PR #25396 完整报告

sgl-project/sglang

fix: fix deepseek v4 CP error

合并时间：2026-05-19 17:04

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/25396

执行摘要

- 一句话：修复 DeepSeek V4 CP 中张量不连续崩溃
- 推荐动作：值得精读，尤其是理解 JIT 内核与张量连续性的依赖关系。建议同步检查 _compute_kv_to_cache 的类似问题。

功能与动机

运行 DeepSeek V4 模型时，启用 --enable-nsa-prefill-context-parallel 后，出现 Tensor is not contiguous as expected 错误。错误源于 qkv_a 切片操作导致 kv 张量不连续，而 fused JIT 内核要求连续内存。

实现拆解

1. 定位根因：在 python/sglang/srt/models/deepseek_v4.py 的 _compute_kv_bf16 方法中，第 388 行 kv = qkv_a[..., self.q_lora_rank :] 产生不连续张量，随后传入 fused_norm_rope_inplace 导致 JIT 内核崩溃。
2. 修复：在切片后、调用 fused_norm_rope_inplace 前插入 kv = kv.contiguous()，确保内存连续。
3. 测试验证：通过 curl 请求验证模型输出正常，无 NaN 或崩溃。

关键文件：

- python/sglang/srt/models/deepseek_v4.py（模块 模型；类别 source；类型 data-contract；符号 _compute_kv_bf16）：核心模型文件，修复 _compute_kv_bf16 中张量不连续导致的 JIT 内核崩溃。

关键符号：_compute_kv_bf16, fused_norm_rope_inplace

关键源码片段

python/sglang/srt/models/deepseek_v4.py

核心模型文件，修复 _compute_kv_bf16 中张量不连续导致的 JIT 内核崩溃。

```
# python/sglang/srt/models/deepseek_v4.py
# _compute_kv_bf16 方法：用于 NSA prefill-CP 场景，需要 bf16 kv 进行跨 rank all-gather
    def _compute_kv_bf16(
        self,
```

```
        x: torch.Tensor,
        positions: torch.Tensor,
        qkv_a: Optional[torch.Tensor] = None,
    ) -> torch.Tensor:
        """Bf16-kv path used by the NSA prefill-CP case (needs all-gather)."""
        if qkv_a is not None:
            # 切片操作可能产生不连续张量
            kv = qkv_a[..., self.q_lora_rank :]
        else:
            kv, _ = self.wkv(x)
        # 确保 kv 连续，防止 fused_norm_rope_inplace JIT 内核崩溃
        kv = kv.contiguous()
        fused_norm_rope_inplace(
            kv,
            self.kv_norm.weight.data,
            self.eps,
            self.freqs_cis,
            positions,
        )
        return kv
```

评论区精华

AI 审查建议将 .contiguous() 仅应用于切片路径以提升性能，并指出 _compute_kv_to_cache 方法也存在类似问题。作者选择当前修复方式，认为在调用前统一显式调用更安全。

- contiguous() 调用位置优化 (performance): 作者选择在调用前统一添加，认为更安全。PR 合并后未进一步优化。

风险与影响

- 风险：低风险：contiguous() 在非连续张量上可能产生内存拷贝，但仅在 CP 路径（qkv_a 非 None）时发生，开销可接受。未修 _compute_kv_to_cache 可能导致类似问题，但当前修复已覆盖主要崩溃场景。
- 影响：影响范围：仅限 DeepSeek V4 模型启用 NSA prefill context parallelism 的场景。修复后该路径恢复正常运行，其他路径不受影响。
- 风险标记：缺少测试覆盖，类似问题可能存在于其他路径

关联脉络

- PR #25733 [Bug] Fix V4-Pro NaN on Blackwell by converting fp8_einsum input scale to ue8m0: 同为 DeepSeek V4 模型修复，涉及不同模块的类似稳定性问题。