Shenggui Li

Singapore | (+65) 8817 3562 | shenggui001@e.ntu.edu.sg | https://github.com/FrankLeeeee

EDUCATION

Nanyang Technological University — **Singapore**
PhD in Computer Science — Aug 2024 – Present
Supervisors: Prof. Tianwei Zhang, Prof. Ivor Tsang

Nanyang Technological University — **Singapore**
Bachelor of Engineering in Computer Science — Aug 2017 – June 2021
- CGPA: 4.88 / 5.00 (First Class Honor)

RESEARCH INTEREST

- **Machine Learning:** Foundation models, resource-efficient training and fine-tuning, interpretable and trustworthy AI
- **High Performance Computing:** Efficient distributed system for large-scale model training and inference
- **Cloud Computing:** efficient scheduling system for deep learning jobs, heterogeneous system for efficient model training and inference

ACHIEVEMENT

- A*STAR Computing and Information Science Scholarship (2024-2028)
- 1st place in the GPAW application (my track) and 2nd place overall as the captain of NTU HPC Team in the Student Cluster Competition in the International Supercomputing Conference (ISC 2021)
- 2nd place in IO500 benchmark as the captain of the NTU HPC Team in the Student Cluster Competition of the International Conference for High Performance Computing, Networking, Storage, and Analysis (SC 2020)
- Merit Award at APAC HPC-AI Competition (2020)
- NTU President Research Scholar (2018-2019, URECA programme)
- NTU School of Computer Science and Engineering Dean's List (2017-2018, 2018-2019)
- NTU Science & Engineering Undergraduate Scholarship (2017-2021)
- Singapore SM1 Scholarship (2012-2016)

PUBLICATIONS

- **Large-Scale Distributed Training**
  - **Shenggui Li**, Fuzhao Xue, Chaitanya Baranwal, Yongbin Li, and Yang You. 2023. Sequence Parallelism: Long Sequence Training from System Perspective. In Proceedings of the 61st **Annual Meeting of the Association for Computational Linguistics (ACL '23)** (Volume 1: Long Papers), Association for Computational Linguistics, Toronto, Canada, 2391–2404. DOI: https://doi.org/10.18653/v1/2023.acl-long.134
  - **Shenggui Li**, Hongxin Liu, Zhengda Bian, Jiarui Fang, Haichen Huang, Yuliang Liu, Boxiang Wang, and Yang You. 2023. Colossal-AI: A Unified Deep Learning System For Large-Scale Parallel Training. In Proceedings of the 52nd **International Conference on Parallel Processing (ICPP '23)**. Association for Computing Machinery, New York, NY, USA, 766–775. https://doi.org/10.1145/3605573.3605613
  - Jiarui Fang, Zilin Zhu, **Shenggui Li**, Hui Su, Yang Yu, Jie Zhou, and Yang You. 2023. Parallel Training of Pre-Trained Models via Chunk-Based Dynamic Memory Management. **IEEE Transactions on Parallel and Distributed Systems (TPDS '23)** 34, 1 (2023), 304–315. DOI: https://doi.org/10.1109/TPDS.2022.3219819
- **Efficient LLM Inference**
  - Cunxiao Du, Jing Jiang, Yuanchen Xu, Jiawei Wu, Sicheng Yu, Yongqi Li, **Shenggui Li**, Kai Xu, Liqiang Nie, Zhaopeng Tu and Yang You. GliDe with a CaPE: A Low-Hassle Method to Accelerate Speculative Decoding. **The International Conference on Machine Learning (ICML '24)**, ArXiv abs/2402.02082 (2024): n. pag.
- **Cluster Scheduling**
  - Zhengda Bian, **Shenggui Li**, Wei Wang, and Yang You. 2021. Online evolutionary batch size orchestration for scheduling deep learning workloads in GPU clusters. In Proceedings of the **International Conference for High Performance Computing, Networking, Storage and Analysis (SC '21)**. Association for Computing Machinery, New York, NY, USA, Article 100, 1–15. https://doi.org/10.1145/3458817.3480859
- **Competition Technical Report**

- **Shenggui Li** and Bu-Sung Lee. 2022. Critique of "MemXCT: Memory-Centric X-Ray CT Reconstruction With Massive Parallelization" by SCC Team From Nanyang Technological University. **IEEE Transactions on Parallel and Distributed Systems (TPDS '22)** 33, 9 (2022), 2058–2061. DOI:https://doi.org/10.1109/TPDS.2021.3128040

RESEARCH EXPERIENCE

HPC-AI Lab @ NUS – Research Assistant **April 2021 – July 2022**

- Worked as a full-time research assistant on research projects related to model parallelism for efficient large-scale model training, supervised by Prof. Yang You**.**
- Developed a load balancing algorithm for federated learning to support efficient distributed training for large models on heterogeneous systems.
- Worked closely with Dr. Fuzhao Xue and developed a novel parallelization method for scaling deep learning models for long-sequence modeling called sequence parallelism. Sequence parallelism can achieve 27x longer sequence length with no drop in throughput compared with other methods.
- Supervised an undergraduate student on his final-year project on the integration of linear-complexity attention mechanism with sequence parallelism.
- Our work on sequence parallelism was accepted by ACL 2023 with me as the first author.
- Served as the HPC system expert in my spare time for the lab to build and maintain the lab's internal Nvidia DGX station, and created the oh-my-server project to provide compressive usage guidance for the lab's machine and several well-known national supercomputers.

MMLab @ NTU - Final Year Project **September 2020 – March 2021**

- Worked on self-supervised learning for computer vision, supervised by Dr. Jiahao Xie and Prof. Chen Change Loy.
- Proposed a method called ContrastiveODC to integrate contrastive learning and clustering. ContrastiveODC outperforms the baseline models such as SimCLR and MoCo by 1% in top 1 accuracy on the ImageNet dataset.

SCALE @ NTU - URECA Programme **September 2018 – March 2019**

- Worked on sentiment analysis using machine learning, supervised by Prof. Jagath Chandana Rajapakse.
- Developed a support vector machine and deep neural network model using CNN and RNN to predict the valence and arousal of speech for emotion analysis.

INDUSTRY EXPERIENCE

HPC-AI Technology – Vice President **August 2022 – August 2024**

- Joined the startup as a founding member and raised more than 27 million USD until Series A from top-tier investors including Sequoia China, Zhen Fund, Sinovation Ventures, BlueRun Ventures and Singtel as the Executive Vice-President.
- Led the open-source team to develop Colossal-AI as the top contributor, which is a unified system to offer acceleration to large-scale model training and inference. Colossal-AI outperforms existing solutions and has gained more than 30K stars on GitHub.
    - Led the design and implementation of the ZeRO and Gemini modules to speed up zero-redundancy data-parallel training by 20% compared to the baseline.
    - Led the design and implementation of the ShardFormer module to enable easy conversion from Hugging Face models or customized models to 3D parallel models for large-scale distributed training.
    - Led the design and implementation of the Sequence Parallelism module to support ring-style, gather-split-style, and all-to-all-style implementations, enabling max 200K sequence length on 8 H100 GPUs.
    - Led the design and implementation of the Colossal-Inference module to enable efficient inference of LLMs via optimizations including continuous batching, fast attention kernels, speculative decoding and tensor parallelism.

Colossal-Inference can outperform the existing baseline by more than 30% in throughput on a single H100 GPU.
    - Led the ColossalChat project to replicate ChatGPT's training workflow and speed up the training pipeline by 7 times compared to the PyTorch-native baseline.
- Led the internal project to speed up the inference of Stable Diffusion models including V1, V2 and SDXL variants using Tensor-RT and integrate the optimized inference pipeline with Stable Diffusion Web UI.
- Led the product team as a full-stack engineer to develop ColossalCloud, which is a compute platform for users to gain access to a variety of GPUs for model training and deployment. The platform offers some out-of-the-box AI solutions integrated with Colossal-AI at scale and at a low cost.
- Led the SwiftInfer project to accelerate StreamingLLM with TensorRT, the optimized implementation can speed up multi-round conversation by 46% in throughput.
- Led the Open-Sora project to democratize efficient video generation. As the core system engineer, I implemented and optimized the essential pipelines including data processing, data loading, model training, model inference, Gradio-based application demo, and Next.js-based gallery.
- Built and maintained the internal cluster from scratch.
- Participated in key corporate decision-making processes.

YITU – Computer Vision Engineer Intern **Aug 2019 – July 2020**

- Worked on computer vision projects to deliver business solutions to customers, supervised by Dr. Xulei Yang.
- Developed a pipeline of classification, detection, and OCR models for the smart city traffic monitoring system and identity card recognition system.
- Voluntarily developed and maintained an efficient machine learning framework specialized for OCR tasks with features such as command line tool, auto-experiment by configuration, configurable automatic dataset analysis, distributed training, support for customization and extension, mixed-precision training and one-click model delivery with 80% code contribution. This framework can shorten the OCR model development by 50%.
- Voluntarily and independently developed and maintained a multimedia viewer application in my spare time using React, Redux, Django, and Redis. This application helped the Computer Vision Team to efficiently view media files and ground truth annotations on remote servers located in China under low international bandwidth conditions. This application was used and liked by more than 10 colleagues on the team.
- Voluntarily and independently developed a Python-based command line tool in my spare time which contains useful scripts compatible with the company's internal machine learning system to allow for rapid experimentation and development.

Goldman Sachs - Software Engineering Intern **July 2020 – August 2020**

- Worked as a full-stack engineer to create dashboards to monitor the health status and traffic statistics of the internal corporate services using React, Redux, Ant Design, Chart.js, Spring Boot, MongoDB, and Kafka. (Internship shortened due to Covid 19)

ACADEMIC TALKS

- Colossal-AI: Breakthroughs in Efficient AI, talk at the Fortieth International Conference on Machine Learning (ICML), Hawaii, 2023
- Sequence Parallelism: Long Sequence Training from System Perspective, online sharing at the MLSys Seminar@SG group, Singapore, 2023
- Colossal-AI: Scaling AI Models in the Big Model Era, online tutorial at the Nvidia GTC, 2023
- Colossal-AI: Scaling Large AI Models on Distributed Systems and Supercomputers, tutorial at the International Conference for High Performance Computing, Networking, Storage and Analysis (SC), Dallas, 2022
- Sequence Parallelism: Long Sequence Training from System Perspective, online sharing at Biren Technology, Shanghai, 2021
- Sequence Parallelism: Long Sequence Training from System Perspective, online sharing at WeChat Group, Beijing, 2021

CO-CURRICULAR ACTIVITIES

- Captain & Member @ NTU High Performance Computing Team Jan 2020 – June 2021
- Core committee member @ NTU Google Developer Student Club Oct 2019 – July 2020
- President @ The Institution of Engineering and Technology (IET) - NTU Student Branch Aug 2018 – July 2019

SKILLS

- Programming Language – Python, Java, JavaScript, C/C++, Golang
- Technologies – PyTorch, TensorFlow, React, Redux, Node.js, FastAPI, Go-ZeRO, MongoDB, MySQL, Firebase, Kubernetes, Docker, Ansible, Slurm, AWS, Alibaba Cloud, Bytedance Volcengine