Yuxuan Lou

National University of Singapore | yuxuanlou@u.nus.edu | +65 82600153

yuxuanlou.info | Google Scholar | Github

Research Interest

- Efficient large language model scaling with Mixture of Experts
- Multimodal Foundation Model Adaptation from large language models
- Multimodal Embodied LLM Agent

Education

National University of Singapore, School of Computing, HPC-AI Lab 2023 – Present
- Ph.D. in Computer Science, Advised by Prof. Yang You

National University of Singapore, School of Statistics and Probability 2020 – 2022
- M.Sc. in Statistics

Fudan University, School of Mathematical Science 2016 – 2020
- B.Sc. in Applied Mathematics

Research Experiences

Speech-Text Foundation Model with Mixture of Experts, NUS - Apple Jan 2025 - May 2025
- Developed **MoST**, a novel speech-text foundation model featuring a Modality-Aware Mixture of Experts (MAMOE) architecture which directs tokens to specialized pathways for enhanced cross-modal understanding
- Engineered an efficient, three-stage transformation pipeline to adapt a pre-trained Mixture of Experts (MoE) language model for speech-text tasks
- Achieved competitive performance across multiple speech-text benchmarks using exclusively open-source data, contributing to reproducible AI research through full code and model release

Mixture of Reasoning Students Distilled from Dense Model, NUS - Apple Sep 2024 – Jan 2025
- Developed **MoRS** (Mixture of Reasoning Students), a four-stage distillation method that compresses large language models (70B parameters) into efficient mixture-of-experts architectures (12B parameters, 3B activated parameters) while preserving specialized reasoning capabilities across multiple domains
- Achieved better or comparable results compared with comparable models by significant margins - up to +14.5% on reasoning benchmarks (ARC Challenge: 78.0%, MMLU: 62.2%, HumanEval: 40.4%) while requiring fewer training tokens than competitors.
- Created the first framework to distill dense language models into MoE architectures without relying on pre-existing small models, using domain-specific expert specialization (mathematics, coding, scientific reasoning) with a shared-expert design for optimal knowledge integration

Multimodal LLM Agent with Retrieval Augmented Planning, NUS - Panasonic Oct 2023 - May 2024
- Developed **RAP**, a Multimodal planning agent which leverages past successful experiences to enhance decision-making process
- Developed **EnvBridige**, a Multimodal embodied agent which can transfer knowledge from diverse embodied environments and enhance planning ability
- SOTA results on text-only environments(ALFWorld, Webshop), Significant improvements on multimodal robotics benchmarks(Franka Kitchen, Meta-World, RLBench)

Vision Model Scaling with Mixture of Experts, HPC-AI Lab Mar 2021 – Jan 2022
- Developed large-scale vision models: **Sparse-MLP**, **Widenet** based on Mixture of Experts
- Proposed a fully-MLP architecture with conditional computation in two directions and extended MoE to spatial dimension of image representation.

Selected Publications

MoST: Modality-Aware Mixture of Experts for Efficient Speech-Text Foundation Model(2025)

Yuxuan Lou, Kai Yang, Yang You

Project Page

MoRS: Distill Large Language Model into Mixture of Reasoning Students(2025)

Yuxuan Lou, Yang You

In Submission

EnvBridge: Bridging Diverse Environments with Cross-Environment Knowledge Transfer for Embodied AI(2024)

Tomoyuki Kagaya*, **Yuxuan Lou***, Thong Jing Yuan*, Subramanian Lakshmi*, Jayashree Karlekar, Sugiri Pranata, Natsuki Murakami, Akira Kinose, Koki Oguri, Felix Wick, Yang You

arxiv.org/abs/2410.16919

RAP: Retrieval-Augmented Planning with Contextual Memory for Multimodal LLM Agents(2024)

Tomoyuki Kagaya*, **Yuxuan Lou***, Thong Jing Yuan*, Subramanian Lakshmi*, Jayashree Karlekar, Sugiri Pranata, Natsuki Murakami, Akira Kinose, Koki Oguri, Felix Wick, Yang You

arxiv.org/abs/2402.03610

Cross-token modeling with conditional computation(2022)

Yuxuan Lou, Fuzhao Xue, Zangwei Zheng, Yang You

arxiv.org/abs/2109.02008

Open Source Projects

Colossal-AI: Making large AI models cheaper, faster, and more accessible 41k star

- A collection of parallel components for distributed training of large deep learning models
- Managed and contributed to Colossal-AI examples

awesome mixture-of-experts 1.2k star

- A collection of awesome Mixture of Experts papers and projects

MoST: Modality-Aware Mixture of Experts for Efficient Speech-Text Foundation Model

- Official implementation of MoST, a novel speech-text foundation model featuring a Modality-Aware Mixture of Experts (MAMOE) architecture which directs tokens to specialized pathways for enhanced cross-modal understanding

RAP: Retrieval-Augmented Planning with Contextual Memory for Multimodal LLM Agents

- Official implementation of RAP, a Multimodal planning agent which leverage past successful experiences to enhance decision-making process

Skills & Technologies

Deep learning libraries: Pytorch, Tensorflow, Keras, Deepspeed, Colossal-AI

Parallel Training & Optimization: Model parallel, sequence parallel, data parallel training on GPU/TPU clusters