Yizhen Yao

Email Phone Homepage

EDUCATION

Shanghai Jiao Tong University Shanghai, China

- Major (M.E.) in **Computer Technology** 09/2022 - 03/2025
- ***GPA: 3.84/4 Ranking:5th/40***

Shanghai Jiao Tong University Shanghai, China

- B.E. in **Computer Science and Technology** 09/2018 - 06/2022
- ***GPA: 87.42/100***

ACADEMIC EXPERIENCE

No-Reference Image Quality Assessment 10/2022-02/2023

- Supervised by Prof. Zhenzhe Zheng

- Joint first authorship(first place); Under Review

- Keywords: Computer Vision; Image Quality Assessment

• Developed a novel pluggable and lightweight module for No-Reference Image Quality Assessment (NR-IQA), which evaluates the quality of an image against human evaluation criteria without a reference image.

• The proposed module (PLS) compliments existing backbone neural network model solutions by simultaneously extract local and global information and amplifying critical details to improve assessment accuracy.

• Conducted tests and evaluations on six NR-IQA benchmark datasets and tested PLS with different backbone models, allows flexible generalization of existing backbone models without significant retraining while achieving competitive results.

Domain Generalization in Federated Learning 03/2023-08/2023

- Supervised by Prof. Zhenzhe Zheng

- Joint first authorship(second place); Under Review

- Keywords: 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.

Sublayer Skipping for Accelerating LLM Inference 03/2024-06/2024

- Supervised by Research Scientist Pengfei Zuo

- Joint first authorship(second place); Submitted to NeurIPS, 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.

Multi-level Checkpoint Cache for LLM Training 09/2023-09/2024

- Supervised by Research Scientist Pengfei Zuo

- First authorship; Planned to submit to FAST'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.

INDUSTRY EXPERIENCE

Huawei Cloud Computing Technology Co., Ltd. Chengdu, China

Cloud Storage Innovation Lab Intern *09/2023 –09/2024*

• 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.

Honors

First Class Academic Scholarship, Shanghai Jiao Tong University. 09/2022-09/2024

Huawei Scholarship 12/2023

Merit Student of Shanghai Jiao Tong University. 09/2023

LANGUAGES & SKILLS

Programming Language: Python, C++, C#, Matlab, Bash

Tools and Packages: Pytorch, git, docker, Numpy, Pandas, SciKit Learn, Unity

Languages: Chinese (Native), English (Professional)

Extracurricular

Oxford Prospects Summer Programmes, Grade A- 08/2023

China Basic Psychological Counselor 12/2023

Monitor of Class 5, Department of Computer Science, 2022 09/2022-09/2023

Pupils Teaching Volunteer, Sunflower Association 02/2023-06/2023