

PR #28048 完整报告

sgl-project/sglang

[Intel GPU] DeepSeek V4 10/N : Add sqrtsoftplus support to fused_topk_torch_native

合并时间: 2026-07-03 15:59

原文链接: <http://prhub.com.cn/sgl-project/sglang/pull/28048>

执行摘要

- 一句话: 为 fused_topk_torch_native 新增 sqrtsoftplus 评分函数
- 推荐动作: 小型功能性 PR, 逻辑清晰简单, 但缺少单元测试覆盖, 建议在后续整合中补充针对 sqrtsoftplus 的简单测试。值得关注的是 DeepSeek V4 模型对 MoE 评分函数的定制需求。

功能与动机

DeepSeek V4 模型需要 sqrtsoftplus 作为 MoE 路由的评分函数, 以便在原生 torch topk 路径上实现兼容。PR 标题及 commit message 均体现了这一需求。

实现拆解

在 `python/sglang/srt/layers/moe/topk.py` 的 `scoring_func_impl` 内部条件分支中新增一个 `elif scoring_func == "sqrtsoftplus"` 分支, 返回 `F.softplus(gating_output).sqrt()`。该函数被 `fused_topk_torch_native` 调用, 影响所有使用原生 topk 路径的设备。

关键文件:

- `python/sglang/srt/layers/moe/topk.py` (模块 MoE 路由; 类别 source; 类型 core-logic; 符号 `fused_topk_torch_native`, `scoring_func_impl`): 核心变更文件, 在 `scoring_func_impl` 中新增 `sqrtsoftplus` 分支。

关键符号: `fused_topk_torch_native`, `scoring_func_impl`

关键源码片段

`python/sglang/srt/layers/moe/topk.py`

核心变更文件, 在 `scoring_func_impl` 中新增 `sqrtsoftplus` 分支。

```
def fused_topk_torch_native(
    hidden_states: torch.Tensor,
    gating_output: torch.Tensor,
    topk: int,
    renormalize: bool,
    correction_bias: torch.Tensor = None,
    scoring_func: str = "softmax",
):
```

```
# 内部函数：根据 scoring_func 参数选择对应的评分函数
def scoring_func_impl(gating_output: torch.Tensor) -> torch.Tensor:
    if scoring_func == "softmax":
        return gating_output.softmax(dim=-1)
    elif scoring_func == "sigmoid":
        return gating_output.sigmoid()
    elif scoring_func == "sqrtsoftplus":
        # 新增分支：Softplus 后开平方，用于 DeepSeek V4 对标 SiLU 平滑门控
        return F.softplus(gating_output).sqrt()
    else:
        raise ValueError(f"Invalid scoring function: {scoring_func}")
```

评论区精华

1. 通用性确认：reviewer 指出变更并非 XPU 特有，而是影响所有使用该函数的设备，被作者接受并更新了 PR 标题。
 2. 测试覆盖讨论：作者最初添加了单元测试，但 reviewer 认为测试用例（对比 native_biased_topk 与 native_fused_topk）不够合理，且已有 CPU 测试文件 test/registered/cpu/test_topk.py，最终作者移除了测试文件。
- 变更通用性与 PR 标题修正 (documentation): 标题已修正，明确变更通用性。
 - 单元测试覆盖必要性 (testing): 未保留独立单元测试，依赖集成测试。

风险与影响

- 风险：低风险。变更仅新增一个条件分支，不影响现有 softmax 和 sigmoid 分支的逻辑。但该路径缺少显式测试覆盖，依赖集成测试或模型运行来验证。
- 影响：影响所有调用 fused_topk_torch_native 且传入 scoring_func="sqrtsoftplus" 的 MoE 层，主要用于 DeepSeek V4 模型。不影响默认行为。
- 风险标记：缺少测试覆盖

关联脉络

- PR #27350 Support Waterfill with MegaMoE backend: 同样修改了 python/sglang/srt/layers/moe/topk.py 中的 MoE 路由逻辑，属于同一功能区域。