PR #25407 完整报告

sgl-project/sglang

Fix Mistral Large 3 nightly test

合并时间：2026-05-16 08:28

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/25407

执行摘要

- 一句话：修复 Mistral Large 3 测试因 scale shape 不匹配失败
- 推荐动作：建议合并。该修复针对性强，改动极小且经过测试验证。可考虑在后续 PR 中增强切片安全性，例如对空 tensor 做 fallback 处理。

功能与动机

Mistral Large 3 的 nightly 测试因 fp4_quantize 的 scale 参数形状不匹配而失败。原代码对 layer.w13_input_scale_quant 不做切片，但 Mistral Large 3 的该 tensor 形状可能不是 [1]，导致 cute-dsl 后端报错。PR body 中贴出的测试通过日志表明修复后测试正常。

实现拆解

修改位于 compressed_tensors_w4a4_nvfp4_moe.py 的 apply_weights 方法中，仅一处变更：

1. 定位问题：在 apply_weights 函数中调用 fp4_quantize 时，传入 layer.w13_input_scale_quant 作为 scale 参数。该 tensor 在 Mistral Large 3 上形状为 [N, ...]（N 为专家数或分组数），而 cute-dsl 后端要求 scale 为 shape [1] 的标量。
2. 修改代码：将 layer.w13_input_scale_quant 改为 layer.w13_input_scale_quant[:1]，取第一个元素作为全局 scale，确保形状为 [1]。
3. 更新注释：将原注释 # Quantize input hidden states using fp4_quantize 改为 # global_scale must be shape [1] (strict in cute-dsl backend).，明确约束条件。
4. 回归影响：该修改仅影响 use_flashinfer_trtllm 为 True 的 NVFP4 MoE 量化路径，其他路径不受影响。

关键文件：

- python/sglang/srt/layers/quantization/compressed_tensors/schemes/compressed_tensors_w4a4_nvfp4_moe.py（模块 量化层；类别 source；类型 core-logic）：唯一修改的文件，核心变更：将 layer.w13_input_scale_quant 切片为 [:1] 以适应 cute-dsl 后端对 scale shape 为 [1] 的严格要求。

关键符号：未识别

关键源码片段

python/sglang/srt/layers/quantization/compressed_tensors/schemes/compressed_tensors_w4a4_nvfp4_moe.py

唯一修改的文件，核心变更：将 layer.w13_input_scale_quant 切片为 [:1] 以适应 cute-dsl 后端对 scale shape 为 [1] 的严格要求。

```
# 位于 python/sglang/srt/layers/quantization/compressed_tensors/schemes/compressed_tensors_w4a4_nvfp4_moe.py
# 方法：apply_weights，约第 314-321 行

# global_scale must be shape [1] (strict in cute-dsl backend).
hs_fp4_bytes, hs_sf_bytes = fp4_quantize(
    x,
    layer.w13_input_scale_quant[:1], # 取第一个元素确保 shape [1]
    self.group_size, # sf_vec_size
    False, # use_ue8m0
    False, # is_sf_swizzled_layout
)
```

评论区精华

Review 中 gemini-code-assist[bot] 提出现有切片方式在分布式中 rank 无本地 expert 时（如 TP degree > expert count）可能导致空 tensor (shape [0])，建议增加保护逻辑。作者 b8zhong 反问该场景是否可能存在，表明当前认为 Mistral Large 3 (128 experts) 不会触发。Fridge003 已 approve。

- 切片在无本地 expert 时的安全性 (correctness): 作者 b8zhong 反问该场景是否可能，未获得明确确认。当前 PR 作为热修复被 approve，但未解决边角案例。
- 根本修复位置讨论 (design): 该建议未被采纳到当前 PR，PR 作者选择在调用点做最小修复。

风险与影响

- 风险：修改风险较低，因为：
  1. 仅一行切片操作，非常直观。
  2. 覆盖了 Mistral Large 3 的测试场景（已通过）。
  3. 主要风险是极端配置下 TP degree 大于实际 expert 数时，[:1] 可能产生空 tensor。但此场景极罕见，且当前 PR 作为 hotfix 可接受。
  4. 未修改测试文件，回归风险需依赖已有 CI 覆盖。 - 影响：直接影响：修复 Mistral Large 3 在 FP4 量化路径下的运行问题。间接影响：所有使用 flashinfer_trtllm 后端的 NVFP4 MoE 模型，包括 DeepSeek 等，其量化行为不变（原来 tensor 已是 [1] 的模型不受影响）。团队影响极小，属于单行热修复。 - 风险标记：单行修复但缺少测试，极端配置下可能空 tensor

关联脉络

- 暂无明显关联 PR