PR #20564 完整报告

sgl-project/sglang

fix: torch-native LoRA for multi-adapter case

合并时间：2026-03-27 05:34

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/20564

执行摘要

- 一句话：修复 torch-native LoRA 后端在批处理中多适配器请求时的张量大小匹配错误。
- 推荐动作：此 PR 值得快速浏览，特别是对于处理类似去重逻辑的开发者。关注 prepare_lora_batch 中变量一致性的修复模式，以及如何通过测试更新确保覆盖边缘案例。

功能与动机

根据 PR body，动机是修复 RuntimeError：'The size of tensor a (3) must match the size of tensor b (2) at non-singleton dimension 0'，该错误发生在批处理中部分连续请求共享相同适配器时，由 torch.unique_consecutive 去重后大小不匹配引起。

实现拆解

修改了两个文件：1) torch_backend.py 中的 prepare_lora_batch 函数：将 batch_info.num_segments 从 forward_batch.batch_size 改为 num_segments（基于唯一段数）；在复制 weight_indices 时，将索引范围从 [:bs] 改为 [:num_segments]。2) test_torch_backend.py：更新测试用例，使用 weight_indices=[0,0,1] 和 batch_size=3 来模拟多适配器段合并，确保修复路径被覆盖。

关键文件：

- python/sglang/srt/lora/backend/torch_backend.py（模块 lora/backend）：修复了 prepare_lora_batch 函数中的核心 bug，确保 num_segments 和 weight_indices 复制使用正确的唯一段数而非批大小。
- test/manual/lora/test_torch_backend.py（模块 test/lora）：更新单元测试以覆盖多适配器段合并场景，使用 weight_indices=[0,0,1] 和 batch_size=3 验证修复正确性。

关键符号：prepare_lora_batch

评论区精华

Review 过程中没有实质性讨论，只有 approvals。Issue 评论中 claude-pr-review-bot 指出：'Risk Level: Low. Summary: Correct and minimal fix for a real bug...'，确认修复正确且风险低。

- 修复正确性验证 (correctness): 修复被接受并合并。

风险与影响

- 风险：风险低，因为修复针对特定 bug，且测试更新以覆盖场景。潜在风险是如果其他代码假设 num_segments 等于 batch_size，但在此上下文中，num_segments 是基于唯一段数计算的，不会影响其他部分。修改在核心 LoRA 路径上，但范围小，回归风险可控。
- 影响：修复了 torch-native LoRA 后端在多适配器场景下的运行时崩溃，提升了系统稳定性和可靠性。影响范围仅限于使用此后端的批处理请求，对用户透明，但避免了服务中断。
- 风险标记：核心路径变更，测试覆盖增强

关联脉络

- PR #20562 Use torch.addmm instead of separate mm and add_ calls for LoRA torch.native: 同为 torch-native LoRA 后端优化，修改了相同文件 python/sglang/srt/lora/backend/torch_backend.py，属于同一模块的近期改进。