ASF MEMBER | APACHE TVM PMC MEMBER | SJTU PH.D. CANDIDATE me@syfeng.net | # syfeng.net | GitHub | in LinkedIn

Education

• Ph.D. Candidate in Computer Science at School of Electronic, Information and Electrical Engineering

- Research Area: Machine Learning System, Machine Learning Compiler
- Advisor: Prof. Yong Yu and Prof. Weinan Zhang

Ph.D. Candidates - Shanghai Jiao Tong University

- Research Assistant at APEX Data & Knowledge Management Lab
- Working closely with Prof. Tianqi Chen, on Apache TVM and MLC community

Research Intern - University of Washington

- Research Intern at SAMPL Lab
- Advisor: Prof. Luis Ceze and Prof. Tianqi Chen
- Better tensorization support for Apache TVM, initializing TensorIR project

| Bachelor of Engineering | Shanghai Jiao To | ong University | |
|-------------------------|------------------|----------------|--|
| | | | |

- B.Eng. in Computer Science and Technology (Zhiyuan Honored Program)
- Member of ACM Honors Class, which is an elite CS program for top 5% talented students

Open Source Projects

Apache TVM - **○** GitHub ★ 11.0k ¥ 3.3k

- Open source machine learning compiler, enabling deployment models on diverse hardware backends
- \circ Leading TensorIR project, the next generation Tensor-level IR for tensor hardware.
- Co-leading TVM Unity/Relax project, the next generation Graph-level IR for dynamic models.
- Contributing to several key features: TVMScript, Meta-Schudule, runtime, frontend
- Serving in Apache TVM Program Management Committee (PMC)

MLC-LLM - 🗘 GitHub ★ 16.1k 1.2k

- Compile LLMs and depoly models natively on every device
- Supported hardware: NVIDIA GPU, AMD GPU, Apple GPU, mobile GPUs and Intel iGPU
- Supported runtime: CUDA, ROCm, Metal, Vulkan, OpenCL, WebGPU
- Support distributed inference for large models on CUDA and ROCm
- Support LLM serving with OpenAI API compatibility

Web-LLM - ♥ GitHub ★ 8.7k 9.5k

• Bringing large-language models and chat to web browsers with local GPU capabilities.

- Leverage emerging WebGPU API to run LLMs inside the browser
- Serving as a backend runtime of MLC-LLM

Open Source Community

| Apache Software Foundation Member - ASF | Mar 2024 - Present |
|--|--------------------|
| Apache TVM PMC Member Committer - Apache TVM | Mar 2022 - Present |
| Leading the TVM and ML Compiler community in China | |

- Organized 4 TVM Meetups in 2023 at Shanghai, Beijing and Shenzhen
- Have been nominating individuals for 14 positions as reviewers, committers, or PMC members.
- Have been actively contributing and reviewing code for more than 3 years



2023 - Present

2019 - Present

2023 - Present

Sep 2020 - Present

Jul 2019 - Dec 2019

Sep 2016 - Jul 2020

Served as TA at the open source MLC course

• Prepared the tutorials in English and Chinese for MLC community

Publications

Relax: Composable Abstractions for End-to-End Dynamic Machine Learning - Preprint2023Ruihang Lai[†], Junru Shao[†], Siyuan Feng[†], Steven S. Lyubomirsky[†], Bohan Hou, Wuwei Lin, Zihao Ye, Hongyi
Jin, Yuchen Jin, Jiawei Liu, Lesheng Jin, Yaxing Cai, Ziheng Jiang, Yong Wu, Sunghyun Park, Prakalp Srivastava,
Jared Roesch, Todd C. Mowry, Tianqi Chen2023TensorlR: An abstraction for automatic tensorized program optimization - ASPLOS2023Siyuan Feng[†], Bohan Hou[†], Hongyi Jin, Wuwei Lin, Junru Shao, Ruihang Lai, Zihao Ye, Lianmin Zheng, Cody Hao
Yu, Yong Yu, Tianqi Chen2023Effectively Scheduling Computational Graphs of Deep Neural Networks toward Their Domain-
Specific Accelerators - OSDI2023Jie Zhao, Siyuan Feng, Xiaoqiang Dan, Fei Liu, Chengke Wang, Sheng Yuan, Wenyuan Lv, Qikai Xie2023

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

CityFlow: A Multi-Agent Reinforcement Learning Environment for Large Scale City Traffic Scenario - WWW 2019

Huichu Zhang, **Siyuan Feng**, Chang Liu, Yaoyao Ding, Yichen Zhu, Zihan Zhou, Weinan Zhang, Yong Yu, Haiming Jin, Zhenhui Li

Invited Talks

| Apache TVM: Unity Strategy for Emerging Needs - TVM Meetup Shanghai | Dec 2023 |
|---|----------|
| Opportunities and Challenges of ML Systems in the Era of LLMs - TVM Meetup Shenzhen | Sep 2023 |
| Bringing LLMs to Everywhere via ML Compilation - CommunityOverCode Asia, XiaoMi Inc. | Jun 2023 |
| TensorIR: An abstraction for automatic tensorized program optimization - ASPLOS, TVMCon | Mar 2023 |
| A Unified Abstraction for Machine Learning Compilations - TVM Meetup Shanghai | Mar 2023 |
| Machine Learning Compilation - GAIDC | Feb 2023 |