PR #28118 完整报告

sgl-project/sglang

【bugfix】The NPU's forward_dsa_prepare_npu also needs special handling for is_nextn

合并时间：2026-06-15 21:31

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/28118

执行摘要

- 一句话：修复 NPU 平台 DeepSeek 模型 is_nextn 特殊处理缺失
- 推荐动作：该 PR 值得精读，尤其是对 NPU 后端和 speculative decoding 感兴趣的开发者。它是一个典型的平台特定 bugfix，展示了如何处理不同硬件后端之间的行为差异。

功能与动机

在 NPU 平台上运行 DeepSeek 模型时，如果启用了 speculative decoding (Eagle draft 模式)，会触发 RuntimeError: split_with_sizes expects split_sizes to sum exactly to 512。错误栈显示从 ascend_backend.py:1587 的 q.split 开始，问题根因是当 is_nextn 为 True 且 prev_topk_indices 为 None 时，之前的逻辑错误地跳过了 indexer 调用，导致 topk_indices 未被正确计算，进而影响了注意力层的输入维度。PR 描述明确提到"The NPU's forward_dsa_prepare_npu also needs special handling for is_nextn"。

实现拆解

1. 修改 python/sglang/srt/hardware_backend/npu/modules/deepseek_v2_attention_mla_npu.py 中 forward_dsa_prepare_npu 函数的条件判断（第 406 行）。
2. 将原来的条件 if m.skip_topk: 改为 if not m.skip_topk or (m.is_nextn and prev_topk_indices is None):。
3. 同时将 else 分支包裹的 topk_indices = prev_topk_indices 移到新的 else 分支中。 这样当处于 nextn 阶段且之前没有保存的 prev_topk_indices 时，会强制调用 m.indexer 重新计算 topk_indices，避免了因 skip_topk 而跳过 indexer 导致后续错误。

关键文件：

- python/sglang/srt/hardware_backend/npu/modules/deepseek_v2_attention_mla_npu.py（模块 NPU 后端；类别 source；类型 core-logic；符号 forward_dsa_prepare_npu）：核心修改文件，在 forward_dsa_prepare_npu 函数中增加了对 is_nextn 的特殊处理，确保在 nextn 阶段且 prev_topk_indices 为空时强制调用 indexer。

关键符号：forward_dsa_prepare_npu

关键源码片段

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

核心修改文件，在 forward_dsa_prepare_npu 函数中增加了对 is_nextn 的特殊处理，确保在 nextn 阶段且 prev_topk_indices 为空时强制调用 indexer。

```
def forward_dsa_prepare_npu(...):
    ...
    # 原逻辑：if m.skip_topk: topk_indices = prev_topk_indices; else: topk_indices = m.indexer(...)
    # 新逻辑：当 is_nextn 且 prev_topk_indices 为 None 时，也强制调用 indexer
    if not m.skip_topk or (m.is_nextn and prev_topk_indices is None):
        topk_indices = m.indexer(
            hidden_states,
            q_lora,
            positions,
            forward_batch,
            m.layer_id,
            layer_scatter_modes,
            dynamic_scale,
        )
    else:
        topk_indices = prev_topk_indices

    return (
        q_pe,
        k_pe,
        q_nope_out,
        k_nope,
        topk_indices,
        forward_batch,
        zero_allocator,
        positions,
    )
```

评论区精华

没有 review 评论。只有机器人和作者触发的 /tag-and-rerun-ci 和 /run_ci 命令。

- 暂无高价值评论线程

风险与影响

- 风险：变更仅涉及一个文件中的一行逻辑，风险较低。但需要确保：
  1. 该修改不会影响非 NPU 平台（因为函数名明确为 forward_dsa_prepare_npu，只在 NPU 后端使用）。
  2. 在 is_nextn 为 True 且 prev_topk_indices 不为 None 时，行为不变（仍走 topk_indices = prev_topk_indices）。

3. 没有测试覆盖该特定场景，回归风险依赖于集成测试。 - 影响：影响范围：NPU 平台上所有使用 DeepSeek-V2/MoE 模型并启用 speculative decoding (Eagle) 的用户。修复后，这些用户不会再遇到 split 维度不匹配的运行时错误。对非 speculative decoding 场景无影响。 - 风险标记：缺少测试覆盖

关联脉络

- PR #27802 bugfix revise interface get cpu copy for npu mem pool to align with gpu: 另一个 NPU 平台的相关 bugfix，显示了 NPU 后端的持续维护。