PR #23268 完整报告

sgl-project/sglang

【NPU】【bugfix】accuracy fix when enable both nsa cp and prefixcache

合并时间：2026-04-28 09:08

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/23268

执行摘要

- 一句话：修复 NSA CP 和 Prefix Cache 同时开启时的精度问题
- 推荐动作：这是针对特定硬件后端（NPU）和配置组合的定向修复，逻辑清晰。建议在 NPU CI 中增加同时启用 nsa_cp 和 prefixcache 的精度测试，防止未来回归。对于 GPU 用户无需关注。

功能与动机

PR body 明确指出：'when enable both nsa cp and prefixcache, inference accuracy is abnormal.' 这是一个在 NPU 上使用 DeepSeek V2/V3 模型时遇到的精度缺陷修复。

实现拆解

1. python/sglang/srt/layers/attention/nsa/nsa_indexer.py：在 Prefill 且启用 nsa_cp 的分支中，修正 actual_seq_lengths_kv 的计算。当存在前缀缓存时（即 sum(forward_batch.extend_prefix_lens_cpu) > 0），需要在 kv_len_prev_tensor 和 kv_len_next_tensor 的基础上加上 forward_batch.extend_prefix_lens.squeeze()，以使 KV 长度正确反映缓存前缀的长度。
2. python/sglang/srt/hardware_backend/npu/modules/deepseek_v2_attention_mla_npu.py：在非 Scatter 模式的 MLA 预处理中，为 fused_split_qk_norm 调用增加条件 not nsa_use_prefill_cp(forward_batch)。当启用 nsa_cp 时，跳过 fused 的 Q/K 拆分和 LayerNorm 操作，回退到非 fused 的 split 加逐层 norm 路径，避免因融合操作与 cp 元数据不兼容导致的精度问题。

关键文件：

- python/sglang/srt/layers/attention/nsa/nsa_indexer.py（模块 注意力机制；类别 source；类型 core-logic；符号 forward_npu）：核心修复文件：修正了当 prefixcache 与 nsa_cp 同时启用时 KV 序列长度的计算，在原有 kv_len 基础上累加 extend_prefix_lens 以正确反映缓存前缀长度。
- python/sglang/srt/hardware_backend/npu/modules/deepseek_v2_attention_mla_npu.py（模块 注意力机制；类别 source；类型 core-logic；符号 forward_dsa_prepare_npu）：辅助修复：通过添加条件 not nsa_use_prefill_cp(forward_batch) 禁用 fused_split_qk_norm 优化路径，避免在 nsa_cp 下因融合操作与 cp 元数据冲突导致精度异常。

关键符号：forward_npu, forward_dsa_prepare_npu

关键源码片段

python/sglang/srt/hardware_backend/npu/modules/deepseek_v2_attention_mla_npu.py

辅助修复：通过添加条件 not nsa_use_prefill_cp(forward_batch) 禁用 fused_split_qk_norm 优化路径，避免在 nsa_cp 下因融合操作与 cp 元数据冲突导致精度异常。

```
# python/sglang/srt/hardware_backend/npu/modules/deepseek_v2_attention_mla_npu.py
# 跳过 fused_split_qk_norm 的条件：当开启 NSA CP 时，该融合操作与 cp 元数据不兼容
# 原条件只检查 batch 大小，现在额外检查是否使用 prefill cp
if fused_qkv_a_proj_out.shape[0] < 65535 and not nsa_use_prefill_cp(
    forward_batch
):
    # nsa_cp 未启用时走融合优化路径
    q_lora, k_nope, k_pe = fused_split_qk_norm(
        fused_qkv_a_proj_out,
        m.q_a_layernorm,
        m.kv_a_layernorm,
        m.q_lora_rank,
        m.kv_lora_rank,
        m.qk_rope_head_dim,
        eps=m.q_a_layernorm.variance_epsilon,
    )
else:
    # nsa_cp 启用或 batch 太大时走标准的 split + 逐层 norm 路径
    q, latent_cache = fused_qkv_a_proj_out.split(
        [m.q_lora_rank, m.kv_lora_rank + m.qk_rope_head_dim], dim=-1
    )
    q = m.q_a_layernorm(q)
    q_lora = q.clone()
    k_nope, k_pe = latent_cache.unsqueeze(1).split(
        [m.kv_lora_rank, m.qk_rope_head_dim], dim=-1
    )
    k_nope = m.kv_a_layernorm(k_nope)
```

评论区精华

PR 没有 review 评论，只有 maintainer 的两次 Approval。不过从实现可以推断，直接原因是在 cp 和 prefix cache 同时启用时，之前未考虑 prefix 对 KV 序列长度的影响，导致 attention 范围不正确。

- 暂无高价值评论线程

风险与影响

- 风险：

1. 回归风险：条件分支修改了 fused_split_qk_norm 的触发条件，当 nsa_cp 启用时强制走非 fused 路径，可能带来轻微性能开销。
2. 兼容性：此修复仅针对 NPU 后端，修改集中在 NPU 专用模块，不影响 GPU/AMD 等其他后端。
3. 测试覆盖：无直接测试文件变更，建议补充同时开启 nsa_cp 和 prefix cache 的精度测试用例。 - 影响：直接影响面较小，仅修复了在 NPU 上使用 DeepSeek 模型且同时开启 NSA CP 和前缀缓存时的精度问题。影响的用户群体是 SGLang 的 NPU 用户（如华为昇腾）。风险低，修复明确。 - 风险标记：NPU 专用路径，缺少测试覆盖

关联脉络

- PR #23885 [Disagg] Finalize routed_experts_output in process_batch_result_disagg_prefill: 同为 NPU + DeepSeek 相关的精度修复，修改了相似的 attention 模块，可能存在重叠或依赖关系。