PR #40794 完整报告

vllm-project/vllm

[Bugfix][MoE] Unpad routed output before shared expert add [Fixes #35949]

合并时间：2026-04-24 19:53

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/40794

执行摘要

- 一句话：修复 MoE 路由输出未截断导致张量形状不匹配
- 推荐动作：PR 改动小但重要，修复了一个影响 NVFP4 量化模型的回归。建议快速合入。对于 MoE runner 的维护者，建议后续添加对填充场景的单元测试，覆盖 _maybe_pad_hidden_states 不同填充量的情况。

功能与动机

修复由 PR #35949 引入的回归：当 TRTLLM NVFP4 MoE 通过 align_trtllm_fp4_moe_hidden_dim_for_fi 填充隐藏维度时，路由专家输出（如 2816）与共享专家输出（2688）相加会导致形状不匹配，Dynamo 在 fake tensor 追踪时抛出异常。PR body 引用了 NVIDIA-Nemotron-3-Nano-30B-A3B-NVFP4 模型作为复现场景。

实现拆解

1. 记录填充前维度：在 moe_runner.py 的 forward() 中，调用 _maybe_pad_hidden_states 之前，记录 routed_hidden_dim = hidden_states.shape[-1]。
2. 标记是否填充：调用 _maybe_pad_hidden_states 后，通过比较 hidden_states.shape[-1] > routed_hidden_dim 得到 hidden_dim_was_padded 布尔值。
3. 截断路由输出：在 _unpack(result) 后，若 hidden_dim_was_padded 为真，则执行 fused_output = fused_output[..., :routed_hidden_dim]，将路由输出截断回原始维度。
4. 后续流程不变：截断后的 fused_output 与 shared_output 按原逻辑相加、变换和规约。

关键文件：

- vllm/model_executor/layers/fused_moe/runner/moe_runner.py（模块 MoE Runner；类别 source；类型 data-contract；符号 MoERunner.forward）：MoE 前向逻辑入口，修复路由输出维度截断问题。

关键符号：MoERunner.forward

关键源码片段

vllm/model_executor/layers/fused_moe/runner/moe_runner.py

MoE 前向逻辑入口，修复路由输出维度截断问题。

```
def forward(self, hidden_states, router_logits):
```

```
# ... 路由输入变换 ...
# Record original hidden dim before potential padding
routed_hidden_dim = hidden_states.shape[-1]
hidden_states, og_hidden_dim = self._maybe_pad_hidden_states(
    shared_experts_input, hidden_states)
hidden_dim_was_padded = hidden_states.shape[-1] > routed_hidden_dim

result = self._forward_entry(hidden_states, router_logits, ...)

shared_output, fused_output = _unpack(result)
# Truncate fused output back to original dim if padding was applied
if hidden_dim_was_padded:
    fused_output = fused_output[..., :routed_hidden_dim]

# ... 规约、缩放、输出变换、加法 ...
if shared_output is not None:
    result = shared_output + fused_output # now same shape
else:
    result = fused_output
# ...
```

评论区精华

主要讨论围绕潜在的回归风险：@bnellnm 指出此修改可能再次破坏 lora/test_gptoss_tp.py::test_gpt_oss_lora_tp2 测试。维护者 @netanel-haber 在评论中引用了修复该测试的关联 PR #40865。此外，@tomeras91 询问该方案在 latent MoE 下的工作方式，得到图文解释，确认无影响。

- GPTOSS LoRA 测试回归 (testing): @netanel-haber 引用关联 PR #40865 解决了该问题。
- Latent MoE 兼容性 (design): @netanel-haber 提供了 latent MoE 架构图，说明 latent 投影操作会先对齐维度，因此不受影响。

风险与影响

- 风险：
  1. 回归风险：@bnellnm 指出可能破坏 GPTOSS LoRA 测试，但已由关联 PR #40865 修复。
  2. 仅影响填充场景：逻辑仅在 hidden_dim_was_padded 为真时执行截断，非填充路径无变化。
  3. 低风险：修改仅 6 行，逻辑清晰，且仅影响 NVFP4 等启用填充的量化配置。
- 影响：
  1. 用户影响：修复了使用 TRTLLM NVFP4 量化且 MoE 隐层被填充的模型（如 NVIDIA-Nemotron-3-Nano-30B-A3B-NVFP4）的推理错误，对普通模型无影响。
  2. 系统影响：无性能退化，仅增加一次维度比较和可能的切片操作。
  3. 团队影响：与 PR #35949 构成完整功能链，需合入同一版本。 - 风险标记：回归风险，缺少测试覆盖

关联脉络

- PR #35949 [MoE] Refactor MoE runner to handle shared experts: 本 PR 修复了 #35949 引入的回归：共享 / 路由输出加法移入 MoERunner 后未先截断路由输出。
- PR #40865 [Bugfix] Fix GPTOSS LoRA test regression: 关联修复本 PR 可能导致的 GPTOSS LoRA 测试回归。