辛杰

[email] [github] [homepage]

工作经历

NVIDIA 上海
Senior Compute Architect 2022 – Present

- 研究下一代 GPU（Hopper、Blackwell、Rubin）上的端到端模型性能自动化优化
- 开发深度学习编译器（CuTile、Triton-to-CuTile）及 MLPerf 基准测试优化
- 与 PyG、Megatron-LM、OpenAI Triton 团队合作开发编译器基础设施

教育背景

华中科技大学 武汉
计算机科学与技术・硕士 2019 – 2022

华中科技大学 武汉
计算机科学与技术・学士 2015 – 2019

项目经历

Agent for Compiler 2026 – Present
设计专为编译器优化的轻量级 harness 系统。

Triton-to-CuTile 2024 – 2025
OpenAI 孵化项目；连接 Triton 前端与 CuTile 后端的编译器桥接层。 [code]
使 Triton 用户能够使用 NVIDIA 优化的 tile 编程基础设施。

CuTile 2023 – Present
NVIDIA GPU 的 CUDA Tile 编程标准。 [code] [blog]
设计并实现了深度学习工作负载的优化 pass。

NvFuser 2022 – 2023
PyTorch JIT 编译器，在 GNN 模型上实现 1.2x–2.5x 加速。 [code] [blog]
为稀疏工作负载启用了 gather/scatter/index_select 图操作。

OpenFold2 Training 2023 – 2023
MLPerf HPC v3.1 榜首，端到端加速 7.5 倍。 [code] [paper] [blog]
优化大规模 GPU 集群上的蛋白质结构预测训练。

GPT-3 Training 2023 – 2023
MLPerf Training v4.0；为 175B 模型贡献了 1.02x 训练加速。 [code]
优化大语言模型的分布式训练流水线。

PyTorch Geometric TorchScript 2022 – 2022
为 PyG 社区添加 TorchScript 支持，实现 GNN 模型的 JIT 编译。 [code]
贡献了图神经网络生产部署的核心基础设施。

SpMM on GPU – Champion, MIT/IEEE Graph Challenge 2021 – 2021
高性能稀疏矩阵乘法；相比 cuSPARSE 最高加速 652 倍。 [code] [paper]
获得 2021 年 MIT/IEEE/Amazon Graph Challenge 冠军。

当前方向

我的研究兴趣已全面转向 AI Agent ——探索 LLM 如何自主优化编译器、自动化性能调优，并变革传统软件开发流程。