Ruihang **Lai**

ruihangl@cs.cmu.edu | ruihanglai.com | MasterJH5574

Education

Carnegie Mellon University — *Pittsburgh, United States*

Ph.D. in Computer Science — *Aug. 2022 - Present*

- Advised by Prof. Tianqi Chen and Prof. Todd C. Mowry.
- Member of Catalyst Group.

Shanghai Jiao Tong University — *Shanghai, P.R. China*

B.Eng. in Computer Science — *Sep. 2018 - Jun. 2022*

- Member of ACM Honors Class, a pilot CS program for top talented students.
- Advised by Prof. Yong Yu.

Research Interests

- **Machine Learning Systems (Systems for Large-Scale Workloads)**
- **Machine Learning Compilation**
- **Systems for Emerging Computation (E.g., Sparse Computing)**

Experiences

OctoML — *Jun. 2022 - Aug. 2022*

Research Intern

- Working on the Machine Learning Compilation online course, as a teaching assistant.

Catalyst Research Group, Carnegie Mellon University — *Oct. 2021 - Dec. 2021*

Research Intern

- Worked on automatic tensor program optimization and sparse tensor computing.
- Advised by Prof. Tianqi Chen.

Publications

SparseTIR: Composable Abstractions for Sparse Compilation in Deep Learning [arxiv][code] — *ASPLOS 2023*

Zihao Ye, *Ruihang Lai*, Junru Shao, Tianqi Chen, Luis Ceze

TensorIR: An Abstraction for Automatic Tensorized Program Optimization [arxiv] — *ASPLOS 2023*

Siyuan Feng, Bohan Hou, Hongyi Jin, Wuwei Lin, Junru Shao, *Ruihang Lai*, Zihao Ye, Lianmin Zheng, Cody Hao Yu, Yong Yu, Tianqi Chen

Tensor Program Optimization with Probabilistic Programs [arxiv] — *NeurIPS 2022*

Junru Shao, Xiyou Zhou, Siyuan Feng, Bohan Hou, *Ruihang Lai*, Hongyi Jin, Wuwei Lin, Masahiro Masuda, Cody Hao Yu, Tianqi Chen

Talks

TensorIR: An Abstraction for Tensorized Program Optimization [video]

- Jan. 2022 @ System, Architecture, Machine learning, and Programming language (SAMPL) Lab.

SparseTIR: A Unified Abstraction for Sparse Workload Representation and Optimization

- Nov. 2021 @ CMU Automated Learning System (Catalyst) Group.
- Dec. 2021 @ TVM Conference 2021.

Selected Projects

MLC LLM, Web LLM and Web Stable Diffusion

Spring 2023

- Web LLM and Web Stable Diffusion bring large language models and stable diffusion models completely to people's web browsers. **Everything runs inside the browser accelerated by WebGPU with no server support.**
- MLC LLM is a universal solution that allows any language model to be deployed natively on a diverse set of hardware backends and native applications, plus a productive framework for everyone to further optimize model performance for their own use cases. **Everything runs locally with no server support and accelerated with local GPUs on your phone and laptop.**

SparseTIR, A Tensor-Level Abstraction for Sparse Operator Optimization in Deep Learning

Fall 2021 - Spring 2022

- In close collaboration with Zihao Ye from University of Washington.
- Documentation

Apache TVM, An End-to-End Machine Learning Compiler Framework

Committer

Fall 2020 - Present

- Author of over 40+ PRs, 13000+ lines of code.
- Reviewer of over 80+ PRs.

Mx-Compiler

Course Project

Spring 2020

- A toy compiler implemented in Java, from Mx* (a C- and Java-like language) to RISC-V assembly code.
- Implemented many effective optimizations. The generated code has performance close to GCC O2.
- More than 15k+ lines of code overall.

Distributed Hash Table

Course Project

Summer 2019

- Implemented two DHT protocols, Chord and Kademlia, in Go Language.
- Implemented an instant chat room system based on the Chord protocol.

Teaching

Machine Learning Compilation

Online course

Teaching Assistant

Summer 2022

- Prepare and release course assignments. Answer questions in the discussion page.

Principle and Practice of Computer Algorithms

SJTU

Leader Teaching Assistant

Summer 2020

- Advised students to implement a RISC-V simulator.
- Advised students to implement two Distributed Hash Table protocols, Chord and Kademlia, in Go Language.

Data Structure (Honor)

SJTU

Leader Teaching Assistant

Spring 2020

- Taught advanced data structures which students usually do not learn in class.
- Prepared the course assignments, projects and programming exams.

Honors & Awards

Scholarship

2022 **Shanghai Excellent Graduate Award** *Shanghai, China*

2020 **National Scholarship** (Top 0.2% nationwide) *P.R. China*

Programming Competitions

2018 **The 4th Place** The 2018 ICPC Asia Singapore Regional Contests *Singapore*