Yao Fu

1.45, Informatics Forum
The University of Edinburgh
Edinburgh - EH8 9AB, United Kingdom

Email: Y.Fu@ed.ac.uk
LinkedIn: Yao Fu
Google Scholar: Yao Fu

Research

I study the intersection of **machine learning** and **distributed systems** with the goal of developing systems that are affordable for everyone to deploy **large language models**, fostering broader involvement in AI innovation. My current research primarily focuses on the **efficient inference** of large language models.

Education

The University of Edinburgh Edinburgh, UK
Ph. D. - Computer Science; Supervisor: Prof. Luo Mai Sept. 2021 - present

Sun Yat-sen University Guangzhou, China
B. E. - Computer Science and Technology Sept. 2017 - Jun. 2021

Selected Research Projects

ServerlessLLM: Locality-Enhanced Serverless Inference for Large Language Models Apr. 2022 - Present

- Developed a loading-optimized checkpoint format and a fast checkpoint loader. (4X faster than SafeTensor)
- Designed a live-migration mechanism for locality-driven LLM inference. (2X better than serverless scheduling policies)
- Designed a model loading scheduler for locality-aware server allocation (reducing start-up latency by 1.86X).
- Evaluated Phantom against SOTA serverless inference systems: **Ray Serve** and **KServe**, showing 10 - 200X speed up.
- GitHub Project: ServerlessLLM/ServerlessLLM (https://github.com/ServerlessLLM/ServerlessLLM)

MoE-Infinity: Activation-Aware Expert Offloading for Efficient MoE Serving Apr. 2022 - Present

- Evaluated tensor prefetching and caching policies for **Mixture of Experts** model inference.
- Developed Archer's **Python binding**, compatible with **HuggingFace Transformers**.
- Evaluated Archer with DeepSpeed Infinity, showing 9X performance improvement.
- GitHub Project: TorchMoE/MoE-Infinity (https://github.com/TorchMoE/MoE-Infinity)

Selected Open-Source Projects

Open MoE LLM Leaderboard Mar. 2024 - Present

- Developed a benchmark suite to measure the performance of state-of-the-art (SOTA) Mixture of Experts (MoE) LLMs including Grok-1, DBRX, Mixtral-8x7B.
- Deployed on various hardware models including A100, H100, A5000, and RTX 4090.
- Benchmarked state-of-the-art MoE LLMs on diverse tasks including reasoning, coding, and long-context generation.
- Link: https://huggingface.co/spaces/sparse-generative-ai/open-moe-llm-leaderboard

Machine Learning Systems: Design and Implementation Oct. 2022 - Present

- Contributed to the chapters on Deep Learning Recommendation Systems and the section on Federated Learning.
- Link: https://github.com/openmlsys/openmlsys-zh

Work Experience

Tencent Guangzhou, China
Research Intern, MLSys Team; Mentor: Feng Lin May 2021 - Jan. 2022

- Designed an SLO-aware model update scheduler for a large-scale **Deep Learning Recommender System**
- Proposed an inference model state manager to monitor model health and implement low-latency rollbacks.
- Mitigated a 2.32% SLO drop during network congestion in real-world **Short Video** services with over one billion users.

Selected Publications

*Co-primary authors

[1] **Yao Fu**, Leyang Xue, Yeqi Huang, Andrei-Octavian Brabete, Dmitrii Ustiugov, Yuvraj Patel, and Luo Mai. Serverlessllm: Locality-enhanced serverless inference for large language models. *OSDI*, 2024.

[2] Leyang Xue, **Yao Fu**, Zhan Lu, Luo Mai, and Mahesh Marina. Moe-infinity: Activation-aware expert offloading for efficient moe serving. *arXiv preprint arXiv:2401.14361*, 2024.

[3] Jie Ren*, Xidong Feng*, Bo Liu*, Xuehai Pan*, **Yao Fu**, Luo Mai, and Yaodong Yang. Torchopt: An efficient library for differentiable optimization. *Journal of Machine Learning Research*, 24(367):1–14, 2023.

[4] Chijun Sima*, **Yao Fu***, Man-Kit Sit, Liyi Guo, Xuri Gong, Feng Lin, Junyu Wu, Yongsheng Li, Haidong Rong, Pierre-Louis Aublin, and Luo Mai. Ekko: A large-scale deep learning recommender system with low-latency model update. *OSDI*, 2022.

[5] Yipeng Zhou, Xuezheng Liu, **Yao Fu**, Di Wu, Jessie Hui Wang, and Shui Yu. Optimizing the numbers of queries and replies in convex federated learning with differential privacy. *IEEE Transactions on Dependable and Secure Computing*, 2023.

Programming Languages

Advanced: C/C++, Python, Go
Intermediate: Java, Haskell, Matlab

Technical Skills

- Model serving libraries: Triton Inference Server, Ray Serve
- ML frameworks: PyTorch, HuggingFace ecosystem (Transformers, Accelerator, Safetensors)
- Containerization and orchestration: Docker, Kubernetes, Knative serving, KServe
- Network programming and API technologies: network sockets, gRPC, Flask, FastAPI
- System profiling & debugging tools: NVIDIA Nsight Systems, fio, iostat, perf

Awards

- Outstanding Graduates of Yat-sen Honors School, Sun Yat-sen University, 2021
- Outstanding Undergraduate Thesis Award, Sun Yat-sen University, 2021
- First Prize, Sun Yat-sen University Scholarship, 2020
- Parallel Fund Award, The 7th “Intel Cup” Parallel Application Challenge, 2019
- First Prize in Guangdong Province, China Undergraduate Mathematical Contest in Modeling, 2019
- Second Prize, Sun Yat-sen University Scholarship, 2018-2019

Talks

16th USENIX Symposium on Operating Systems Design and Implementation (OSDI)	Jul. 2022
Sixth Annual UK System Research Challenges Workshop	Dec. 2021

Teaching and Academic Services

- HPCA 2024, Artifact Evaluation Committee
- Extreme Computing 2023 Fall/2021 Fall, Marker
- Oxford Machine Learning Summer School 2021, Teaching Assistant