program interfaces (API).
- Participated in drafting technical reports, invention patents, and technical white papers to communicate results with management and general public.

CS307: Operating System Course in Shanghai Jiao Tong University Shanghai, China
Teaching Assistant 02/2023 – 06/2023

- Guide undergraduate students to complete homework, answer questions, and evaluate ultimate students' project presentations.

Awards and Honors

First Class Academic Scholarship, Shanghai Jiao Tong University, 2000 USD/years. 09/2022 – 09/2024
Huawei Scholarship, 1333 USD. 12/2023

Languages and Skills

Languages: English (IELTS 7.0), Chinese (Native)
Programming Language: Python, C++, C#, Matlab, Bash
Tools and Packages: Pytorch, Git, Docker, Cuda, Numpy, Pandas, SciKit Learn, Unity

Extracurricular

China Basic Psychological Counselor, Chinese Institute of Psychology 12/2023
Oxford Prospects Summer Programmes, Grade A- 08/2023
Pupils Teaching Volunteer, Sunflower Association 02/2023 – 06/2023