MINGYU GUAN

(+1)470-334-6144 · my.guan@outlook.com

EDUCATION

Georgia Institute of Technology Aug. 2019 - Present
Ph.D in Computer Science (GPA:4.0) **Atlanta, GA**

The Chinese University of Hong Kong(CUHK) Aug. 2017 - May. 2019
B.S. in Computer Science with Honours, First Class (GPA:3.5) **Hong Kong, China**

Sun Yat-Sen University(SYSU) Sep. 2015 - Jun. 2017
B.S. in Electronic Information Science (2+2*) (GPA:4.0) **Guangzhou, China**
*A joint program offered by SYSU and CUHK.

RESEARCH EXPERIENCE

Graduate Research Assistant May. 2020 - Present
Advisors: Taesoo Kim and Anand Iyer *Georgia Institute of Technology, Atlanta*

- Built graph deep learning and graph processing systems for real-world graphs with billions of edges;
- Built general ML training and serving systems optimized for large-scale data and models;
- Leveraged machine learning techniques such as GNNs and Generative AI (LLMs) to cross-domain problems such as blockchain analytics and cloud security.

Research Intern May. 2022 - Aug. 2022
Mentor: Jay Stokes *Microsoft Research, Redmond*

- Designed and implemented a novel heterogeneous Graph Neural Network (GNN) for compromised email detection, which encodes heterogeneity of graphs efficiently by considering both path and hop information;
- Outperformed state-of-the-art solutions in terms of accuracy and scalability;
- Cooperated with a research team and a product team to construct heterogeneous graphs from a large-scale noisy enterprise email data set and built an automatic system for detecting compromised email accounts.

Undergraduate Research Assistant May. 2018 - Apr. 2019
Advisor: James Cheng *The Chinese University of Hong Kong, Hong Kong*

- Supported Distributed Online Analytical Processing (OLAP) on Husky, which is a general-purpose distributed computing system developed by the system laboratory at CUHK;
- Used the platform of Husky to implement the By-Layer cubing algorithm in Apache Kylin;
- Implemented SQL engine and customized query optimization rules on Husky using Apache Calcite.

SELECTED PROJECTS

Model Training Provenance with Confidential Computing March. 2024 - Present

- Identified the performance and privacy challenges in model training provenance;
- Designed proof generation and verification protocol with low computation and memory overhead with secutity guarantees;
- Implement proof of concept on Intel TDX, supporting latest LLMs including Bert and Llama family models.

High-throughput Inference for Early-exit LLMs May. 2023 - Sep. 2024

- Enabled batching techniques for existing early-exit (EE) LLMs, e.g. CALM (based on google T5);
- Extended non-EE LLMs and compressed models, e.g., Llama family models and distilled Bert, to their early-exit counterparts;
- Implemented Griphook's techniques to autoregressive LLMs;
- Accelerated goodput of early-exit model inference for autoregressive LLMs (2.8-3.8x) and compressed models (1.67x).

Distributed System for Dynamic Graph Neural Networks May. 2021 - Dec. 2023

· Supported efficient dynamic GNN training in large-scale distributed settings;
· Leveraged computational structure in the GNN-RNN approach to propose cross-layer optimizations;
· Enabled efficient distributed training that reserves both structure and time dependencies in dynamic graphs;
· Outperformed existing state-of-the-art GNN frameworks by up to 10.7x on a number of dynamic GNN architectures and workloads.

Processing Billion-scale Dynamic Graphs on a Single Machine Jan. 2020 - Jul. 2021

· Introduced the design of cell abstraction, allowing a significant reduction in overall storage space as well as enabling a simple, yet effective load-balancing strategy;
· Proposed an API and execution model tailored for streaming graphs by incorporating a hybrid edge- and vertex-centric API coupled with the *edgeChanged* API to allow a timely reaction to graph changes;
· Designed a technique for concurrent analytics on streaming graphs, which fully exploits the similarities in data access among concurrent graph processing jobs.

Automating Massively Parallel Heterogeneous Computing Jan. 2020 - May. 2021

· Modeled input program as a hierarchical data flow graph (HDFG) to perform a set of graph-based operations and transformations for automatic optimization and parallelization;
· Performed purity checking automatically by traversing abstract syntax tree(AST) module;
· Inferred types of variables and objects automatically with both static analysis and dynamic analysis.

PUBLICATION AND PREPRINTS

1. Anand Iyer, **Mingyu Guan**, Yinwei Dai, Rui Pan, Swapnil Gandhi, and Ravi Netravali. Improving DNN Inference Throughput Using Practical, Per-Input Compute Adaptation. *In Proceedings of the 30th Symposium on Operating Systems Principles (SOSP)*, Austin, TX, USA, Nov 2024.
2. **Mingyu Guan**, Jack W. Stokes, Qinlong Luo, Fuchen Liu, Purvanshi Mehta, Elnaz Nouri, and Taesoo Kim. HetTree: Heterogeneous Tree Graph Neural Network. *arXiv preprint* arXiv:2402.13496, 2024.
3. **Mingyu Guan**, Anand Padmanabha Iyer, and Taesoo Kim. DynaGraph: Dynamic Graph Neural Networks at Scale. *In Proceedings of the 5th ACM SIGMOD Joint International Workshop on Graph Data Management Experiences & Systems and Network Data Analytics (GRADES-NDA)*, Philadelphia, PA, June 2022.

SERVICES

- **Artifact Evaluation Committee**, SOSP '24.
- **External Review Committee**, ATC '24.

HONORS AND GRANTS

2019 **Deans List**, CUHK Hong Kong, China
2019 **Rev Mak Shuet Kwong Memorial Scholarship**, CUHK Hong Kong, China
2017 **First Prize Academic Scholarship**, SYSU Guangzhou, China
2016 **Second Prize Academic Scholarship**, SYSU Guangzhou, China
2016 **Jetta Scholarship for Outstanding Students**, SYSU Guangzhou, China

TEACHING EXPERIENCE

Graduate Teaching Assistant Georgia Institute of Technology, Atlanta

· CS8803 Systems for AI: Large Language Models, Spring 2024
· CS3251 Computer Networking, Spring 2020

SKILLS

Language C++, C, Python, SQL
Frameworks PyTorch, TensorFlow, JAX/Flax, DGL, PyG, gRPC, Hadoop
Tools LaTeX, Docker, Git, OpenAI API