AML 作业一：GLM4-9B 的基础代码复现

2024 年 10 月 15 日

1 作业背景

在课程中，我们已经介绍了机器学习和深度学习的基础知识，以及语言模型 (LLM, Large Language Models) 的基本概念。尽管业界已有各种框架对相关算法进行了封装，但了解底层算法的实现对加强我们对模型的全面理解至关重要。通过手动实现这些底层代码，我们将更好地掌握 LLM 的工作原理，提高编程和调试技巧。

2 作业目标

1. 根据课上所讲内容以及相关的阅读材料了解当前大语言模型的基础结构，包括但不仅限于 Multi-Group Attention、RMSNorm、SwiGLU、RoPE 等。
2. 基于给定的代码模板进行补全，复现 GLM4 的基础代码，支持推理用的 KV-Cache 等参数。Attention、RMSNorm 等主要的模块需要手动实现，其余如 Softmax，Linear 等基础模块可以直接调用 torch 自带的函数。
3. 加载 GLM4-9B 官方 hugging face 权重代码，并通过改名的方式将其权重转换到自己的模型中。保证模型在使用 hugging face 的 generation 接口后可以进行正常对话。（**提示：作业用国内的 hugging face镜像**）
4. 尝试将 attention 部分改为 flash attention，并测试模型在不同长度下的空间、时间加速比。

3 提交材料

1. 一份报告，解释你的调研过程以及代码实现方式，加速比测试的结果，以及遇到的任何挑战或困难。
2. 一份补全后可以成功运行的 run_glm4.py 代码，在报告中需要附上成功运行后对话的截图。

4 评分标准

1. 基础代码实现（40 分）：如果实现的有问题，根据实现内容的多少可能会给一部分分。
2. 成功进行 checkpoint 转换并实现对话（20 分）

3. 实现 flash attention（20 分）: flash attention 的代码需要提交一份, 即提交一个正常实现 attention 的代码以及基于 flash attention 的代码。

4. 报告（20 分）

请将你的所有作业内容打包成一个 zip 压缩文件并提交。

5 参考文献

- Attention Is All You Need
- GPT2 blog
- GQA: Training Generalized Multi-Query Transformer Models from Multi-Head Checkpoints
- Root Mean Square Layer Normalization
- GLU Variants Improve Transformer
- hugging face glm4-9b
- ChatGLM: A Family of Large Language Models from GLM-130B to GLM-4 All Tools
- pip 环境参考 glm4 github