黄杨

2679frank@gmail.com | github.com/frank2679.io | Linkedin/FrankYoung

个人总结

- 有六年以上异构平台 (CPU/GPU/DSP/NPU) 算子加速库经验，包括 DNN, BLAS, FFT 等。深度参与 0-1 芯片软硬件研发；
- 有两年带团队经验，团队规模 5 人；
- 超强的学习能力，追求卓越，铁三爱好者。

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。

[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.

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

技术能力

- **技术内容**: 熟悉算子优化，深度学习模型部署优化，包括 CNN, 大语言模型;
- **技术栈**: 熟悉 C++, Python, cuda, cmake, verdi, vim, git, JIRA, 英文工作环境.

工作经历/6 年

国内头部 GPU 厂商/3 年 | *软件工程师/team leader/算子加速库* 2021.1 至今

- 基于 GPGPU 平台从零开发 AI/HPC 领域加速库（DNN, BLAS, RAND, FFT）
- 带领 5 人团队

安防领域龙头企业/3 年半 | *软件工程师/AI 加速库/商业应用* 2017.7—2020.12

- **异构平台高性能卷积神经网络 CNN 库**开发，调研评估芯片性能；
- 端到端负责算法侧项目，深入梳理业务需求，构建高效的应用方案，加速智能算法落地；

项目经历

BLAS 加速库 |*0-1 开发维护* 2021.11-至今

- lead 项目，负责设计，构建，开发，测试，CICD，文档，管理；
- GEMM 优化，优化多种大模型场景，性能对标 A100。
- kernel selection 算法设计与实现；
- 实现了算子注册框架，日志系统，核心算子开发，

GEMM 极致优化实现 | *基于 zebu/verdi 极致优化* 2021.9—2021.10

- 手写汇编在 vcore 上实现极致性能 GEMM，通过 verdi 看波形，做到特定 shape 下利用率打满。
- 基于 tcore 实现模拟 fp32 GEMM 算法，也做到特定 shape 下利用率打满，性能超过 A100。

开发 **DNN 框架** | *基于 yaml 自定义算子表达* 2022.3—2023.6

- 参考 caffe 开发深度学习框架；
- 参考 onnx 设计基于 yaml 定义算子/图表达；
- 对标业界主流精度对比方案，自定义精度比对方案；
- 自动化部署精度验证，benchmark dashboard；