FPGA 编译项目报告

——技术报告

高琦，苏灿，黎睿正

1. 项目背景

目前大模型部署存在的难点

- 工具结合度低

  现有的量化、打分和部署等工具模块间甚至模块内操作逻辑不一致，许多操作之间仍需手动调整参数。

- 新模型适配成本高

  现有的工作流需要根据所使用的模型手动调整，新模型需要逐一环节调试，无法开箱即用。

- 操作学习成本高

  现有的工具多为命令行工具，缺少直观的交互逻辑和图形化界面，用户操作上手难度较大。

2. 项目内容

2.1 技术特性

- 量化：支持 GPTQ 指定 bits、group_size、desc_act
- 打分
  - 支持 lm-evaluation-harness 测试项

    arc_easy、arc_challenge、gsm8k_cot、gsm8k_platinum_cot、hellaswag、mmlu、gpqa、boolq、openbookqa
  - 支持 EvalPlus 测试项

    humaneval、mbpp
- GPU 部署：支持 vLLM 指定上下文长度、显存占用限制、服务端口、API 密钥
- 权重处理：支持 Compiler-VCU128 指定 bits、group_size、desc_act

- FPGA 服务：支持 Fast API 指定上下文长度、生成温度、服务端口、API 密钥
- GPU、Port 调度器：支持 GPU 设备图分类、调整调度显存占用量、设备锁机制
- WebUI：支持指定服务地址、用户管理

2.2 WebUI

- 用户管理：支持自定义登录用户
- 打分：支持同时开展多框架多测试项测试和原模型、量化模型对比测试
- GPU 部署：支持同时部署原模型、量化模型以便对比
- 客制化：支持定制页面图标、企业名称、语言

2.3 FPGA OpenAI-Style API

3. 开发计划表

- 未来会继续向现有的工具链中添加新的功能
- 除了在各个环节加入新的功能外，还将额外开发 Cli 命令行工具以便在服务器终端操作调用
- 苏灿同学会同时参与量化部分的 llmc 开发

子项目	开发者	已完成内容	6月29日	未来开发
量化	高琦	GPTQ	AWQ	VIT量化工具适配
打分		Lm-Evaluation-Harness EvalPlus	OpenCompass	llmc
GPU部署		vLLM	暂无	
权重处理	苏灿	Compiler-VCU128	多模态模型权重处理和上板	
API服务		Fast API	暂无	
调度器	高琦	GPU, Port	已完成	
WebUI	黎睿正	基础功能	适配后端GPU部署测试功能	完成WebUI开发内容
Cli	高琦	暂不开发	适配完毕已有功能	

图 1 开发计划表