Xiaonan LUO

✉ xluobd@connect.ust.hk
☎ +852 5574-8799

Education

Bachelor of Science — *The Hong Kong University of Science and Technology*
Major in Computer Science & Data Science — *September 2020 – June 2024*

- *GPA: 3.7/4.3 (Around 5% in CS Dept)*
- *Graduate Courses: Optimization, Computer Vision*

Exchange — *Northwestern University*
2022 Fall
Graduate Courses: Operating System, Machine Learning

Publication

- **Xiaonan Luo***, Yichao Fu,* Cheng Wan, Zhifan Ye, Yingyan Lin. *VR-BNS: Variance Reduction for Boundary Nodes Sampling for full-graph training.* (In preparation)
- Minchen Yu, Ao Wang, Dong Chen, Haoxuan Yu, **Xiaonan Luo**, Zhuohao Li, Wei Wang, Ruichuan Chen, Dapeng Nie, Haoran Yang. *FaaSwap: SLO-Aware, GPU-Efficient Serverless Inference via Model Swapping*. arXiv:2306.03622 (Submitted to EuroSys' 24)

Research Experience

Topic in DLRM & CXL-related architecture design — *2023 Summer*
Advised by Prof. Yufei Ding in UCSD — *UCSD, U.S.*

- Research in the area of system/architecture design for memory-intensive DLRM. The project aims to alleviate memory pressure while minimizing training latency overhead. A CXL-GPU heterogeneous memory-tiered system is proposed.
- Design CXL-featured cache mechanism by leveraging the granularity of the CXL-enabled system to mitigate inter-device communication
- Propose a comprehensive memory allocation algorithm(comprehensive compared with SOTA works) over different memory hierarchies to minimize embedding lookup latency

VR-BNS: Variance Reduction for Boundary Nodes Sampling for full-graph training — *2023 Spring*
Advised by Prof. Yingyan Lin in Georgia Institute of Technology — *Gatech, U.S.*

- Research in the area of GNN training optimization. The project is closely related to BNS-GCN, a boundary node sampling based training framework to reduce memory footprint and communication volume. A new approximate computation algorithm will be used to reduce feature variance, induced by the insight of history aggregation embeddings.
- Re-design GAT computation algorithm, followed by the insight of history aggregation embedding to approximate feature prediction under full-graph training.
- Implemented GAT computation and training process, stabilized computation over training iteration with synchronized normalization.
- In addition to sampling-based memory reduction, included tensor compression technique to further reduce memory footprint on accelerators.

FaaSwap: SLO-Aware, GPU-Efficient Serverless Inference via Model Swapping *Jan 2023 – May 2023*
Advised by Prof. Wei Wang in HKUST *HKUST, HK*

- Research in the area of ML inference optimization in the context of serverless computing. The project aims to improve accelerator utilization under latency-aware inference. Submitted to EuroSys 24': FaaSwap
- Design of GPU remoting, model swapping, memory management, asynchronized server-client communication, and scheduling algorithm technique
- Conducted experiments on classical models inference performance(e.g. BERT, Resnet) with factors of GPU remoting(sync or async), model swapping, pipelining(PCIe or NVLink)

Professional Experience

Software Engineer Intern *2022 Summer*
Meituan *Beijing, China*

- Implemented Meituan Network Automatic Platform(MNAP) for switch operation and maintenance

Skills

Coding	C++, Python, Golang
Framework	PyTorch, DGL
Language	Fluent in English, Native Mandarin Chinese