PR #36976 完整报告

vllm-project/vllm

[Bugfix][LoRA] Fix Qwen35 LoRA

合并时间：2026-03-20 11:09

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/36976

执行摘要

- 一句话：修复 Qwen3.5 模型的 LoRA 适配器支持，解决 IndexError 问题。
- 推荐动作：该 PR 值得精读，特别是模型层 LoRA 兼容性设计决策，如分离 in_proj_qkvz 层以处理 GDN 结构，这为其他支持 LoRA 的模型提供参考。同时关注 gemini-code-assist[bot] 提出的 packed_modules_mapping 修复是否已正确实施。

功能与动机

根据 review 评论，musab-mk 确认此 PR 修复了 LoRA 适配器在 Qwen/Qwen3.5-397B-A17B-FP8 模型上导致的 IndexError，标题和标签也表明这是一个针对 Qwen3.5 的 bug 修复。

实现拆解

实现主要包括三个部分：1) 在 qwen3_5.py 中修改 Qwen3_5GatedDeltaNet 类，当 LoRA 启用时（vllm_config.lora_config 非空）使用 MergedColumnParallelLinear 和 ColumnParallelLinear 分离 in_proj_qkv 和 in_proj_z，并在 forward 方法中添加条件分支处理；2) 在 qwen3_next.py 中调整 Qwen3NextGatedDeltaNet 的 __init__方法，新增 create_in_proj_qkvz 参数以支持子类自定义；3) 添加测试文件 test_qwen35_densemoel_lora.py 并更新 CI 配置 lora.yaml 和 conftest.py 以集成新测试。

关键文件：

- vllm/model_executor/models/qwen3_5.py（模块 model_executor/models）：核心模型修改，实现 LoRA 兼容的 in_proj 分离，处理 Gated Delta Network 结构
- vllm/model_executor/models/qwen3_next.py（模块 model_executor/models）：基础类修改，支持 create_in_proj_qkvz 参数，为 LoRA 路径提供灵活性
- tests/lora/test_qwen35_densemoel_lora.py（模块 tests/lora）：新增测试文件，验证 Qwen3.5 密集模型 LoRA 修复的正确性和 TP4 场景

关键符号：Qwen3_5GatedDeltaNet.init, Qwen3_5GatedDeltaNet.forward, Qwen3NextGatedDeltaNet.init

评论区精华

gemini-code-assist[bot] 指出 critical 问题：Qwen3_5ForConditionalGeneration 的 packed_modules_mapping 未正确定义，可能导致 LoRA 在其他模块（如 qkv_proj）上失效，建议从 Qwen3NextForCausalLM 继承映射。此问题在 PR 合并前应已解决，但材料未显示具体修复。musab-mk 测试确认修复有效，DarkLight1337 批准合并。

- packed_modules_mapping 初始化问题 (design): PR 合并前应已解决，但材料未显示具体修复代码
- 测试确认修复有效性 (correctness): 修复有效，PR 被批准合并

风险与影响

- 风险：风险包括：1) 核心模型文件 qwen3_5.py 变更可能引入回归，影响 Qwen3.5 的非 LoRA 路径性能或正确性；2) packed_modules_mapping 问题若未完全解决，LoRA 支持可能仍不完整，导致其他模块失效；3) 新增测试仅覆盖密集模型（Qwen3.5-4B），未测试更大规模或混合专家模型，可能存在覆盖不足。
- 影响：直接影响使用 Qwen3.5 模型和 LoRA 适配器的用户，修复了一个导致 CUDA 图捕获崩溃的 bug，提升模型可用性。对系统整体影响较小，仅限于 Qwen3.5 模型的 LoRA 功能。团队需关注类似模型 LoRA 支持的兼容性设计。
- 风险标记：核心模型变更，LoRA 兼容性风险，测试覆盖有限

关联脉络

- PR #37816 [CI/Build][LoRA] Update Qwen35 LoRA testing: 关联测试更新，可能是同一功能线的后续 PR，扩展测试覆盖
- PR #37810 [Bugfix] Store Qwen3Next A_log in fp32: 涉及 Qwen3Next 模型修复，相关模型结构调整
- PR #37338 [Perf] [Bugfix] Fix Triton autotuning in inference for Qwen3.5: 涉及 Qwen3.5 性能 bugfix，类似模型特定修复