Jiamin Li

Email: jiaminli.icy@gmail.com ⋄ Tel: +852 54847038
Ph.D. Candidate (4^{th} year) ⋄ Computer Science ⋄ City University of Hong Kong

RESEARCH INTERESTS

- Distributed machine learning system: accelerating large-scale distributed DNN tasks
- Adaptive and sparse computation: exploring new computing paradigms to keep scaling DNN models
- Simulation: building accurate performance simulator of DNN training workloads
- Resource scheduling in GPU clusters: designing efficient scheduling algorithms for DNN tasks

EDUCATION

The University of Texas at Austin, Austin, TX, United States Apr. 2023 – Present
Visiting Student Researcher at UTNS Group, Department of Computer Science.
Supervisor: Prof. Aditya Akella

City University of Hong Kong, Kowloon, Hong Kong Sep. 2019 – Present
Ph.D. Candidate, Department of Computer Science.
Supervisors: Prof. Hong Xu (The Chinese University of Hong Kong), Prof. Cong Wang (CityU)
Dissertation Title: TBD

City University of Hong Kong, Kowloon, Hong Kong Aug. 2015 – Jul. 2019
B.S., Department of Computer Science (First Class Honours).
Advisor: Dr. Shiqi Wang
Dissertation Title: "Mobile Face Anti-spoofing with Deep Learning"

University of Missouri, Columbia, MO, United States Jul. 2017 – Aug. 2017
Big Data Analysis Summer Program, Department of Computer Science.

PROFESSIONAL EXPERIENCES

Research Intern, MLSys - AI Lab, ByteDance May. 2019 – May. 2021
Supervisor: Dr. Yibo Zhu
We build a task scheduler for training and inference GPU clusters. The key idea is to exploit cluster-level elasticity by loaning idle inferences servers for training and job-level elasticity by scaling jobs to better utilize the dynamic resource pool.

Part-time Research Assistant, City University of Hong Kong May. 2018 – May. 2019
Supervisor: Prof. Hong Xu
In backpropagation of DNN training, the update of some DNN model gradients is tiny across multiple iterations. We use a threshold to adaptively control the gradients that need to be sent each time so that the transfer time of communication operations can be reduced.

Backend Developer Intern, Jardine Matheson & Co. Limited May. 2017 – May. 2018
Design and develop web services to facilitate employee recruitment for the Group Human Resources department.

PUBLICATIONS

Preprints

[P1] **Jiamin Li**, Cheng Luo, Ziyue Yang, Lei Qu, Peng Cheng, Cong Wang, Hong Xu, "Merak: An Analytical Performance Simulator for Large Scale Distributed Training", under review.

Conference proceedings

[C3] **Jiamin Li**, Yimin Jiang, Yibo Zhu, Cong Wang, Hong Xu, " Accelerating Distributed MoE Training and Inference with Lina", USENIX Annual Technical Conference (**ATC**), 2023. (Acceptance Rate = 18.4%)

[C2] **Jiamin Li**, Hong Xu, Yibo Zhu, Zherui Liu, Chuanxiong Guo, Cong Wang, "Lyra: Elastic Cluster Scheduling for Deep Learning", ACM European Conference on Computer Systems (**EuroSys**), 2023. (Acceptance Rate = 16.1%)

[C1] Kaiwei Mo, Chen Chen, **Jiamin Li**, Hong Xu, Chun Jason Xue, "Two-Dimensional Learning Rate Decay: Towards Accurate Federated Learning with Non-IID Data", IEEE International Joint Conference on Neural Networks (**IJCNN**), 2021.

Journals

[J2] Libin Liu, Hong Xu, Zhixiong Niu, Jingzong Li, Wei Zhang, Peng Wang, **Jiamin Li**, Jason Xue Chun, Cong Wang, "ScaleFlux: Efficient Stateful Scaling in NFV", IEEE Transactions on Parallel and Distributed Systems, 2022 (**TPDS**).

[J1] Libin Liu, Chengxi Gao, Peng Wang, Hongming Huang, **Jiamin Li**, Hong Xu, Wei Zhang, "Bottleneck-Aware Non-Clairvoyant Coflow Scheduling with Fai", IEEE Transactions on Cloud Computing, 2021 (**TCC**).

SELECTED PROJECTS

Accelerating Distributed Training and Inference of Mixture-of-Experts models Ongoing
Collaborate with MLSys team at ByteDance
Distributed training of MoE models is prone to low efficiency, mainly due to the interleaved all-to-all communication during model computation. In Lina, we propose a new communication scheduling scheme based on tensor partitioning that prioritizes the all-to-all operations over other communication.

Performance Simulator for Large-scale Distributed DNN Training Ongoing
Collaborate with Networking Research Group at Microsoft Research Asia
We design a performance simulator to accurately predict the step time of large-scale distributed DNN training tasks. Merak introduces an analytical formulation to compute the all-reduce kernel running time. We also build an ML model to predict the running time slowdown caused by concurrent execution in wait-free backpropagation.

Elastic Cluster Scheduler for DNN training jobs Completed
Collaborate with MLSys team at ByteDance
Organizations often build separate training and inference clusters for deep learning. This leads to low utilization in inference clusters and long queuing for jobs in training clusters. Lyra is designed and built to address these problems by cluster-level capacity loaning and job-level elastic scaling.

SELECTED AWARDS

Research Activity Funds	City University of Hong Kong	2023
EuroSys Student Travel Grant	ACM EuroSys	2023
Full Postgraduate Studentship	City University of Hong Kong	2019 – 2023
Dean's List (College of Engineering)	City University of Hong Kong	2015 – 2019

TEACHING ASSISTANT

2022 Fall	CS2311, Computer Programming
2022 Spring	CS4296 & CS5296, Cloud Computing
2021 Fall	CS4394 & CS5294, Information Security and Management
2021 Spring	CS4293 & CS6290, Topics on Computer Security
2020 Fall	CS5222, Computer Networks and Internets
2020 Spring	CS4296 & CS5296, Cloud Computing
2019 Fall	CS2311, Computer Programming

PROFESSIONAL SERVICES

Artifact Evaluation Committee	MLSys 2023, ACM EuroSys 2023, ACM CoNEXT 2022, USENIX OSDI 2023, USENIX ATC 2023
Technical Program Committee	IEEE IJCNN 2023

TECHNICAL SKILLS

Languages	C++, Python, Bash, Go, LaTeX
Operating Systems	Linux/UNIX
MLSys	PyTorch, MXNet, DeepSpeed, NCCL, CUDA