

PR #47031 完整报告

vllm-project/vllm

[Bugfix] Fix GraniteMoeShared weight loading broken by #41184

合并时间: 2026-07-01 22:39

原文链接: <http://prhub.com.cn/vllm-project/vllm/pull/47031>

执行摘要

- 一句话: 修复 GraniteMoeShared 权重加载路径错误
- 推荐动作: 该 PR 值得合并, 修复精确且影响面小。建议开发者在未来类似重构中更关注子类覆盖的协议一致性, 并考虑增加回归测试。

功能与动机

修复 IBM GraniteMoeShared 模型在 FusedMoE 重构后权重加载失败的问题。用户运行 `ibm-research/moe-7b-1b-active-shared-experts` 时遇到 `KeyError: 'layers.0.block_sparse_moe.experts.routed_experts.routed_experts.w13_weight'`, 导致引擎初始化失败。

实现拆解

1. 定位根因: 在 `vllm/model_executor/models/granitemoeshared.py` 的 `load_weights()` 方法中, 三个 `n.replace()` 调用生成 `expert` 键时多插入了 `routed_experts` 段, 与父类期望的 `experts.{e}.w{n}.weight` 格式不符。
2. 修正替换目标: 将 `w1_name`、`w3_name` 和 `w2_name` 的替换字符串中的 `routed_experts` 删除, 使生成的键路径为 `experts.{e}.w{n}.weight`。
3. 验证修复: 通过加载 `ibm-research/moe-7b-1b-active-shared-experts` 模型并执行一次生成验证, 确认权重加载成功, 无 `KeyError`。

关键文件:

- `vllm/model_executor/models/granitemoeshared.py` (模块 模型加载; 类别 `source`; 类型 `data-contract`): 包含 `GraniteMoeModel.load_weights()` 方法, 是本次 Bug 的根因和修复所在。三处替换字符串中的 `routed_experts` 段被移除, 对齐父类期望的专家键格式。

关键符号: `GraniteMoeModel.load_weights`

关键源码片段

`vllm/model_executor/models/granitemoeshared.py`

包含 `GraniteMoeModel.load_weights()` 方法, 是本次 Bug 的根因和修复所在。三处替换字符串中的 `routed_experts` 段被移除, 对齐父类期望的专家键格式。

```
# vllm/model_executor/models/granitemoeshared.py
```

修复：移除 spurious `routed\_experts` 段，使键路径与父类契约一致

```
def load_weights(self, weights: Iterable[tuple[str, torch.Tensor]]) -> set[str]:
    new_weights = {}
    for n, p in weights:
        if n.endswith(".block_sparse_moe.input_linear.weight"):
            for e in range(p.size(0)):
                # 原: f".block_sparse_moe.experts.routed_experts.{e}.w1.weight"
                # 改后: f".block_sparse_moe.experts.{e}.w1.weight"
                w1_name = n.replace(
                    ".block_sparse_moe.input_linear.weight",
                    f".block_sparse_moe.experts.{e}.w1.weight",
                )
                w3_name = n.replace(
                    ".block_sparse_moe.input_linear.weight",
                    f".block_sparse_moe.experts.{e}.w3.weight",
                )
                w1_param, w3_param = p[e].chunk(2, dim=0)
                assert w1_name not in new_weights
                assert w3_name not in new_weights
                new_weights[w1_name] = w1_param
                new_weights[w3_name] = w3_param
            elif n.endswith(".block_sparse_moe.output_linear.weight"):
                for e in range(p.size(0)):
                    # 原: ...routed_experts.{e}.w2.weight
                    # 改后: ...experts.{e}.w2.weight
                    w2_name = n.replace(
                        ".block_sparse_moe.output_linear.weight",
                        f".block_sparse_moe.experts.{e}.w2.weight",
                    )
                    w2_param = p[e]
                    assert w2_name not in new_weights
                    new_weights[w2_name] = w2_param
            elif n.endswith(".block_sparse_moe.router.layer.weight"):
                gate_name = n.replace(
                    ".block_sparse_moe.router.layer.weight",
                    ".block_sparse_moe.gate.weight",
                )
                assert gate_name not in new_weights
                new_weights[gate_name] = p
            else:
                new_weights[n] = p
    # 委托给父类进行实际权重赋值
    return GraniteMoeModel._load_weights(self, new_weights.items())
```

评论区精华

review 过程简单，主要由 reviewer jikunshang 批准。无额外争议或讨论。

- 暂无高价值评论线程

风险与影响

- 风险：风险极低。改动仅修改三个字符串字面量，移除冗余的 `routed_experts` 段，逻辑完全符合父类契约。由于该修改仅影响权重加载路径，不会引入回归性功能问题。但缺少新增的单元测试用例覆盖该修复。
- 影响：影响范围限定为 `GraniteMoeShared` 模型系列（如 `ibm-research/moe-7b-1b-active-shared-experts`），修复了该模型无法加载权重的阻断性 Bug。对其他模型无影响。
- 风险标记：缺少测试覆盖

关联脉络

- PR #41184 `FusedMoE/MoERunner inversion refactor`: 本次 Bug 由该 PR 引入，父类 `_load_weights()` 期望的键格式发生变化，但 `GraniteMoeShared` 未同步更新。
- PR #45054 `Fix same class of issue in aria.py, qwen3_vl_moe.py, step3_text.py, step3p5.py`: 解决了同类型问题，但显式排除了 `granitemoeshared.py`，本 PR 补全该遗漏。