

PR #31393 完整报告

sgl-project/sglang

[NPU] Determine the topk norm_type through scoring_func

合并时间: 2026-07-27 15:12

原文链接: <http://prhub.com.cn/sgl-project/sglang/pull/31393>

执行摘要

- 一句话: 根据 scoring_func 动态选择 NPU topk norm_type, 修复 DS coder v2 精度
- 推荐动作: 建议合并。这是一个针对特定回归的精准修复, 改动量小且逻辑清晰。后续可在 NPU 文档中补充 scoring_func 参数说明。

功能与动机

PR #29509 影响了 ds coder v2 lite instruct 的准确性, 因为该模型在 topk 使用了 softmax 函数。为 NPU 重构 topk 部分, 需要使 norm_type 与评分函数匹配。

实现拆解

1. 模型侧配置: 在 python/sglang/srt/models/glm4_moe_lite.py 中, 为 GLM4 MoE Lite 模型在 NPU 环境下初始化 TopK 时传递 scoring_func="sigmoid", 明确该模型使用 sigmoid 评分函数。
2. 算子侧调整: 在 python/sglang/srt/hardware_backend/npu/moe/topk.py 的 fused_topk_npu 函数中, 将原本硬编码的 norm_type=1 改为根据 topk_config.scoring_func 动态决定: 若为 "softmax" 则设为 0, 否则保持 1 (sigmoid)。
3. 配套改动: 在模型文件中导入 is_npu 工具函数并创建全局变量 _is_npu, 以便条件传递参数。

关键文件:

- python/sglang/srt/models/glm4_moe_lite.py (模块 模型定义; 类别 source; 类型 data-contract; 符号 Glm4MoeLiteMLP.init): 模型定义文件, 为 NPU 环境下的 GLM4 MoE Lite 模型传递 scoring_func='sigmoid', 触发正确的 topk 行为。
- python/sglang/srt/hardware_backend/npu/moe/topk.py (模块 NPU 内核; 类别 source; 类型 core-logic; 符号 fused_topk_npu): NPU 专用 topk 算子实现, 核心修复: 将硬编码的 norm_type 改为根据 scoring_func 动态选择。

关键符号: fused_topk_npu

关键源码片段

[python/sglang/srt/hardware_backend/npu/moe/topk.py](#)

NPU 专用 topk 算子实现，核心修复：将硬编码的 norm_type 改为根据 scoring_func 动态选择。

```
# python/sglang/srt/hardware_backend/npu/moe/topk.py
# 在 fused_topk_npu 函数中，调用 npu_moe_gating_top_k 时
# 之前：norm_type=1 # 固定 sigmoid
# 现在：norm_type=(0 if topk_config.scoring_func == "softmax" else 1)
# 注释：1 表示 sigmoid，0 表示 softmax
norm_type=(0 if topk_config.scoring_func == "softmax" else 1),
```

评论区精华

PR 提交过程中有多轮 lint 修复、冲突解决和 revert，最终合并为清晰的 4 行净改动。Review 由 [sglang-npu-bot](#) 批准，无人工评论。主要技术决策在于不在所有模型中统一设置 scoring_func，而是仅在 GLM 模型侧指定，避免影响其他模型。

- 暂无高价值评论线程

风险与影响

- 风险：风险较低：仅影响 NPU 后端的 MoE topk 路径，且显式根据 scoring_func 选择 norm_type，等价于之前硬编码 1（sigmoid）的行为对于未设置 scoring_func 的模型保持不变。但需确保所有使用 NPU 的模型在 TopK 初始化时能正确传递 scoring_func 或使用默认值；当前默认值可能导致意外走 softmax 路径。
- 影响：影响 NPU 上的 GLM4 MoE Lite 模型及类似使用 sigmoid 评分函数的模型，修复精度回归；对其他后端（CUDA）无影响。影响范围小，变更局限在 2 个文件。
- 风险标记：NPU 特定路径，模型配置变化

关联脉络

- PR #29509 [NPU] Refactor topk part for NPU: 引入的 topk 重构导致精度回归，本 PR 修复了该问题。
- PR #31280 [NPU] Acc fix for afmoe model introduced by topk refactor.: 类似性质的精度修复，同为 topk refactor 导致的 NPU 回归。