第七届“复微杯”全国大学生电子设计大赛

基于 ICRAFT 编译器 AI 算法部署

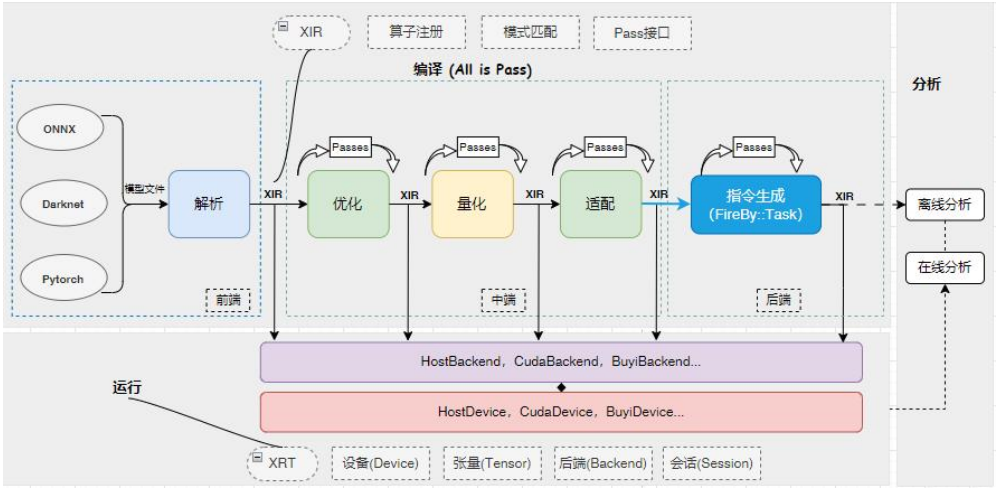
软件赛道赛题一

1 赛题简介

我司 AI 推理加速平台包括 FPAI 芯片和配套的 Icraft ai 编译器&工具包 2 部分。

其中 FPAI 芯片是一款异构芯片，包含了 cpu（ps），npu（icore），FPGA（pl），gpu，vpu 等硬件模块与资源。可灵活的对 AI 算法进行推理加速。目前已有芯片产品为中/小算力，主要面向边缘端场景的应用。

Icraft 是复微智能计算平台基于自主研发的 FPAI 芯片的软件工具套件，包括编译器、运行时等部署工具链。



如上图所示，Icraft 按功能分为多个组件，各组件采用统一的自定义的中间层数据结构（Icraft XIR），组件与组件之间的通信通过统一格式的 Json 文件和 raw 文件，Json 文件描述了网络结构，以及与 raw 文件数据的对应关系；raw 文件描述了网络参数、指令序列等。组件化的方式，便于我们快速开发替换新的组件功能。

其中，编译部分的功能组件为，解析器、优化器、量化器、适配器和指令生成。运行时（Icraft XRT）提供丰富的 API，支持 C++和 Python 接口，可支持用户将编译后的网络在指定的后端(计算资源)上运行。

本赛题要求大家使用 ICRAFT 部署工具在我司 FPAI 芯片进行特定领域的模型或 AI 解决方案进行部署与优化，根据对产品的使用深度与最终部署后的效果进行排名。

2 赛题要求

2.1 方向规定

可以部署某个模型，或者解决某个实际问题的整体 ai 方案

2.1.1 参考方向：

- 长时/遮挡目标跟踪
- 轻量化图像增强（去模糊，纹理增强，超分等）
- 多传感端（融合）检测/跟踪
- 综合视觉模型：检测/分割/跟踪一切等、多种视觉任务
- 多模态：单纯语音/图像转文本、多模态头
- llm 大模型
- 自动驾驶相关算法
- 强化学习&决策算法
- 通讯类算法

2.1.2 其他说明

1） 如果擅长的领域不在上述方向中，也可自行选择，但若选择的模型没有我们模型库中同类模型效果好，会影响评分。已有模型 list：https://gitee.com/mxh-spiger/modelzoo/blob/modelzoo_v3.7.1/index.md

2） 由于当前比赛决赛阶段基于我司边缘端 AI 芯片, 部分对运行效率有要求的领域的模型或整个方案，耗时是重点考察项，因此部署一些实用的轻量化算法会提高评分。

2.2 部署环节与考察点

1~4 步为初赛阶段；5~6 步为决赛阶段；

1. 算法/方案选择(根据选择部署的模型或方案是否实用、效果如何、有自己的创新点评分)

2. 模型导出（导出**不包含前后处理**的静态图模型，支持 torchscript, onnx 2 种格式），并验证导出模型与原模型一致性

3. 使用 icraft 将模型转化成 imodel（过程中可以先用 icraft 提供的 python api 进行编译环节仿真验证。最终模型转换成功，且有量化调优工作加分）

4. 使用 c++实现算法的前后处理或某些中间环节，并与源工程精度对齐；（若第三步编译遇到问题，此环节可用 libtorch&onnxruntime 等

c++推理库进行模型部分推理，实现精度对齐，并对前后处理有代码性能优化；有设计 fpga 模块进行部署环节加速者加分）

5. 上板测试模型 npu 耗时（根据模型耗时与部署后精度综合打分）

6. 上板性能优化（决赛；根据性能优化方案设计与效果打分）

2.3 重要赛制说明：

本赛题依赖 ICRAFT 编译器以及硬件板卡，边缘端 AI 加速芯片需要在灵活性、性能、功耗之间做平衡，因此对算子的支持并未完全覆盖 AI 框架的算子库。部署过程中可能存在算子不支持或其他问题导致无法部署下去而更换模型或方案等情况。本赛题额外设置了贡献度积分榜，将过程中的产出量化成分数，最终根据累积得分排名给予奖励。

2.4 部署文件提交必须遵循相关规范

参考已有模型库形式进行整理，具体细节参赛后培训

3 评分标准

3.1 初赛（积分制）

1） 初赛第一个月，不会获得编译器算子扩充需求或 debug 的支持，可以评估多套算法方案并列表汇报每个方案的情况与遇到的问题。1 个月后进行初审，对通过初审的方案进行支持。

2） 初赛阶段的多套算法部署允许流程走不完，根据每项完成数量与质量累积分数。最终根据分数榜排名给予优秀奖。

3) 有一套以上算法部署完成第 4 步即通过编译且仿真结果正确的队伍有进入决赛的**评审资格。**通过的队伍可选择一套最优的方案进行上板测试与优化。

评分细则：

序号	内容	标准	分值
1	算法方案选择/设计与部署评估	1. 根据模型或方案本身的性能与实用性、是否有原创设计或创新情况、硬件友好性评分； 2. 根据模型或方案的硬件资源分配划分（npu，cpu，fpga）的合理性与整体环节耗时与优化预估情况评分；	0~10
2	算法微调&模型导出	分离前后处理或方案中不利于在 npu 上计算的环节后导出模型，并验证一致性；根据导出难度、导出的静态模型性能一致性评分。	0~20
3	前后处理或中间环节 c++工程迁移	将 python 版的前后处理或中间需要 cpu 处理的环节剥离出来，并用 c++实现。根据移	0~20

		植功能难易程度以及 c++代码性能优化程度评分。	
4	ICRAFT 编译通过，精度调优	通过模型编译。并使用 ICRAFT pyrt① or crt② api 实现仿真精度测试工程；根据仿真精度情况评分	0~20
5	对 ICRAFT 提出有效建议或 bug	根据具体情况每条得分	2

注：①pyrt：icraft 提供的对转化模型进行仿真或上板推理的 python runtime api

②crt：icraft 提供的对转化模型进行仿真或上板推理的 c++ runtime api

3.2 决赛

决赛会结合初赛该方案的

序号	内容	标准	分值
1	上板测试模型推理耗时, 进行基本耗时调优；平衡耗时与混合精度策略；	·完成上板耗时测试获得基础分 ·使用 icraft 提供的内存复用，网络衔接等运行时优化功能进行基础优化工作加分 ·有平衡耗时与混合精度策略相关工作加分	0~50
2	完成完整的算法流程移植, 并对各环节进行耗时分析与效率优化, 或开发 fpga 模块进行优化。	根据方案设计、工作量与效果打分	0~50

4 赛事安排

4.1 参赛队伍要求

每支参赛队伍控制在 5 人以内。

4.2 赛事流程

阶段	时间节点	审核内容
初赛	约 1 个月，初审	部署方案初审，包括：方案规划，模型部署评估，部署遇到问题汇总。确定哪些遇到编译问题的方案可以得到编译器扩充或 debug 支持。
	约 2 个月，二审	各队伍提交工程与报告，结算初赛积分榜，角逐进入决赛的名额。有 1 个以上方案通过 icraft 编译且结果正确才有资格进入决赛。
决赛	约 1 个月，终审	使用编译过的模型进行上板部署、耗时测试、整体优化。
答辩		决赛答辩，确定名次与奖项

注：具体安排以组委会通知为主。

更多实时本赛题赛事信息请关注：https://gitee.com/mxh-spiger/fwb