



≜Website lry89757@gmail.com

Vision: 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

- · Advisor: Prof. Hao Zhang
- · Role: Research Intern
- Profiled the bottleneck of current SOTA LLM Serving framework(e.g., vllm, ppl.llm).
- Improve the GPU SM utilization to accelerate the serving throughtput of LLMs.

WhiteFox: White-box Compiler Fuzzing via LLMs University of Illinois Urbana-Champaign

June. 2023 — Sept. 2023

- · Advisor: Prof. Lingming Zhang
- Role: Research Intern, Third author, paper already submitted to FSE'24
- Use LLMs to generate examples to trigger the optimization in the compiler
- Detect the flaws of Pytorch, XLA, LLVM based on the pattern written in the source code

Efficient Paged Dynamic Graph Serving HUST

Oct. 2022 — June 2023

- Advisor: Prof. Hai Jin, Prof. Zhiyuan Shao
- Role: Research Intern, Co-first author, paper will be submitted to ICDE'24
- Remap the PageTable of OS Kernel and propose a new B+ Tree-based memory algorithm to accelerate the dynamic graph processing system.
- Speed up existing SOTA algorithms by more than 10x times.

INDUSTRIAL EXPERIENCE

Optimize the LLVM Backend of SenseTime TPU, GPU Compiler, Sensetime Company

April 2023 — Aug.2023

• Role: LLVM Backend Developer

Shanghai. China

- · Mentor: Wengiang Yin
- GPU Compiler Optimization, Instruction Selection, Instruction Pattern Match, CodeGen Emitter

Develop High Performance Neural Network Inference Engine, Tencent Company

July 2022 — Nov. 2022

• Role: **Top 15** committer of 263(util Nov.2022)

Shenzhen, China

- Mentor: nihui, with 6k+ followers in Github
- Optimize high performance operators and math library for ncnn, 18k+ stars in Github, a neural network library handcraftly optimized 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 Wuhan, China

• Deploy YOLOX/LiteHRNet on Snapdragon 870(Arm CPU), Achieve 20 FPS.

SKILLS

ΑI LLM/CV Model Deployment

CUDA, Intel SSE, Arm NEON, Assembly **HPC** Compiler Compiler Infra like LLVM, MLIR, Triton

More Info

For better reading experience and more detailed information, please feel free to visit my \(\overline{\pi} \) website :)