PR #24760 完整报告

sgl-project/sglang

[Bug Fix] Fix broken sgemm_lora_a_graph_fwd due to invalid torch.mm() call

合并时间：2026-05-13 03:29

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/24760

执行摘要

- 一句话：修复 sgemm_lora_a_graph_fwd 中 torch.mm() 多余参数导致的 TypeError
- 推荐动作：PR 虽小，但修复了一个明确的 bug，值得快速合入。对于深入学习 LoRA 或图模式执行的开发者，可以查看该函数理解 torch.mm 的正确用法。

功能与动机

在 python/sglang/srt/lora/torch_ops/graph_lora_ops.py 的 sgemm_lora_a_graph_fwd 函数中，torch.mm(x_seq, w_seq.t(), 0) 传入了第三个参数 0，但 torch.mm 不支持这种签名（期望 (Tensor, Tensor) 或 (Tensor, Tensor, dtype)），导致运行时 TypeError，使得对应测试无法通过。

实现拆解

1. 定位问题：发现 sgemm_lora_a_graph_fwd 函数中调用 torch.mm 时多传了一个整数 0。
2. 修改代码：在 python/sglang/srt/lora/torch_ops/graph_lora_ops.py 中，将 torch.mm(x_seq, w_seq.t(), 0) 改为 torch.mm(x_seq, w_seq.t())，删除多余的第三个参数。
3. 验证：运行 test/manual/lora/test_lora_ops.py 中的 test_sgemm_lora_a_graph_fwd 测试，确认通过；同时全部 9 个 sgemm_lora 相关测试均通过，无回归。

关键文件：

- python/sglang/srt/lora/torch_ops/graph_lora_ops.py（模块 LoRA；类别 source；类型 core-logic；符号 sgemm_lora_a_graph_fwd）：唯一变更文件，修复了 sgemm_lora_a_graph_fwd 函数中的 torch.mm 参数错误。

关键符号：sgemm_lora_a_graph_fwd

关键源码片段

python/sglang/srt/lora/torch_ops/graph_lora_ops.py

唯一变更文件，修复了 sgemm_lora_a_graph_fwd 函数中的 torch.mm 参数错误。

```
# python/sglang/srt/lora/torch_ops/graph_lora_ops.py
# 函数 sgemm_lora_a_graph_fwd 的一部分，展示了修复前后的关键行

# @brief 执行 LoRA A 的图模式前向传播
```

```
# 错误修复前：torch.mm(x_seq, w_seq.t(), 0) # 第三个参数 0 不合法
# 修复后：
output.add_(scaling_tensor[lora_idx] * torch.mm(x_seq, w_seq.t())) # 正确调用

# 上下文说明：
# x_seq = torch.where(batch_token_mask, inputs, 0)
# w_seq = weights[lora_idx]
# output 为累加结果
```

评论区精华

审查者 ping1jing2 在多次 CI 失败后，检查了所有排队和失败的 CI 并确认失败与本次更改无关，最终合并 PR。作者 flutist 也指出 CI 失败是预先存在的 flaky 测试。

- CI 失败与更改无关的确认 (other)：确认 CI 失败为与 PR 无关的 flaky 测试，正常合并。

风险与影响

- 风险：该 PR 仅修改一行代码，删除一个多余参数，逻辑清晰，风险极低。回归测试覆盖全部 9 个 sgemm_lora 测试用例，未发现新问题。但若其他调用路径也隐含类似错误，可能未被当前测试覆盖。
- 影响：直接影响 LoRA 模块的图模式前向传播（sgemm_lora_a_graph_fwd），修复了此前运行时的 TypeError，使相关功能恢复正常。对非 LoRA 场景无影响。
- 风险标记：单行修复，已有测试覆盖

关联脉络

- 暂无明显关联 PR