PR #47083 完整报告

vllm-project/vllm

[Bug] Fix sparse attention issue for GLM5.2 non-torch compile path

合并时间：2026-06-30 06:45

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/47083

执行摘要

- 一句话：修复 GLM5.2 非 torch.compile 路径 sparse attention 内存分配问题
- 推荐动作：该 PR 是典型的细小但关键的控制流修复，变更仅 4 行。对于维护 deepseek_v32 注意力模块的开发者可以快速理解，对于普通使用者无需深入。建议部署 GLM5.2 的用户及时合入此修改。

功能与动机

在 GLM5.2 FP8 模型使用 expert parallel 和 tensor parallel 部署时，dummy run 过程中 sparse_attn_indexer 会分配大量内存（如 5280 MB），而 workspace 已被锁定（仅 1 MB），导致 AssertionError。PR body 中附带了完整的错误调用栈和修复前后的 lm_eval gsm8k 评测结果。

实现拆解

1. 调整 early return 顺序：在 vllm/models/deepseek_v32/nvidia/attention.py 的 _fused_attention 函数中，将原本位于 sparse_attn_indexer 调用之前的 if attn_metadata is None: output.zero_(); return 代码块移动到 sparse_attn_indexer 调用之后。这样在 dummy run（attn_metadata 为 None）时，sparse_attn_indexer 会被正常调用，不会因为提前返回而跳过，从而避免了 workspace 锁定后因 indexer 未调用而后续分配大内存导致的崩溃。
2. 保持逻辑等价性：非 dummy run 路径下，sparse_attn_indexer 调用后依然会执行 attn_metadata 检查，行为与原来一致。
3. 测试验证：未新增单元测试，但 PR body 提供了 lm_eval gsm8k 评测结果，修复后准确率约 94.09%-94.24%，符合预期。

关键文件：

- vllm/models/deepseek_v32/nvidia/attention.py（模块 模型层；类别 source；类型 core-logic；符号 _fused_attention）：核心修复文件，在 _fused_attention 函数中调整了早期返回的位置，确保 sparse_attn_indexer 在 dummy run 中被调用。

关键符号：_fused_attention

关键源码片段

vllm/models/deepseek_v32/nvidia/attention.py

核心修复文件，在 _fused_attention 函数中调整了早期返回的位置，确保 sparse_attn_indexer 在 dummy run 中被调用。

```
# vllm/models/deepseek_v32/nvidia/attention.py
# 在 _fused_attention 函数中，调整 early return 顺序
# 确保 dummy run (attn_metadata is None) 时 sparse_attn_indexer 仍被调用

if self.indexer is not None:
    sparse_attn_indexer(
        q_c,
        self.indexer.k_cache.prefix,
        self.indexer.k_cache.kv_cache,
        index_q_fp8,
        None, # q_scale folded into weights on the fp8 path
        None, # k unused when skip_k_cache_insert=True
        index_weights_out,
        self.indexer.quant_block_size,
        self.indexer.scale_fmt,
        self.indexer.topk_tokens,
        self.indexer.head_dim,
        self.indexer.max_model_len,
        self.indexer.max_total_seq_len,
        self.topk_indices_buffer,
        True, # skip_k_cache_insert
        False, # use_fp4_cache
        True, # skip_topk_buffer_clear (fused_norm_rope already did it)
    )

# 原本该代码块位于 sparse_attn_indexer 调用之前，现在移动至此
if attn_metadata is None:
    output.zero_()
    return
```

评论区精华

WoosukKwon 批准了该 PR 并表示感谢。无其他 review 讨论。

- 暂无高价值评论线程

风险与影响

- 风险：
  1. 回归风险低：变更仅调整了 early return 的位置，保持了原有的逻辑分支。非 torch.compile 路径下，dummy run 时原本不会走到 indexer 调用，现在提前返回前会调用 indexer，但 indexer 本身在 dummy run 下应有合理的空操作处理（如跳过 topk 缓存清除等参数已默认设置）。
  2. 性能影响无：正常推理路径下（attn_metadata 非 None）代码执行顺序不变。

3. 仅影响特定模型：仅适用于 deepseek_v32 的 nvidia 注意力实现中的 GLM5.2 模型。 - 影响：影响范围：仅针对使用 GLM5.2 FP8 模型并启用 expert parallel 和 tensor parallel 的部署场景。修复后这些场景不再因 workspace 内存分配限制而崩溃。影响程度：中等，对受影响用户是关键的运行时修复。 - 风险标记：核心路径变更

关联脉络

- 暂无明显关联 PR