Songrun Xie

☎ (+86)13669971205 ✉ sory.xie@gmail.com soryxie

Research Interest

I am generally interested in *distributed systems* and *machine learning systems*. Recently, I have been working on improving the efficiency and scalability of machine learning systems by optimizing communication and caching strategies, as well as enhancing operator performance.

Education

Tongji University **2021 - 2025 (expected)**
B.S. School of Computer Science *GPA: 91/100*

Publication

[1] Morpheus: High-Throughput Framework of Diffusion Model with Distributed Cache Scheduling.
Songrun Xie, Chengsong Zhang, Ping He, Songhao Zhang, Zhengzhong Tu, Fan Lai.

[2] **(COLM'24)** Beyond Correctness: Exercising Language Models for Efficient Code Generation.
Jiawei Liu, *Songrun Xie*, Junhao Wang, Yuxiang Wei, Yifeng Ding, Lingming Zhang. paper ◇ code

Experience

Research intern, GAEA Lab at UIUC **Apr. 2024 - Now**
Advised by Prof. Fan Lai *Topic: Machine Learning System*

- This system is based on our findings: for image generation, we can split an image into multiple parts, which can be generated in parallel and with different levels of complexity (different denoising steps). Thereby, we proposed an distributed *inference system* for diffusion models (like U-Net, DiT), in which we divide each generation task into patches and schedule sub-tasks to maximize the overall good output.
  ➢ Role: Led and scheduled team members for the project, and designed and implemented core components, such as operators (e.g., attention and convolution) capable of handling *sparse* input tensors, as well as a distributed cache manager.

Research intern, iSE Lab at UIUC **Apr. 2023 - Mar. 2024**
Advised by Prof. Lingming Zhang *Topic: Software Engineering*

- **Evaluating LLM code efficiency**: We built an evaluation methodology, namely *Differential Performance Evaluation* (DPE), which rigorously evaluates LLM-generated code via synthetic performance-exercising test input and a stable and intuitive compound efficiency metric. Specifically, I contributed to the framework implementation, such as CPU-instruction-based profiling and performance-exercising input selection.
- **Testing ML Systems**: Prior to EvalPerf, I explored the topic of ML system testing and contributed to NNSmith, a solver-aided random model generator to test ML frameworks. Specifically, I added an oracle in NNSmith to support gradient checks, detecting 2 high-priority bugs in the PyTorch framework.
- Besides research explorations, I also learned and practiced how to present and advertise research outcomes by building academic project pages. Specifically, I co-designed the leaderboard pages for the EvalPlus/EvalPerf project, where my template has been adopted by various projects, such as CRUXEval from MIT and MHPP from Edinburgh.

Research intern, ByteDance **May. 2023 - Oct. 2023**
TikTok Infrastructure *Topic: Cloud Performance, Stability*

- Responsible for identifying *performance bottlenecks* and optimizing data-intensive services, developing the *anomaly detection* system and *service call graph system*. Mainly, I actively contributed to Feed and Pack services, the backbone (500k+ LoC) for packaging and forwarding data streams on the TikTok server.

Projects

- **NEMU+NanOS**: a simple but complete full-system x86 emulator designed for teachings. Many x86-specific programs can run on NEMU. And a simple operating system built on it
- **SimpleC Compiler**: a compiler implemented with Python, can transform C code into LLVM IR and, subsequently, multiple assemblies.

Skill Set

I am experienced with C++, C, CUDA, Python, Pytorch, Go, JavaScript, parallel computation, machine learning systems, and distributed systems