

PR #30458 完整报告

sgl-project/sglang

[NPU] [BUGFIX] Fix input parameters of swiglu_oai operator

合并时间: 2026-07-14 14:16

原文链接: <http://prhub.com.cn/sgl-project/sglang/pull/30458>

执行摘要

- 一句话: 修复 NPU SwiGLU OAI 算子输入参数错误
- 推荐动作: 该 PR 改动量小且方向正确, 建议精读以了解 NPU 上 MoE 激活函数的调用链路。但由于缺少测试覆盖, 建议后续补充针对 swiglu_oai_triton 的回归测试。

功能与动机

PR body 明确指出: 'Since pre-transposed storage is removed for layer.w13_weight, the swiglu_oai operator throws errors when fetching dimension information from layer.w13_weight.' 解决 NPU 运行时 MoE 激活函数层的崩溃问题。

实现拆解

1. 修改导入语句: 将 `from sgl_kernel_npu.activation.swiglu_oai import swiglu_oai` 改为 `from sgl_kernel_npu.activation.swiglu_oai import swiglu_oai_triton`。
2. 改变调用签名: 原来传入 `layer` 对象让算子内部从 `layer.w13_weight.shape` 推导 `gate_up` 维度, 现在改为直接传入: `hidden_states, hidden_states.shape[-1], gemm1_alpha, gemm1_clamp_limit`。
3. 消除对 `weight` 形状的依赖: 由于 `w13_weight` 现在按非转置格式存储 (在 `grouped_matmul` 中动态转置), 其 `shape[2]` 是 `hidden` 维度而非 `gate_up` 维度, 显式传参避免了这一歧义。
4. 仅修改一个文件 `python/sglang/srt/layers/quantization/unquant.py`, 正向传播路径未涉及测试文件变更。

关键文件:

- `python/sglang/srt/layers/quantization/unquant.py` (模块 量化层; 类别 `source`; 类型 `dependency-wiring`): 核心修复文件, 修改了 `swiglu_oai` 激活函数的导入和调用方式, 修复因 `w13_weight` 预转置存储移除导致的维度错误。

关键符号: `forward_npu`

关键源码片段

`python/sglang/srt/layers/quantization/unquant.py`

核心修复文件，修改了 swiglu_oai 激活函数的导入和调用方式，修复因 w13_weight 预转置存储移除导致的维度错误。

```
# python/sglang/srt/layers/quantization/unquant.py
# 在 forward_npu 函数中，activation == "npu_swiglu_oai" 分支
if self.moe_runner_config.activation == "npu_swiglu_oai":
    from sgl_kernel_npu.activation.swiglu_oai import swiglu_oai_triton

    # `hidden_states` 是 gmm1 输出，形状为 [num_tokens, 2 * inter]。
    # 原版本传入 layer 对象，由算子内部通过 `layer.w13_weight.shape[2]` 获取
    # gate_up 维度；但 w13_weight 现已改为非转置存储（在 grouped_matmul 中动态
    # 转置），shape[2] 是 hidden 维度而非 gate_up 维度，导致 view 失败。
    # 修复后改为从 activation 自身显式传入 gate_up 维：hidden_states.shape[-1]。
    hidden_states = swiglu_oai_triton(
        hidden_states,
        hidden_states.shape[-1],
        self.moe_runner_config.gemm1_alpha,
        self.moe_runner_config.gemm1_clamp_limit,
    )
```

评论区精华

无 reviewer 讨论记录。

- 暂无高价值评论线程

风险与影响

- 风险：该 PR 仅修改了 swiglu_oai 分支的调用约定，且原 layer 参数在算子内部仅用来获取维度，显式传入 shape 后逻辑等价。但由于未添加对应的测试用例，若 swiglu_oai_triton 函数签名或行为与旧版本存在差异（如参数顺序、默认值），可能导致推理结果错误。
- 影响：影响范围限于 NPU 平台上使用 npu_swiglu_oai 激活函数的 MoE 模型。修复前该分支会直接崩溃，修复后可正常运行。对其他激活函数分支（silu/gelu）无影响。
- 风险标记：缺少测试覆盖，NPU 特定逻辑

关联脉络

- 暂无明显关联 PR