PR #36170 完整报告

sgl-project/sglang

[NPU] [BugFix] Fix discontinuous input for FIA operator in GLM4.7-Flash

合并时间：2026-08-29 15:03

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/36170

执行摘要

- 一句话：修复 NPU FIA 算子因非连续输入导致的运行时报错
- 推荐动作：建议精读该 PR，了解 FIA 算子对输入连续性的要求以及 NPU 后端对张量布局的处理方式。虽然变更很小，但体现了硬件后端适配中常见的连续性约束问题，值得记录。

功能与动机

PR 描述指出，GLM4.7-Flash 在调用 FIA 算子时，k_rope / k_value / v_value 张量可能不连续，而新版算子已移除内部的 AutoContiguous 逻辑，需要调用方保证输入张量连续性，否则会引起算子运行时失败。此修复正是为恢复该场景的正常推理。

实现拆解

1. 定位问题：在 ascend_backend.py 的 forward_extend 方法中，FIA 算子的调用路径上，v 张量通过切片 v[None, q_len_offset : q_len_offset + q_len] 传入，该切片视图不保证连续。
2. 应用补丁：对该切片视图调用 .contiguous() 方法后再传入 FIA 算子，确保输入张量满足新版本算子的连续性要求。
3. 配套测试：本次变更未包含新增单元测试，但后续通过 CI 和 review 确认修复有效，且未引入回归。

关键文件：

- python/sglang/srt/hardware_backend/npu/attention/ascend_backend.py（模块 NPU 后端；类别 source；类型 core-logic）：核心修复文件，对 FIA 算子的 v 输入追加 contiguous() 调用，解决非连续输入导致的运行失败。

关键符号：未识别

关键源码片段

python/sglang/srt/hardware_backend/npu/attention/ascend_backend.py

核心修复文件，对 FIA 算子的 v 输入追加 contiguous() 调用，解决非连续输入导致的运行失败。

```
# python/sglang/srt/hardware_backend/npu/attention/ascend_backend.py
# 调用 FIA 算子时，v 张量经切片后可能不连续，需保证传入算子的张量连续
```

```
attn_output[q_len_offset : q_len_offset + q_len] = (
    torch.ops.npu.npu_fused_infer_attention_score(
        q[None, q_len_offset : q_len_offset + q_len],
        k[None, q_len_offset : q_len_offset + q_len],
        # 关键修复：对 v 的切片追加 .contiguous()，满足新版 FIA 算子对输入连续性的要求
        v[None, q_len_offset : q_len_offset + q_len].contiguous(),
        num_heads=layer.tp_q_head_num,
        num_key_value_heads=layer.tp_k_head_num,
        input_layout="BSND", # 注意：TND 布局不支持 q_heads != k_heads
        atten_mask=self.fia_mask.unsqueeze(0),
        sparse_mode=3 if q_len != 1 else 0,
        scale=layer.scaling,
        next_tokens=0,
    )[0]
)
```

评论区精华

本 PR 的 review 评论为空，仅有一个自动 bot 的批准，因此没有实质性的技术讨论记录。

- 暂无高价值评论线程

风险与影响

- 风险：风险较低。变更仅对传递给 FIA 算子的 v 张量追加 contiguous() 调用，该操作在 v 已是连续张量时几乎零开销。但需要注意，当前仅对 v 做了处理，而 PR 描述中提到 k_rope / k_value 也可能不连续，若后续需要，可能需同步处理。另外，由于缺少单元测试，如果未来重构该调用路径，可能回归此问题。
- 影响：影响范围限于 NPU 硬件后端上运行 GLM4.7-Flash 模型时的 extend 阶段，修复了算子运行失败问题，使该场景恢复可用。对 CPU、CUDA 等其他后端无影响，也不影响 decode 路径。
- 风险标记：缺少测试覆盖

关联脉络

- PR #35021 [NPU] add causal conv1d for ascend kda backend: 同一 NPU 后端代码路径，涉及 ascend 后端注意力相关修改，可参考其性能优化方式。