

PR #28872 完整报告

sgl-project/sglang

【NPU】adapt_fused_rope_qk_mqa_optimize

合并时间: 2026-06-25 08:55

原文链接: <http://prhub.com.cn/sgl-project/sglang/pull/28872>

执行摘要

- 一句话: NPU fused_rope_qk_mqa 阈值条件修复
- 推荐动作: 建议关注此改动是否引入 NPU 上的精度或性能回归, 可通读 fused_rope_qk_mqa 算子实现确认其上限约束。

功能与动机

NPU 上 fused_rope_qk_mqa 的调用条件过于严格, 使用 `query.shape[0] * query.shape[1] < 65535` 可能导致某些形状无法触发融合算子, 本质是修复条件使其与算子的实际支持范围一致。

实现拆解

1. 在 `python/sglang/srt/layers/rotary_embedding/base.py` 的 `forward_npu` 方法中, 将 `fused_rope_qk_mqa` 的调用条件从 `query.shape[0] * query.shape[1] < 65535` 简化为 `query.shape[0] < 65535`, 移除对第二维的乘法检查。
2. 同时删除原条件周围多余的空格和换行, 使代码更紧凑。
3. 该变更只涉及一行核心逻辑改动, 其余为格式化清理。

关键文件:

- `python/sglang/srt/layers/rotary_embedding/base.py` (模块 旋转位置编码; 类别 source ; 类型 core-logic) : 核心修改文件, 调整 `fused_rope_qk_mqa` 的调用条件, 直接影响 NPU 上的 RoPE 计算路径。

关键符号: `forward_npu`

关键源码片段

`python/sglang/srt/layers/rotary_embedding/base.py`

核心修改文件, 调整 `fused_rope_qk_mqa` 的调用条件, 直接影响 NPU 上的 RoPE 计算路径。

```
# 变更前: 条件为 query.shape[0] * query.shape[1] < 65535
# 变更后: 仅检查 query.shape[0] < 65535
if fused_rope_qk_mqa is not None and query.shape[0] < 65535:
    return fused_rope_qk_mqa(
        query,
```

```
        key,
        cos_sin,
        self.rotary_dim,
        self.is_neox_style,
    )
else:
    return self.forward_native(positions, query, key, offsets)
```

评论区精华

无相关 review 讨论，PR 获得批准后合并。

- 暂无高价值评论线程

风险与影响

- 风险：风险极低：改动仅涉及一个条件表达式的简化，不会改变无融合算子时的回退行为。但需注意若 `fused_rope_qk_mqa` 内部对第二维有隐式限制，放宽条件可能导致运行时错误，不过从上下文看该算子应为通用实现。
- 影响：影响范围仅限于 NPU 后端上使用 `fused_rope_qk_mqa` 的场景，缩小了原有限制，使更多形状可以受益于融合算子加速。
- 风险标记：条件表达式修改

关联脉络

- PR #29042 [NPU] Fix the DeepSeek-V2-Coder model accuracy issue: 同为 NPU 相关的 bugfix，涉及模型精度修复。
- PR #26724 [NPU] adaptation to support operator FA3 in deterministic inference on NPU.: 同为 NPU 上算子适配的 feature，可能涉及类似的条件逻辑调整。