Github
Transcripts

Runyu Lu

AI HPC Compiler

lry89757.github.io
lry89757@gmail.com

Hope: Applying for a PhD program in the field of MLsys, especially about LLM Serving and Compilers.

Education

Bachelor of Computer Science Huazhong University of Science and Technology, GPA: **3.95/4.00** **Sept. 2020 — June 2024**

Academic Experience

LLM Serving, Inferencing and Profiling University of California San Diego **Aug. 2023 — Present**

- Role: Research Intern
- Mentor: Hao Zhang , Advisor: Hao Zhang
- Profiling the bottleneck of current SOTA LLM Inferencing framework(vllm, ppl.llm). And now coding for one project about acclerating the serving throughtput of LLMs.

WhiteFox, LLM Fuzzing LLVM University of Illinois Urbana-Champaign **June. 2023 — Sept. 2023**

- Role: Research Intern, **Third author**, paper already submitted to **FSE'24**
- Mentor: Chenyuan Yang , 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

Efficient Paged Dynamic Graph Structure HUST **Oct. 2022 — June 2023**

- Role: Research Intern, **Co-first author**, paper will be submitted to **ICDE'24**
- Mentor: Hongru Gao, Advisor: Zhiyuan Shao, Hai Jin
- Based on the the memory bound of graph processing SOTA algorithm, a more efficient dynamic-graph-friendly data storage format is proposed, which involves modifications to the page table operating system kernel(Linux).

Industrial Experience

Optimize the LLVM Backend of SenseTime TPU, GPU Compiler, Sensetime Company **April 2023 — Aug.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 **18k+** stars in Github, mainly aligned with pytorch, specificially accleration for X86/ARM/RISCV/GPU platforms.

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

- Role: **Leader**
- Mentor: Yayu Gao, Xinggang Wang
- YOLOX/Lite-HRNet, 20FPS on Arm CPU Snapdragon 870

Skills

AI	Familiar with deploying common LLM and CV Models on device like GPU, CPU(x86,arm)
HPC	Have developed many high performance neural network operators(CUDA, C++, Asm)
Compiler	Optimized the backend of LLVM and familiar with DL Compiler Infra like MLIR, Triton

More Info

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