Transformer 算子开发

目录

1. 合作需求/任务

1.1. 研究背景

Transformer 由谷歌于 2017 年提出，整个网络结构完全由 attention 机制组成。由于其出色的性能以及对下游任务的友好性，或者说对下游任务仅仅微调即可取得不错的效果，得到了广泛的应用。当前主流的 NLP 模型，例如 Bert、GPT 系列等，都基于 Transformer 衍生而来。昇腾（Ascend）是华为推出的一系列 AI 加速器产品，旨在为人工智能应用提供高性能、低功耗、高效能的计算能力，支持 PyTorch、MindSpore 等多种 AI 框架。基于昇腾毕昇计算平台开发和优化上述算子，将能对 Transformer 网络带来巨大加速。

1.2. 研究目标

通过分析 Transformer 网络在英伟达 cuda 及 AMD ROCm 等计算平台的 profiling 数据，得出了 Transformer 网络的关键算子，包括矩阵乘(Matmul)、高斯误差线性单元（Gelu）、自适应动量优化器(Adam)、归一化指数函数(Softmax)，以及训练场景中用到的反向算子，如高斯误差线性单元反向(GeluGrad)、偏置加反向(BiasAddGrad)、批归一化反向(BatchNormGrad)等。其中矩阵乘算子华为 CANN 已提供可用的二进制实现。另外一些需要用毕昇 C++重写的算子例举如下：

算子类型	算子	优先级（数值约小优先级约高）
激活函数	Gelu	0
	GeluGrad	0
	Softmax	0
elementwise	BiasAddGrad	1
归一化	BatchNormGrad	1
优化器	Adam	2

我们希望通过本众智项目，实现 Gelu、GeluGrad、Softmax、BiasAddGrad、BatchNormGrad 和 Adam 一共 6 个算子的实现，并对接框架实现 Transformer 网络的端到端验证。在算子开发过程中，需要面向昇腾体系结构的，为这 6 个算子做有针对性的优化，以达到加速的目的。性能提升方面，从两个维度来衡量。功能正确，性能按两个维度评价。首先，相对于传统 CPU（鲲鹏 920）上实现的算子，性能提升 10+倍。第二方面，相对于在 CANN 中发布的基于 TBE 的实现，我们要求性能不低于对应的 CANN 中二进制的实现，挑战目标平均性能提升 5%。

1.3. 项目交付方式

兼容性需求

硬件：在鲲鹏+昇腾的多算力硬件架构下均完成；

软件：openEuler 22.03 lts， 毕昇融合编译器；使用毕昇 C++语言开发；、

其他要求

- 输出代码采用木兰 V2 协议。
- 使用的第三方软件清单列表，引入和刷新第三方软件要及时知会华为方并经过华为方同意。
- 开源项目，非必要不推荐使用第三方非开源项目依赖。
- 项目结束后，开发者需要继续 6 个月的 bug 维护期，维护期间对 bug 应做到 2 天内响应给出方案，1 周内解决。
- 故意放置恶意、安全漏洞代码的，将保留追究责任的一切权利。
- 源代码必须经过业界主流静态扫描工具扫描，并清零。
- 乙方完成项目以后，需要将代码提交到甲方指定仓库。

文档交付

根据华为提交的文档模板输出：

序号	交付
1	特性设计文档（根据附件文档模板编写）
2	特性测试方案
3	特性测试报告（包含功能+性能）

2. 项目验收方法

2.1 验收环境

硬件环境：

硬件	Kunpeng 920+Ascend 910
网络	IB/RoCE 100GE
存储	OceanStor 100D

软件环境：

OS	openEuler	20.03-SP3
编译器	毕昇编译器	支持异构计算平台编译
MPI	HMPI	1.1.1

2.2 一致性

3 次运行软件，每次得到构建、性能测试结果需保持不变

2.3 验收标准

文档规范性

文档所需的各个章节完备，语言简练准确。表述错误、遗漏低于 0 处。

功能性

*使用上述软硬件环境，不额外安装任何软件、rpm 包，即可成功编译得到所需的二进制文件。

性能和准确性对比

为了更强力的去支撑各种各样的应用，在算子的构建和优化中我们需要达到更高的性能，让船脸识别系统在华为 AI 芯片毕昇计算平台适配使用。Gelu、GeluGrad、Softmax、BiasAddGrad、BatchNormGrad 和 Adam 算子功能正确，精度达到 e-06 以上，性能优于现有对应算子的性能 5%，对比 X86(intel 6354(3.0GHz))+ AMD(MI210 GPU)同类算子性能达到 80%以上，整体船脸识别模块使用新算子后性能提升 10%以上。

	场景一	场景二
硬件	Kunpeng 920+Ascend 910	Intel i7+ AMD
OS	OpenEuler 20.03.sp3	CentOS 7.4
编译器	毕昇编译器（支持异构计算平台编译）	-

3. 交付计划

乙方应在中国境内（“工作地点”），按照下表的各阶段开展协议工作。各阶段工作的详细计划、交付件及验收标准如下所示：

注：T 为合同签署生效日

阶段	开始时间	任务描述	交付件	验收标准
开工会	项目开始	对齐项目计划与初始方案	1. 开工会纪要 2. 项目实施计划书	包括项目需求分析，初始方案，开发团队
1 阶段	6 个月	功能实现，性能达标	采用毕昇 C++异构编程模型改写网络，实现昇腾对关键计算部分进行加速	功能和性能达到验收指标

4. 项目监控和管理

本项目采用如下项目管理机制：

1) 项目月报。
2) 不定期面谈项目进展。
3) 交付后进行设计方案串讲。
4) 根据实际需要安排的其它交流，如电话、邮件、电话会议等。

5. 附录（供参考）

1) 考虑到应用开发的过程中，存在系统库的依赖情况。针对改造的 6 个函数，成功率越高验收评分越高；

2） SOW 中提供的软件包列表为参考列表，在与业务放沟通达成一致意见后，可以进行替换

3） 此附件是 NRE 合作协议的补充，与 NRE 协议具有同等法律效力，未约定事项按照 NRE 协议约定执行。因履行本协议甲乙双方产生纠纷的，应当另行协商解决。

(以下无正文)