PR #42630 完整报告

vllm-project/vllm

gemma3 multi-gpu bug-fix

合并时间：2026-05-15 17:32

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/42630

执行摘要

- 一句话：修复 Gemma3 多 GPU 下形状不匹配 bug
- 推荐动作：值得合并，属于典型的一行修复 bug。可快速批准合并。建议未来在类似 RowParallelLinear 的使用处也显式指定 input_is_parallel 参数以增强可读性。

功能与动机

修复使用 --tensor-parallel-size 2 运行 Gemma-3n 模型时出现的 RuntimeError: mat1 and mat2 shapes cannot be multiplied (1024x2048 and 1024x2048) 错误。此问题在 unsloth/gemma-3n-E4B-it 和 google/gemma-3n-E4B-it 上均被复现。

实现拆解

1. 定位问题：在 vllm/model_executor/models/gemma3n_mm.py 的 Gemma3nMultimodalProjector.__init__ 中，embedding_projection 使用 RowParallelLinear 从 multimodal_hidden_size 投影到 text_hidden_size。当张量并行数 >1 时，RowParallelLinear 默认会将输入视为已沿并行维度分片（即 input_is_parallel=True），但该处输入是来自 VocabParallelEmbedding 的全宽张量，导致形状不匹配。
2. 添加参数：在构造函数调用 RowParallelLinear 时显式设置 input_is_parallel=False，告知该类需要对全宽输入进行内部散射，而非期望并行分片后的输入。
3. 变更范围：仅修改 gemma3n_mm.py 中的一行代码，无其他文件、测试或配置配套变更。

关键文件：

- vllm/model_executor/models/gemma3n_mm.py（模块 模型执行器；类别 source；类型 data-contract；符号 Gemma3nMultimodalProjector.init）：仅修改此文件，在 Gemma3nMultimodalProjector.__init__ 的 RowParallelLinear 调用中增加 input_is_parallel=False 参数，修复张量并行时形状不匹配的 bug。

关键符号：Gemma3nMultimodalProjector.init

关键源码片段

vllm/model_executor/models/gemma3n_mm.py

仅修改此文件，在 Gemma3nMultimodalProjector.__init__ 的 RowParallelLinear 调用中增加 input_is_parallel=False 参数，修复张量并行时形状不匹配的 bug。

```
# 修改位置：vllm/model_executor/models/gemma3n_mm.py:420-425
# 在 __init__ 中，embedding_projection 使用 RowParallelLinear 进行投影。
# RowParallelLinear 默认假设输入已沿并行维度分片 (input_is_parallel=True)，
# 但此处的 emb_norm 输出是全宽张量，因此需要显式设置 input_is_parallel=False
# 让 RowParallelLinear 在内部自行散射，避免形状不匹配。
self.embedding_projection = RowParallelLinear(
    self.multimodal_hidden_size,
    self.text_hidden_size,
    bias=False,
    input_is_parallel=False, # 修复：传入 False 确保正确分片
)
```

评论区精华

Review 中未出现争议讨论。tjtanaa 询问原版权重是否也会遇到此问题，作者确认 google/gemma-3n-E4B-it 同样存在，并已更新 PR 描述。tjtanaa 随后批准变更。

- 官方权重是否也存在相同问题 (question): 作者确认官方权重同样存在问题。

风险与影响

- 风险：风险极低。变更仅添加一个已明确定义的默认参数，且语义与已有逻辑一致。若 RowParallelLinear 默认行为在未来被修改（可能性小），此参数可确保兼容性。无性能或安全风险。
- 影响：影响范围明确：仅修复 Gemma-3n 模型在多 GPU（张量并行 >1）场景下的崩溃 bug。对单卡用户、其他模型、或非张量并行场景无影响。未引入新特性，也未影响 API 或配置接口。
- 风险标记：暂无

关联脉络

- 暂无明显关联 PR