PR #26634 完整报告

sgl-project/sglang

[CPU] fix incorrect index of b_ptr in fused_sigmoid_gating_delta_rule…

合并时间：2026-05-29 16:08

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/26634

执行摘要

- 一句话：修复 CPU 核中 b_ptr 索引错误
- 推荐动作：建议尽快合并。该修复为明确的 bugfix，且已有充分测试验证。对于关注 CPU 推理性能的团队值得关注。

功能与动机

本 PR 源于 issue #19484 在 review 中发现的 bug：fused_sigmoid_gating_delta_rule_update_kernel_impl 函数中 b_ptr 使用了错误的索引 ni（仅表示头索引），导致在多批次场景下读取到错误的 b 值。PR 作者 @blzheng 在 body 中特别感谢 @fadara01 指出该问题。

实现拆解

1. C++ 内核修复：在 sgl-kernel/csrc/cpu/mamba/fla.cpp 的 fused_sigmoid_gating_delta_rule_update_kernel_impl 函数中，将 beta_val = 1 / (1 + std::exp(-b_ptr[ni])) 改为 b_ptr[bi * v_num_heads + ni]，修正了索引，使其与其他参数（如 a_ptr 的索引方式）保持一致，确保在 batch_size > 1 时能正确获取每个样本的 b 值。
2. Python 测试调整（test/registered/cpu/test_mamba.py）：
   - 在 sigmoid_gating_delta_rule_update 函数中，将 g.unsqueeze(0) 和 beta.unsqueeze(0) 改为 unsqueeze(1)，使得 g 和 beta 的维度与 torch_recurrent_gated_delta_rule 的预期一致（在批次维度上添加新维度），避免维度不匹配导致的错误。
   - 将 test_fused_sigmoid_gating_delta_rule_update 测试用例从单参数改为使用 @parametrize 装饰器，支持 batch_size=[1, 4] 等多种参数组合，增强测试覆盖。同时修正了 query、key、value 的 reshape 维度以匹配新参数，并将 query_start_loc 从固定 [0, 1] 改为 torch.arange(batch_size + 1) 以支持动态批量大小。
3. 依赖导入调整：在测试文件中从 utils 导入 parametrize 以支持参数化测试。

关键文件：

- sgl-kernel/csrc/cpu/mamba/fla.cpp（模块 CPU 内核；类别 source；类型 core-logic；符号 fused_sigmoid_gating_delta_rule_update_kernel_impl）：内核的主实现文件，修复了 b_ptr 索引错误。单行变更，但影响计算正确性。
- test/registered/cpu/test_mamba.py（模块 测试；类别 test；类型 test-coverage；符号 sigmoid_gating_delta_rule_update, test_fused_sigmoid_gating_delta_rule_update）：

测试文件，同时修复了 Python 函数中的 unsqueeze 维度错误，并增强了参数化测试，覆盖多批次场景。

关键符号：fused_sigmoid_gating_delta_rule_update_kernel_impl, sigmoid_gating_delta_rule_update, test_fused_sigmoid_gating_delta_rule_update

关键源码片段

test/registered/cpu/test_mamba.py

测试文件，同时修复了 Python 函数中的 unsqueeze 维度错误，并增强了参数化测试，覆盖多批次场景。

```
# Python 参考函数，修正 g 和 beta 的 unsqueeze 维度：
# 之前使用 unsqueeze(0) 在批次前插入维度，但内核期望在批次后插入（unsqueeze(1)）。
# 修正后与 torch_recurrent_gated_delta_rule 的输入布局一致。
def sigmoid_gating_delta_rule_update(...):
    beta = b.sigmoid()
    g = -A_log.float().exp() * softplus(a.float() + dt_bias)
    return torch_recurrent_gated_delta_rule(
        query, key, value,
        g.unsqueeze(1), # 原来为 unsqueeze(0)，修正为 unsqueeze(1)
        beta.unsqueeze(1), # 同上
        initial_state, output_final_state,
        use_qk_l2norm_in_kernel=use_qk_l2norm_in_kernel,
    )

# 测试用例被参数化，现在同时测试 batch_size=1 和 4：
@parametrize(
    batch_size=[1, 4], # 新增参数化，确保多批次正确性
    num_value_heads=[32],
    head_k_dim=[128],
    head_v_dim=[128],
    num_heads=[16],
    seq_len=[1],
    attn_tp_size=[1],
)
def test_fused_sigmoid_gating_delta_rule_update(self, batch_size, ...):
    # ... 内部 reshape 使用 batch_size 替代固定值 1
    query = query.view(1, batch_size, num_heads, head_k_dim)
    key = key.view(1, batch_size, num_heads, head_k_dim)
    value = value.view(1, batch_size, num_value_heads, head_v_dim)
    # query_start_loc 也动态生成
    query_start_loc = torch.arange(batch_size + 1, dtype=torch.int32)
```

评论区精华

无 review 评论。仅有一条来自 gemini-code-assist bot 的警告（达到每日配额限制）。未发现争议点。

- 暂无高价值评论线程

风险与影响

- 风险：风险较低。核心修复仅修改一行 C++ 索引，已通过参数化测试覆盖 batch_size=1 和 4。回归风险小。但请注意，该内核仅在 CPU 路径上生效，GPU 或其他硬件平台不受影响。
- 影响：影响范围局限：主要影响 CPU 上使用 fused_sigmoid_gating_delta_rule_update 内核的 Mamba/SSM 模型推理。修复后保证了多批次（batch_size > 1）下计算的正确性。
- 风险标记：单行修改，测试覆盖

关联脉络

- PR #19484 [Mamba] Add fused_sigmoid_gating_delta_rule_update_kernel_impl cpu kernel: 本 PR 修复了在该 PR review 中发现的 bug，是该内核的后续修正。