

Vision: Applying for a PhD program in the field of MLsys, especially improve the latency and throughput of LLM Serving and the instruction scheduling of ML Compilers.

EDUCATION

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

PUBLICATIONS

WhiteFox: White-box Compiler Fuzzing via Large Language Models, Under Review, Arxiv soon **FSE'24**

• Authors: Chenyuan Yang, Yinlin Deng, **Runyu Lu**, Jiayi Yao, Jiawei Liu, Reyhaneh Jabbarvand, Lingming Zhang

Efficient Memory Management for Large Dynamic Graph Serving with PagedMapping, To be submitted **ICDE'24**

• Authors: *Hongru Gao, ***Runyu Lu**, Zhiyuan Shao, Hai Jin

ACADEMIC EXPERIENCE

LLM Serving, Inferencing and Profiling University of California San Diego 

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

LMSYS Lab, Aug. 2023 — Present

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

- Research Intern advised by **Prof. Lingming Zhang**
- Test optimization in compilers with white-box fuzzing technique by leveraging LLMs
- Detect 80 bugs of Pytorch, TensorFlow XLA, TensorFlowLite, LLVM based on the optimization source code

ISE Lab, June. 2023 — Sept. 2023

Efficient Paged Dynamic Graph Serving Huazhong University of Science and Technology 

- Research Intern advised by **Prof. Hai Jin, Prof. Zhiyuan Shao**
- Remap the PageTable of OS Kernel to accelerate the dynamic graph processing system.
- Speed up existing SOTA algorithms by more than 10x times.

CGCL Lab, Oct. 2022 — June 2023


INDUSTRIAL EXPERIENCE

Optimize the LLVM Backend of SenseTime TPU, GPU Compiler Sensetime , Shanghai.China 

- Role: LLVM Backend Developer
- Mentor: Wenqiang Yin
- GPU Compiler Optimization, Instruction Selection, Instruction Pattern Match, CodeGen Emitter

April 2023 — Aug.2023

Develop High Performance Neural Network Inference Engine Tencent , Shenzhen.China 

- Role: **Top 15** committer of 263(util Nov.2022)
- Mentor: nihui, with **6k+** followers in Github
- Optimize high performance neural network operators and math library for NCNN , **18k+** stars in Github, handcraftly optimized for X86/ARM/RISCV/GPU platforms.

July 2022 — Nov. 2022

Deploy High-FPS AI Models on Arm Chips FiberHome , Wuhan.China 

- Role: **Leader** of HUST.Dian.AI Group
- Mentor: Yayu Gao, Xinggang Wang
- Deploy YOLOX/LiteHRNet on Snapdragon 870(Arm CPU), Achieve 20 FPS.

Dec. 2021 — June 2022

SKILLS

AI	LLM/CV Model Deployment
HPC	CUDA, Intel SSE, Arm NEON, Assembly
Compiler	Compiler Infra like LLVM, MLIR, Triton

MORE INFO

For better reading experience and more detailed information, please feel free to visit my website :)