PR #26515 完整报告

sgl-project/sglang

Allow Optional key/value in unified_attention_with_output split-op (MLA absorb fix)

合并时间：2026-05-28 16:04

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/26515

执行摘要

- 一句话：修复 MLA 吸收路径在分段 CUDA 图中因 key/value 为 None 崩溃
- 推荐动作：值得精读，特别是理解分段 CUDA 图（PCG/BCG）与 MLA 吸收路径的交互。设计上保持与非分段路径的兼容性，但缺少单元测试验证是一个风险。

功能与动机

修复 unified_attention_with_output 在 MLA 吸收路径且使用分段 / 可中断 CUDA 图（PCG/BCG）运行器时因 key/value 为 None 而崩溃的问题。非分段路径（直接后端调用）不受影响。

实现拆解

1. 修改参数类型签名：在 python/sglang/srt/layers/radix_attention.py 的 unified_attention_with_output 函数中，将 key 和 value 参数类型从 torch.Tensor 改为 Optional[torch.Tensor]，与 RadixAttention.forward 的契约一致。
2. 添加 None 守卫：在 real_num_tokens 切片逻辑中，对 key 和 value 分别添加 if key is not None 和 if value is not None 判断，仅在非 None 时执行切片。
3. 保持兼容：非 MLA 后端始终传递非 None 的 K/V，行为无变化；仅 MLA 吸收路径（传递 None）受影响。
4. 测试验证：在 PR body 中提供了 MLA Triton EXTEND 在 PCG 和 BCG 运行器下的通过率测试。

关键文件：

- python/sglang/srt/layers/radix_attention.py（模块 注意力层；类别 source；类型 core-logic；符号 unified_attention_with_output）：核心变更文件，修改了 split-op 入口函数的参数类型和切片逻辑，修复 MLA 吸收路径在 PCG/BCG 下的崩溃。

关键符号：unified_attention_with_output

关键源码片段

python/sglang/srt/layers/radix_attention.py

核心变更文件，修改了 split-op 入口函数的参数类型和切片逻辑，修复 MLA 吸收路径在 PCG/BCG 下的崩溃。

```
# python/sglang/srt/layers/radix_attention.py
# split-op 入口函数，被分段 CUDA 图运行器调用以分派到活跃的注意力后端。
# MLA 吸收路径调用 RadixAttention.forward(q, k=None, v=None) 时，
# 由于压缩潜在 KV 从 token-to-kv-pool 中读取，无需外部传入。
@register_custom_op(mutates_args=["output"])
@register_split_op()
def unified_attention_with_output(
    query: torch.Tensor,
    key: Optional[torch.Tensor], # 变更为 Optional，兼容 MLA 传 None
    value: Optional[torch.Tensor], # 同上
    output: torch.Tensor,
    save_kv_cache: bool,
    layer_id: int,
    *,
    q_rope: Optional[torch.Tensor] = None,
    k_rope: Optional[torch.Tensor] = None,
    sinks: Optional[torch.Tensor] = None,
    # MLA / TRT-LLM / NSA 路径通过 RadixAttention.forward(**kwargs) 传递；
    # 当 --enforce-piecewise-cuda-graph 启用时，它们必须出现在 schema 中。
    cos_sin_cache: Optional[torch.Tensor] = None,
    is_neox: Optional[bool] = None,
    llama_4_scaling: Optional[torch.Tensor] = None,
    topk_indices: Optional[torch.Tensor] = None,
) -> None:
    context = get_forward_context()
    forward_batch = context.forward_batch
    attention_layers = context.attention_layers
    attention_layer = attention_layers[layer_id]
    real_num_tokens = forward_batch.num_token_non_padded_cpu

    query = query[:real_num_tokens]
    # 仅当 key/value 非 None 时才切片，避免 MLA 传 None 时崩溃
    if key is not None:
        key = key[:real_num_tokens]
    if value is not None:
        value = value[:real_num_tokens]
    # ... 后续 kwargs 组装和后端调用保持不变
```

评论区精华

无 review 讨论。PR 作者自审自合，仅有 bot 自动评论每日配额已达上限。

- 暂无高价值评论线程

风险与影响

- 风险：低风险。变更集中在一处函数，仅添加了可选类型和条件切片，不会影响非 MLA 路径。但没有新增单元测试，长期依赖后续的注意力后端单元测试矩阵 PR。

- 影响：影响范围：仅影响使用 MLA 吸收路径且启用分段 CUDA 图（PCG/BCG）的用户。修复前崩溃，修复后正常运行。对非 MLA 模型无影响。
- 风险标记：缺少测试覆盖

关联脉络

- PR #24737 Support Flashinfer Cute-DSL MLA attention: 同为 MLA 注意力相关 PR，引入了新的 MLA 后端，本 PR 修复了该后端在分段 CUDA 图下的兼容性问题。
- PR #26382 Enable Kimi-K2.5 piecewise CUDA graph: 启用了分段 CUDA 图，本 PR 修复了分段 CUDA 图与 MLA 交互时的崩溃问题。