阮庭峰

phone：13979475804
mail：ruantingfeng22s@ict.ac.cn
github: github.com/tfruan2000

教育经历

中国科学院计算技术研究所 (研究生)	计算机技术	2022.8-至今
西安交通大学 (本科)	自动化	2018.8-2022.6

专业技能

- 熟练使用 CPP、Python，熟悉 PyTorch 编程
- 熟悉 MLIR 编程、了解 LLVM、了解 Triton 编程与调优
- 熟悉常见推理优化技术和 AI 编译器相关的“调度与自动调优”技术

项目经历

Triton kernel 性能优化 **2024.6-至今**

- 优化 FlagGems 中部分 reduce 类算子性能
  - 浅层优化：通过替换算子、合并 kernel、拆时间片循环等方式实现初步优化。
  - 深层优化：分析下降所得 IR，使用 perf 工具，对照算子库实现等方式，优化 kernel 的下降行为。
- 优化后，多数 reduce 类算子（all、sum、max 等）达到算子库性能的 1.3-1.6 倍。此外，优化前后 vector_norm、var_mean 算子有 6-8 倍 speedup，layernorm backward 提升 10 倍。

基于 MLIR 的 AI 编译器开发 **2023.8-至今**

- 性能优化
  - tensor 层面：分别为 linalg 和 triton 实现 ext-canonicalize，包含诸如 fold unit dim、fold dense mask 等 fold pattern。实现 linalg decompose、layout transpose 等 pass 以优化 IR 下降。
  - memref 层面：实现 continuity-enhancement 以避免不连续 memref 生成低效指令。实现跨 block 的 RE 以支持复杂控制流下的冗余 copy 删除。实现 barrier 合并。
- 语义合法性：实现 reposition pass 用于调整特定 op 的位置，使用 use-def 分析和 SliceAnalysis 依次完成 wrap 和 clone。实现部分 op 的 legalize pattern。
- 支持新 features：实现 output alias，支持非 dense value constant 下降，支持 i1 类型的下降。
- 完备性测试：修复多个算子下降流程，设计 stablehlo、linalg 算子测试方案并补充测试。

ICT-TX 跨平台编译联合项目 **2023.9-2024.3**

- 完善基于 MLIR 的编译器对 DCNMix 网络的部分算子以及特性支持。优化 DCNMix 网络推理时间。
  - 优化 slice、reduce 等算子下降行为，将其转换为对底层指令更友好的 IR，例如将 reduce 转为性能更好的 pooling 计算。
  - 通过 perf 分析单算子的运行时间，为 conv-like、pooling、element-wise 等算子实现自适应 tile pattern，自适应获得不同 shape 下的最优 tile 策略。
- 10 万数据量下，DCNMix 网络推理时间缩短到最初的 1/30。累计完成该项目 10 次交付。

其他

- 参与实验室项目：RISCV Intrinsic 扩展、基于 IREE 构建 IPRC-Scheduler
- 参与编写《智能计算系统》“编程框架机理”一章
- 多次获院三好学生，拥有赛艇国家二级运动员证书