Tingfeng Lan

🏠 antlera.github.io Antlera ✉ erc8gx@virginia.edu

Research Interests

- I am interested in co-designing systems and algorithms for **efficient large-scale machine learning**. I am particularly interested in applying my research in the development and deployment of foundation models, such as GPT and LLaMA.
- Current research: 1) rethinks the design of large-scale systems for LLM applications in the interaction between computing and storage systems, and 2) optimizes/offloads/accelerates critical operations of LLM apps to the most appropriate hardware to harmonize heterogeneity, efficiency, and performance.

Education

Department of Computer Science, University of Virginia **Sep 2024 - Present**

Ph.D. in Computer Science *VA, United States*

- Advisor: Prof. Yue Cheng

Department of Computer Science, Sichuan University (SCU) **Sep 2020 - Jun 2024**

B.Eng in Computer Engineering *Sichuan, China*

- Advisor: Prof. Mingjie Tang (Ph.D. Purdue University)

Work Experience

AntGroup AI Infrastructure Group **Sep 2023 - July 2024**

Mentor: Qinglong Wang *Research Intern*

- Build efficient training system over heterogeneous GPUs.
- Optimize distributed parallelism for Parameter Efficient Fine-tuning (PEFT).
- (co-)Design quick checkpoint scheme for Large Language Models training.

Publications

- **ZenFlow: Enabling Stall-Free Offloading Training via Asynchronous Updates** Preprint, 2024.
  Tingfeng Lan, Yusen Wu, Bin Ma, Zhaoyuan Su, Rui Yang, Tekin Bicer, Dong Li, Yue Cheng
- **λScale: Enabling Fast Scaling for Serverless Large Language Model Inference.** Preprint, 2024.
  Minchen Yu, Rui Yang, Chaobo Jia, Zhaoyuan Su, Sheng Yao, **Tingfeng Lan**, Yuchen Yang, Yue Cheng, Wei Wang, Ao Wang, Ruichuan Chen.
- **mLoRA: Fine-Tuning LoRA Adapters via Highly-Efficient Pipeline Parallelism in Multiple GPUs.**
  51th International Conference on Very Large Data Bases (**VLDB'25**).
  Zhengmao Ye*, Dengchun Li*, Zetao Hu, **Tingfeng Lan**, Jian Sha, Sicong Zhang, Lei Duan, Jie Zuo, Hui Lu, Yuanchun Zhou and Mingjie Tang. (To appear)
- **DLRover-RM: Resource Optimization for Deep Recommendation Models Training in the Cloud.**
  50th International Conference on Very Large Data Bases (**VLDB'24**).
  Qinglong Wang*, **Tingfeng Lan***, Yinghao Tang, Bo Sang, Haitao Zhang, Jian Sha, Hui Lu, Ke Zhang, and Mingjie Tang.
- **PathBee: Accelerating Shortest Path Querying via Graph Neural Networks.**
  Jiale Lao, Yinghao Tang, **Tingfeng Lan**, Mingjie Tang, Yuanchuan Zhou, and Jianguo Wang. (Preprint)

* denotes equal contribution

Research Experience

Efficient Serverless Inference Scaling for Large Language Models **Aug 2024 - Present**

Advisors: Prof. Yue Cheng (UVA); Prof. Wei Wang (HKUST) *Research Assistant*

- Developed λScale, a serverless inference platform leveraging high-speed RDMA networks to significantly reduce model startup overhead for large language models (LLMs).
- Implemented λPipe, enabling adaptive multicast of model parameters and dynamic construction of execution pipelines to perform distributed inference during model loading.
- Optimized memory efficiency across GPU and host memory through a locality-driven model startup and efficient memory management, achieving up to 5x improvement in tail latency and 31.3% resource cost reduction compared to state-of-the-art methods.

Efficient LLM Fine-tuning and Serving via Multi-LoRA Optimization 🔗 **Aug 2023 - Mar 2025**

Advisors: Prof. Hui Lu (UTA); Prof. Mingjie Tang (SCU) *Research Assistant*

- Design and implement m-LoRA, an innovative framework enabling fine-tuning multi-task Large Language Models (LLMs) with multiple LoRA adapters.
- Enhanced traditional LoRA fine-tuning methods, achieving parallel training across multiple LoRA adapters and drastically reducing memory redundancy by sharing base model.
- Optimized memory usage efficiency through meticulous data sharding alignment and scheduling in LoRA multitasking.

Resource-aware Optimization on Distributed Machine Learning System 🔗 **Apr 2023 - Oct 2023**

Advisors: Prof. Hui Lu (UTA); Prof. Mingjie Tang (SCU) *Research Assistant*

- Developed DLRover, a cloud-native system for training Deep Learning Recommendation Models (DLRM), integrating resource-aware optimizations to boost performance and efficiency.
- Constructed in-depth memory consumption and resource-throughput models for DLRM training, accounting for I/O overheads and computational demands.
- Designed a tri-phase algorithm to dynamically allocate resources throughout the DLRM training lifecycle, based on the performance models.

Open Source Projects

mLoRA: An Efficient "Factory" to Build Multiple LoRA Adapters 🔗 **Sep 2023 - Present**

Received 300+ ⭐ *on GitHub*

mLoRA is an open-source framework designed for efficient fine-tuning of multiple Large Language Models (LLMs) using multiple LoRA adapters.

- Designed and implemented a training mechanism "BatchLoRA" which allows multiple LoRA adapters to share the pre-trained base model concurrently with reduced kernel launch overhead.

DLRover: An Automatic Distributed Deep Learning System 🔗 **Jun 2023 - Present**

Received 1.4k+ ⭐ *on GitHub, Joined LF AI & Data Foundation* ⚡ *Top 3 Contributor*

DLRover is an automatic system aiming to train large AI models easy, stable, fast, and green.

- Designed and implemented a hyper-parameter autotuner to optimize performance-relevant configurations, like micro-batch size, for maximum hardware utilization. Achieved over 95% memory utilization within a 30s estimation and re-configuration time.
- Create an elastic trainer, allowing for real-time hyper-parameter configuration during training sessions, thereby eliminating the restart overheads typically necessary in conventional training frameworks.