PR #42169 完整报告

vllm-project/vllm

[Bugfix] Fix DeepSeek v4 topk numerical issue for unaligned max-model-len

合并时间：2026-05-10 11:30

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/42169

执行摘要

本 PR 修复了 DeepSeek V4 在设置特定 --max-model-len（如 900000，非 256 对齐）时出现的 token 分布偏移和准确性下降问题。根因是 CUDA topk 内核中错误使用了 logits.size(1) 获取 stride，而 deepgemm 的 logits 输出可能因填充而不连续，导致 stride 计算错误。修复仅两行，将 size(1) 替换为 stride(0) 后恢复正常。

功能与动机

DeepSeek V4 在 --max-model-len 900000 等非 256 对齐值时，生成 token 分布出现显著偏移（平均 token 数从 ~12000 飙升至 ~26000），GPQA 准确率从 ~89% 降至 ~82%，且出现 5% 无响应。调查发现，deepgemm 的 fp8_fp4_paged_mqa_logits 函数（参考）会对 max_model_len 维度进行 256 对齐填充，然后切片到实际长度，导致 logits tensor 的内存不连续。而 topk kernel 使用 logits.size(1) 计算 stride，在连续时等于 stride(0)，但不连续时则错误。

实现拆解

1. 定位 root cause：在 csrc/topk.cu 中，launch_persistent_topk 和 persistent_topk 两个函数均使用 const int64_t stride = logits.size(1) 来获取 logits 在行方向上的内存步长。当 logits 最后一维被 padding 后不连续时，size(1) 与实际 stride 不匹配。
2. 修复：将两处 stride 定义改为 logits.stride(0)，即使用 tensor 的内存步长而非维度大小。这样无论 logits 是否连续，都能正确获取相邻行之间的实际字节偏移。
3. 验证：修复后 token 分布恢复正常，GPQA 准确率恢复至 ~88.9%，与不设置 --max-model-len 时的结果一致。

csrc/topk.cu

修复核心：将 stride 计算从 logits.size(1) 改为 logits.stride(0)，解决非连续 logits 导致的 topk 数值错误。

```
// csrc/topk.cu

void launch_persistent_topk(const torch::Tensor& logits,
                            ...) {
  const int64_t num_rows = logits.size(0);
  // 修复：使用 stride(0) 代替 size(1)，因为 deepgemm 可能对最后一维
  // 进行 256 对齐填充，导致 size(1) != stride(0) 当 max_model_len 未对齐时。
  const int64_t stride = logits.stride(0);
```

```
  // ... kernel launch ...
}

void persistent_topk(const torch::Tensor& logits,
                     ...) {
  const int64_t num_rows = logits.size(0);
  // 同上，统一使用 stride(0) 确保内存步长正确。
  const int64_t stride = logits.stride(0);
  // ... kernel 执行 ...
}
```

评论区精华

无 reviewer 讨论，仅有 bot 自动评论和 zyongye 的批准。

风险与影响

- 风险：极低。两行简单赋值修改，不影响其他逻辑。在 logits 连续时语义等价。
- 影响：影响 DeepSeek V4 在非标准 --max-model-len 配置下的推理正确性，对使用大 max_model_len（如 900k）的用户至关重要。

关联脉络

- 与 PR #41428（DSv4 fused Indexer Q quant kernel 优化）同为 DeepSeek V4 的 kernel 层改进。
- 与 PR #41957（DSv4 PD 模式修复）同为 DeepSeek V4 bugfix。
- 属于 DeepSeek 模型系列持续优化的一部分。