Portfolio
Github

Runyu Lu

AI HPC Compiler

lry89757.github.io
lry89757@gmail.com

Hope: Applying for a Ph.D. program in the field of **AI Compiler** in the United States.

Skills

AI	Familiar with training and deploying common CV Models
HPC	Have developed many high performance neural network operators
Compiler	Optimizing the backend LLVM and interested in the DL Compiler Infra like TVM, MLIR

Education

Bachelor of Computer Science Huazhong University of Science and Technology, GPA: **3.96/4.00** **Sept. 2020 — June 2024**

Academic Experience

WhiteFox, Fuzzing University of Illinois Urbana-Champaign **July. 2023 — Augu. 2023**

- Role: Research Intern
- Mentor: Chenyuan Yang Jiawei Liu Advisor: Lingming Zhang
- Responsible for the LLVM part of this project. Use LLMs to infer what kind of test inputs could trigger the optimization in the compiler based on the pattern written in the source code

Explore More Efficient PMA/PCSR Dynamic Graph Structure HUST **Nov. 2022 — June. 2023**

- Role: Research Intern, Co-author
- Mentor: Hongru Gao Advisor: Zhiyuan Shao, Hai Jin
- Based on the current dynamic graph storage formats of PMA/CSR, a more dynamic-graph-friendly data storage format is proposed, which involves modifications to the operating system kernel

Industrial Experience

Optimize the Backend of LLVM, Sensetime Company **April 2023 — Augu.2023**

- Role: LLVM Backend Developer
- Mentor: Wenqiang Yin
- Optimize the llvm backend based on the Self-Develop TPU of Sensetime
- ISA just like NV PTX
- Instruction Selection, Instruction Pattern Match
- CodeGen Emitter, GPU Compiler Optimization

Develop High Performance Neural Network Inference Engine, Tencent Company **July 2022 — Nov. 2022**

- Role: Top 15 committer of 253(util Nov.2022),
- Mentor: nihui, with **6k** followers in Github
- Write and Optimize high performance operators and math library for ncnn, an open source project with **17.6k** stars in Github, mainly aligned with pytorch, some examples I built: GridSample: Given an input and a flow-field grid, computes the output using input values and pixel locations from grid. To be noted, the PNNX of ncnn, a new PyTorch Neural Network eXchange, draw on the design concept of MLIR GELU: Implement sse/avx/avx512 version of gelu, with a fast version of erfc.

Deploy High-FPS AI Models on Arm Chips, FiberHome Telecommunication Company **Dec 2021 — June 2022**

- Role: **Leader**
- Mentor: Yayu Gao
- Arm CPU / 20FPS / Snapdragon 870
- YOLOX/Lite-HRNet
- pattern match algorithm/Hungarian Algorithm

More Info

For better reading experience and more detailed information, you're welcome to visit lry89757.github.io :)