

PR #29493 完整报告

sgl-project/sglang

[NPU][Bugfix] Add scoring_func for mimo_v2

合并时间: 2026-06-29 16:34

原文链接: <http://prhub.com.cn/sgl-project/sglang/pull/29493>

执行摘要

- 一句话: 修复 MiMo-V2.5 模型 NPU 上 softmax 归一化错误
- 推荐动作: 值得合并的一行修复, 无技术争议。可作为后续添加 NPU 专用单元测试的参考点。

功能与动机

PR #29042 更新了 NPU topk 内核, 使其根据 `topk_config.scoring_func` 选择 `norm_type`, 但 `mimo_v2.py` 中的 TopK 初始化未传递 `scoring_func` 参数, 导致默认使用 softmax。对于配置了 `scoring_func='sigmoid'` 的 MiMo-V2.5 模型, NPU 内核使用了不正确的 softmax 归一化, 影响推理准确性。

实现拆解

1. 在 `python/sglang/srt/models/mimo_v2.py` 的 `MiMoV2DecoderLayer.__init__` 方法中, 创建 TopK 对象时新增关键字参数 `scoring_func=config.scoring_func`。
2. 该参数直接源自 HuggingFace 模型配置中的 `scoring_func` 字段, 与 PR #29042 中 NPU 后端对 topk kernel 的修改对齐, 确保底层内核能根据模型配置选择正确的 softmax 或 sigmoid 归一化逻辑。
3. 无其他文件变更, 无测试配套改动。

关键文件:

- `python/sglang/srt/models/mimo_v2.py` (模块 模型层; 类别 source; 类型 data-contract; 符号 `MiMoV2DecoderLayer.init`): 唯一变更文件, 在 TopK 初始化的构造函数参数列表中新增 `scoring_func=config.scoring_func`, 以匹配 PR #29042 中 NPU topk 内核的增强逻辑。该参数决定了 NPU 后端选择 softmax 还是 sigmoid 归一化, 直接影响模型准确性。

关键符号: `MiMoV2DecoderLayer.init`

关键源码片段

`python/sglang/srt/models/mimo_v2.py`

唯一变更文件, 在 TopK 初始化的构造函数参数列表中新增 `scoring_func=config.scoring_func`, 以匹配 PR #29042 中 NPU topk 内核的增强逻辑。该参数决定了 NPU 后端选择 softmax 还是 sigmoid 归一化, 直接影响模型准确性。

```
# python/sglang/srt/models/mimo_v2.py - MiMoV2DecoderLayer.__init__
# 关键变更：在创建 TopK 对象时传入 scoring_func 参数
# 该参数控制 NPU topk kernel 使用 softmax 还是 sigmoid 归一化
self.topk = TopK(
    top_k=config.num_experts_per_tok,
    renormalize=config.norm_topk_prob,
    use_grouped_topk=True,
    num_expert_group=config.n_group,
    topk_group=config.topk_group,
    correction_bias=self.gate.e_score_correction_bias,
    scoring_func=config.scoring_func, # <-- 新增行, 修正 PR #29042 的遗漏
    quant_config=quant_config,
    routed_scaling_factor=1.0,
    apply_routed_scaling_factor_on_output=(
        self.experts.should_fuse_routed_scaling_factor_in_topk
    ),
    output_format=(
        TopKOutputFormat.STANDARD if quant_config is None else None
    ),
)
```

评论区精华

该 PR 审查过程简短：review 评论为 0，仅由 [sglang-npu-bot](#) 批准。主要讨论发生在 PR 正文中：作者通过详细的精度测试结果证明修复有效性——在 MiMo-V2.5-W8A8 模型上 GSM8K 数据集准确率从错误值（作者未展示修复前数值）恢复至 0.985。

- 暂无高价值评论线程

风险与影响

- 风险：风险极低：仅添加一行参数传递，不会影响未配置 `scoring_func` 的模型（`config.scoring_func` 在 HuggingFace 配置中通常有默认值，如 `softmax`），且变更处于 MiMo 解码层初始化路径，不改变运行时逻辑。缺少单元测试覆盖是唯一可提及的不足。
- 影响：直接用户是使用 NPU 运行 MiMo-V2.5 等配置了 `scoring_func='sigmoid'` 模型的用户，错误归一化将导致推理结果异常。修复后可恢复正确行为。对其他硬件后端（CUDA）无影响，因为 TopK kernel 的 `scoring_func` 参数仅在 NPU ascend 后端生效。影响范围限定于 NPU + MiMo 组合。
- 风险标记：缺少测试覆盖

关联脉络

- PR #29042 [NPU] update topk kernel to respect scoring_func: 该 PR 更新 NPU topk kernel 以识别 `scoring_func` 参数，但未在 MiMo 模块的 TopK 调用处传递该参数，导致回归。本 PR 修复了此遗漏。