Shuo Yang

andy_yang@sjtu.edu.cn https://andy-yang-1.github.io/

Education

Shanghai Jiao Tong University Shanghai, China

Bachelor of Computer Science *Sept. 2020 - Present*

- Member of ACM Honors Class, which is an elite CS program for top 5% talented students
- Member of LMSYS
- GPA (All): 91.53/100, **Ranking: 2/36**
- Scores of some courses:
  - Mathematical Analysis (MATH1607H): 97/100, Ranking: 2/36
  - Data Structure (CS1951): 98/100, Ranking: 1/36
  - Algorithm Design And Analysis (CS3950): 98.40/100, Ranking: 1/36
  - Computer Operating Systems (CS2952): 95/100, Ranking: top 5
  - Computational Complexity (CS3954): 97/100
  - Graph Theory and combinatorics (CS2962): 94/100, Ranking: 3/36
  - Reinforcement Learning (CS3316): 98.4/100, Ranking: 1/36
  - Compiler Design (CS2966): 95/100

Research Interests & Technical Skills

Interests

- Machine Learning System
- Machine Learning Compiler
- Large Language Model

Skills

- Programing Languages: Proficient in **CUDA**, **OpenAI Triton**, C/C++, Python, Java, Verilog and Golang
- ML Frameworks: Experienced in PyTorch (including Dynamo and Inductor), **TVM**, TensorRT and CUDA Graph
- LLM Frameworks: Contribute to **FastChat**, **VLLM**, Megatron, LightLLM

Experience

University of California, Berkeley Berkeley, CA, USA

Research Assistant, advised by Prof. Ion Stoica *Mar. 2023 - Present*

Research Topic: Low latency inference for LLM; Fast LoRA serving system; Contamination study of LLM

Shanghai Jiao Tong University Shanghai, China

Undergraduate Researcher, advised by Prof. Tianqi Chen *July. 2022 - Mar. 2023*

Research Topic: Building fast and efficient operators and providing strategies to generate the operators

Publications

Rethinking Benchmark and Contamination for Language Models with Rephrased Samples

S. Yang*, *W. Chiang*, L. Zheng*, J. Conzalez, I. Stoica (*equal contributions)*

- Submitted to **ICML 2024**. arxiv
- We demonstrate that simple variation of test data can easily bypass existing detection methods and help models to achieve dramatically high scores. We propose LLM decontaminator and apply it to widely used dataset, revealing significant previously unknown test overlap.

S-LoRA: Serving Thousands of Concurrent LoRA Adapters

Y. Sheng, S. Cao*, D. Li, C. Hooper, N. Lee, **S. Yang**, L. Zheng, J. Gonzalez, I. Stoica(*equal contributions)*

- Submitted to **MLSys 2024**. arxiv
- S-LoRA employs a novel tensor parallelism strategy and highly optimized custom CUDA kernels for heterogeneous batching of LoRA computation. Collectively, these features enable S-LoRA to serve thousands of LoRA adapters on a single GPU or across multiple GPUs with a small overhead.

Projects

LLM Decontaminator

Decontamination system for language model datasets. (LMSYS's 4th Project)

Code for the paper "Rethinking Benchmark and Contamination for Language Models with Rephrased Samples"

OpenAI Compatible Server

Merged into FastChat

Effortlessly port applications based on OpenAI to open-source alternatives without modifying the code.

An SIMT matrix multiplication generating template for TVM

Instructed by Tianqi Chen, Bohan Hou and Siyuan Feng

The template implements double buffer which enable a two-stage pipeline. It achieved a full catch-up on cutlass under regular workload.

A RISC-V CPU implemented in Verilog

SJTU ACM Class Architecture 2021 Assignment (MS108 Course Project)

The cpu employs Out of Order execution and a methodical debugging approach to do away with waveform graphs.

A Compiler from Mx* language (which is a C++ & Java like language) to RV32I Assembly

SJTU ACM Class Compiler Design and Implementation 2022 Assignment (MS208 Course Project)

The whole project went from Semantic to LLVM IR and finally to RV32I Assembly, for which I designed an efficient high-dimensional dynamic array approach that makes the application efficient.

Honors & Awards

Scholarship

- 2020, 2021, 2022 Zhiyuan Honorary Scholarship
- 2021 Ruiyuan-Hongshan Scholarship
- 2022 Hanying-Juhua Scholarship

Other Experience

Great Minds in Computer Science CS1950 Shanghai, China

Teaching Assistant *June. 2021 - Dec. 2021*

Probability Theory MATH2701 Shanghai, China

Teaching Assistant *June. 2022 - Dec. 2022*