杨行

(+86) 151-6143-5537 | 2679frank@gmail.com | github.com/frank2679.io

.

个人总结

- 拥有 7 年异构平台（CPU/GPU/DSP/NPU）算子加速库开发经验，专注 DNN、BLAS 等核心加速库的开发。
- 具有三年团队管理经验，成功带领团队完成交付，得到客户认可，也曾跨团队主导项目开发。
- 拥有超强的学习能力，追求卓越，喜欢钻研技术，乐于分享。

技术能力

- **技术方向**: 加速库开发、算子优化、深度学习模型部署优化（LLM，CNN），熟悉 pytorch，cutlass。
- **编程语言**: C++/C、Python、CUDA。
- **开发工具**: CMake、Verdi、Vim、Git、JIRA、Jenkins、markdown。

工作经历

国内头部 GPU 厂商 | 软件工程师 / *Team Leader* / 算子加速库 2021.01 至今

- 从零开始开发 AI 领域的加速库（DNN、BLAS），在多种大模型场景下优化性能，达到 A100 同等水平。
- 设计并实现了高效的算子注册框架和日志系统，以及 kernel selection 算法，大幅提升计算效率。
- 负责项目从设计、开发、测试到 CICD 的全流程管理，确保高质量交付。

安防领域龙头企业 | 软件工程师 / *AI* 加速库 / 商业应用 2017.07 —2020.12

- 负责异构平台高性能卷积神经网络（CNN）库的开发，优化芯片性能，提升算法落地效率。
- 从业务需求到解决方案的端到端项目负责，构建高效的应用方案，实现智能算法的实际应用。

项目经历

BLAS 支持 llama2 系列 |*llama2* 性能打平 *A100* / 专利 *2* 项 2023.12 - 2024.03

- 通过 trace 分析 llama2 系列模型不同并行策略下 BLAS 所需要优化的各种场景。
- 使能了多种优化策略，包括 warp-specialized 编程范式, fused bias, fused grad 累加，BF16 累加等。

GEMM 极致优化实现 | 芯片利用率打满 / 专利 *2* 项 2021.09 - 2021.10

- 在特定 shape 下，通过手写汇编和 Verdi 波形分析，实现了基于 vcore 和 tcore 平台的 GEMM 算法性能最优化。

DNN 框架开发 | 支撑 *DNN* 全栈验证 / 软著 *1* 项，软著 *1* 项 2022.03 - 2023.06

- 参考 pytorch 开发深度学习框架，并设计基于 YAML 的算子表达方式，实现了自动化精度验证和性能对标。

教育经历/硕士研究生

台湾大学 | 电信研究所 | 学术型硕士研究生 2014.09 —2017.01

WIFI 802.11ax 协议设计，分别在 IEEE access 期刊，及通信领域顶会 globalcom 上发表两篇论文。

重庆大学 | 通信学院 | 工学学士 2010.09 —2014.06

专业前 5%，优秀毕业生，获奖学金多次

其他信息

- **兴趣爱好**: 铁三爱好者，持续学习新技术，积极参与技术社区分享。
- AI 领域专利 5 项，软著 1 项，通信领域专利一个，论文 2 篇。

[1] **Yang, Hang**, D.-J. Deng, and K.-C. Chen, "On energy saving in ieee 802.11 ax," *IEEE Access*, vol. 6, pp. 47 546–47 556, 2018.

[2] **Yang, Hang**, D.-J. Deng, and K.-C. Chen, "Performance analysis of ieee 802.11 ax ul ofdma-based random access mechanism," in *GLOBECOM 2017-2017 IEEE Global Communications Conference*, IEEE, 2017, pp. 1–6.