傅申

(+86) 157-7969-2697 fushen@mail.ustc.edu.cn Fr4nk1inCs

教育经历

中国科学技术大学，计算机科学与技术专业，计算机系统结构方向，硕士在读 2024.09 – 至今

- 实验室：先进数据系统实验室（ADSL），导师：李诚副教授
- 研究方向：人工智能系统，特别是面向 MoE 模型的系统设计与优化
- GPA: 4.13/4.30

中国科学技术大学，少年班学院，计算机科学与技术专业，本科 2020.09 – 2024.06

- GPA: 3.92/4.30 (前 8%)

研究经历

动态 Top-K 路由 MoE 模型推理并行策略自动搜索 2025.05 – 2025.08

- 对于采用动态 Top-K 路由的 MoE 模型，针对其不同层之间计算开销存在的差异，设计并实现了一种并行策略自动搜索方法，将 Prefill-Only 任务和 Prefill 密集型任务的推理吞吐量分别提升至多 31% 和 16%
- 负责动态 Top-K 路由在 SGLang 中的适配与实现，参与了模型延迟分析器的实现、并行策略搜索算法的设计以及实验评估工作
- 相关论文已提交至 AAAI 2026

MoE 模型专家并行 All-to-All 通信去冗余 2024.10 – 2024.12

- 针对训练 MoE 模型时，专家并行 All-to-All 通信中存在的冗余跨机数据传输问题，设计并实现了一种去冗余方法，将跨机流量转换为机内跨卡流量，提升模型端到端训练速度至多 33%
- 提出了一种基于匈牙利算法的通信调度方法，以最小化机内跨卡通信开销
- 参与了 Megatron-LM 框架中去冗余方法的实现，对关键模块进行了性能优化

所获奖项

- 2023 年度蔷薇远志奖学金
- 2022 年度江淮蔚来汽车奖学金
- 2021 年度陈林义奖学金

其他经历

- 开源贡献：**sgl-project/sglang PR #6121**（为 Qwen 2/3 MoE 模型添加 DP Attention 支持）
- USENIX ATC '25 Artifact Evaluation Committee 成员
- 中国科学技术大学 2023 秋季学期编译原理与技术课程助教

专业技能

- **编程语言**：Python、C/C++、Lua、Shell
- **框架**：PyTorch、SGLang