learning-lm-rs 报告

仓库地址：[learning-lm-rs GitHub]

一、项目概述

本项目在基础语言模型功能之上实现了两大核心扩展：

1. **混合精度推理系统** - 支持 FP32/FP16/BF16 计算模式
2. **网络服务化接口** - 提供网络 API 服务

开发环境

- **操作系统**：WSL2 (Ubuntu 22.04 LTS)
- **技术栈**：Rust 1.83，Actix-Web 4.0

二、核心功能实现

1. 混合精度推理

技术方案：

- 采用泛型编程范式抽象计算逻辑
- 集成 `half` 库实现 FP16/BF16 数据类型支持

2. 网络服务 API

技术方案：

- 基于 Actix-Web 框架构建异步 REST API
- 使用 session 技术识别用户
- 在 APP 层面为不同用户缓存 KVCache

3. 多线程推理

技术方案：

- **MLP 层**：从 `intermediate_size` 维度切分，分给不同线程处理
- **Attention 层**：从 `head` 层切分，分给不同线程处理

三、功能演示

1. 长文本生成能力

```
Welcome to Llama Chatbot!
Please select a mode:
1. Chat mode
2. Story mode
3. API service
4. Exit
2

Once upon a time, a little girl named Lily lived in a small house with her mom. She loved to play with the other toys with her toy
 car, making noises and pretended it would be like it.
One day, Lily found a pretty box in her room with long things. She wanted to show her mom and dad and dad the special pieces of fo
od. Lily felt happy that she could make the best car clean.
At the park, Lily met a little boy named Tim. She asked, "Can I play with your toy car?" Her mom said, "Yes, but we need to help m
e make them look happy and poking." Lily smiled and said, "Thank you, Tim!" They played together all day long, and Lucy was so hap
py that she could do it by it.
The next day, Lily went back inside and saw the room with the big, beautiful vase with its balloon. Now, Lily was clean and had so
 much fun! She put the bitter car back to Tim. Tim felt proud that she had found herb and went home to her mom. They thanked her m
om for her new mom and promised to be careful with her new friend.<|end_story|>
```

2. 对话交互系统

```
Welcome to Llama Chatbot!
Please select a mode:
1. Chat mode
2. Story mode
3. API service
4. Exit
1

You:
Hi!
Assistant:
 I' m glad to hear that you' re enjoying this amazing and exciting adventure in the world of the arts. Have you tried any other ar
t or art before?
You:
```

3. API 服务接口

```
root@DESKTOP-KUU63C8:~/Simon/InfiniTensorTMP/learning-lm-rs# curl http://localhost:8080/story
Once upon a time, a woman dog named Spot lived in a nest with many church of trees and a nest. It was made of bees in its nest and
 lovely. One day, Spot was playing in the nest. They saw a big red box near a tree. The nest was in the tree. Spot wanted to find
something special. So, he had an idea.
Spot found a pile of circle! Then he found a big red nest. He opened his eyes and found his nest. The nestle said, "Wow, where it
is! Let's use it." They both laughed and played with the nest until it was time to go home.
After they reached the nest, the nest began to blow! It was a magic nest in it. The nest had the nest belonged back to its nest. M
ax was happy too! He could use the nestmir the nest too. Spot was excited to see all the nest too!
The nest was now a little thin nut! Now, the nest was not a meat! It was a very fun game. The nest was not for them! root@DESKTOP-
KUU63C8:~/Simon/InfiniTensorTMP/learning-lm-rs# curl http://localhost:8080/chat/hi
 I am unable to write any specific information. However, I can provide some general information regarding the specific techniques
and techniques employed by a professor or professor or instructor to support your claims.root@DESKTOP-KUU63C8:~/Simon/InfiniTensor
TMP/learning-lm-rs#
```

4. 多线程推理

将 `top_k` 设置为 1 后，使用 Chat 模型，在用户输入为 `hi` 的条件下，进行单线程与多线程推理对比：

- **单线程推理**：231 秒
- **多线程推理**：167 秒
- **提速效果**：提升 37%

```
Welcome to Llama Chatbot!
Please select a mode:
1. Chat mode
2. Story mode
3. API service
4. Exit
1

You:
hi
Assistant:
 I am not capable of creating visual content. However, I can provide you with some general gu
idelines for creating visual content for your website. Here are some guidelines:

1. Use a clear and conc ise style that is easy to read and understand. Use a clear and conc i
se style that is easy to read and understand.

2. Use a clear and conc ise style that is easy to read and understand. Use a clear and conc i
se style that is easy to read and understand.

3. Use a clear and conc ise style that is easy to read and understand. Use a clear and conc i
se style that is easy to read and understand.

4. Use a clear and conc ise style that is easy to read and understand. Use a clear and conc i
se style that is easy to read and understand.

5. Use a clear and conc ise style that is easy to read and understand. Use a clear and conc i
se style that is easy to read and understand.

6. Use a clear and conc ise style that is easy to read and understand. Use a clear and conc i
se style that is easy to read and understand.

7. Use a clear and conc ise style that is easy to read and understand. UseTime taken: 231.931
855482s
```
单线程下需要231秒

```
Welcome to Llama Chatbot!
Please select a mode:
1. Chat mode
2. Story mode
3. API service
4. Exit
1

You:
hi
Assistant:
 I am not capable of creating visual content. However, I can provide you with some general gu
idelines for creating visual content for your website. Here are some guidelines:

1. Use a clear and conc ise style that is easy to read and understand. Use a clear and conc i
se style that is easy to read and understand.

2. Use a clear and conc ise style that is easy to read and understand. Use a clear and conc i
se style that is easy to read and understand.

3. Use a clear and conc ise style that is easy to read and understand. Use a clear and conc i
se style that is easy to read and understand.

4. Use a clear and conc ise style that is easy to read and understand. Use a clear and conc i
se style that is easy to read and understand.

5. Use a clear and conc ise style that is easy to read and understand. Use a clear and conc i
se style that is easy to read and understand.

6. Use a clear and conc ise style that is easy to read and understand. Use a clear and conc i
se style that is easy to read and understand.

7. Use a clear and conc ise style that is easy to read and understand. UseTime taken: 167.953
068622s
```
启用多线程(4线程)后时间缩短为167s,提速37%

四、待提高

1. **混合精度** 不支持 TF32。
2. **Chat 模型的网络服务 API** 等待时间过长。
3. 功能可以进一步扩展。