王易涛

(+86) 153-2458-2149 · Email: wangyitao2001@163.com · **求职意向：AI Infra** · GitHub: walker-ai

教育经历

东南大学 (985), 网络空间安全, *硕士* 2023.9 - 2026.6

东南大学学业二等奖学金 (2 次), 研究生模拟国际会议展演三等奖

南京林业大学 (双一流), 物联网工程, *本科* 2019.9 - 2023.6

GPA: 4.1/4.5, RANK: 2/105(1.9%), Mathorcup 数学建模挑战赛全国一等奖, 美国大学生数学建模二等奖 (H 奖), 团体程序设计天梯赛团体三等奖, 蓝桥杯 C++ 程序设计省级一等奖, JSCPC'21 RANK 30%, 南京林业大学优秀毕业生, 南京林业大学三好学生 (4 次), 南京林业大学一等奖学金

实习经历

蚂蚁集团（杭州）| 基础研发工程师, Mentor: Peng Zhang 2025.5-2025.8

- 参与维护 AWQ, GPTQ, Compressed-tensor 等 W4A16 量化方法的集成, 支持 AI 推理云 H20 推理性能优化相关目标
- 独立负责 SageAttention 适配 SGLang 需求, 测试 FP8 的 SageAttention 性能分析, 预期将在内部平台和开源社区同步上线

科研经历

面向移动端设备的 LoRA-based LLM 高效推理技术 ***ACL 2025, CCF-A***

- 该工作主要聚焦于端侧设备上的 LLM 推理系统的 KV cache 优化，旨在减少 LLM 部署时的延迟和内存占用；其基于 SGLang 框架进行二次开发，是首个提出针对不同 LoRA 进行 KV cache 存储，以及考虑 LLM 应用上下文进行联合优化的 LLM 推理工作
- 本人独立完成整个系统的开发，以及前期调研、方案设计和部分论文写作部分，已被 ACL 2025 主会（CCF-A）接收。另同时在投一篇 CCF-A 期刊论文

开源经历

sgl-project/sglang https://github.com/sgl-project/sglang

- 参与 SGLang 量化板块中 AWQ / GPTQ 的维护，集成 AWQ / GPTQ 相关 CUDA kernel，并移除 vllm 依赖，迁移算子到 SGLang 中，并使用 Nsight system, Nsight compute 及 Torch Profiler 等工具分析在不同并发程度下的性能
- Related Issue & PR: #6312, #6842, #5639

项目经历

CMU10-414/714: Deep Learning System (Zico Kolter, Tianqi Chen) 2023.7-2023.10

- 该项目来自于 CMU 公开课 Deep Learning System，要求实现一个简易版的深度学习框架（Needle），内容包括：计算图，自动微分系统，优化器（SGD，Adam）；实现常用算子包括 softmax，conv，gemm 等及其前向计算函数；实现硬件加速相关包括 Ndarray Backend, CPU / CUDA Backend；利用实现好的功能实现一些简易的神经网络模型包括 CNN、RNN、LSTM、Transformer 等；
- 项目地址：https://github.com/walker-ai/CMU10714-Fall2022

技能和其他

- 专业技能：熟悉 Python，C++ 等语言及 PyTorch，vLLM 和 SGLang 等训推和深度学习部署框架，熟练运用 Nsight Compute / Nsight System, Torch Profiler 等内核性能分析工具, 了解集合通信和 CUDA / Triton 编程及算子调优经验, 熟悉 Linux 开发环境, 掌握 CMake, Git, GDB 等日常开发工具
- 英语水平：CET-4（547）、CET-6（531）
- 暑期实习 offer 情况：腾讯（深圳）TEG 云架构平台部后台开发实习生（AI 芯片推理优化方向）；荣耀（南京）AI Infra 实习生；蚂蚁集团基础平台研发实习生（最终去向）