PR #25180 完整报告

sgl-project/sglang

Fix AMX GQA extend attention

合并时间：2026-05-18 09:30

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/25180

执行摘要

- 一句话：修复 AMX CPU GQA extend attention 概率布局错误
- 推荐动作：值得精读：展示了硬件特定 bug 的定位和修复方法，以及如何通过精确控制数据布局解决问题，对理解 AMX CPU 加速细节有帮助。

功能与动机

修复 Intel AMX CPU extend attention 路径中，之前的通用转换路径写入完整的 BLOCK_N 打包块，但 GEMM 只消费 padded_n_size 列，导致 GQA extend 场景下有效概率布局被破坏，产生无效输出（如重复 '!' token）。PR body 明确说明了该 bug 的表现和根因。

实现拆解

1. 修改 softmax 输出转换逻辑（sgl-kernel/csrc/cpu/flash_attn.h）：将 copy_stub<scalar_t, BLOCK_N>(s_delta2 + row * BLOCK_N, s_delta + row * BLOCK_N) 改为 copy_stub<scalar_t>(s_delta2 + row * BLOCK_N, s_delta + row * BLOCK_N, 1.f, padded_n_size)，只复制 brgemm 需要的列数，保持行优先布局。
2. 新增回归测试（test/srt/cpu/test_extend.py）：添加 test_extend_attention_gqa_partial_extend_with_prefix 测试用例，使用 GQA 配置（H_Q=16, H_KV=4）和部分前缀（prefix=97, extend=37），验证修复后的正确性。

关键文件：

- sgl-kernel/csrc/cpu/flash_attn.h（模块 AMX 内核；类别 source；类型 core-logic；符号 flash_attn_softmax::apply）：核心修复文件：修改 softmax 输出转换，仅复制 brgemm 实际消费的列数，保持行优先布局。
- test/srt/cpu/test_extend.py（模块 测试；类别 test；类型 test-coverage；符号 test_extend_attention_gqa_partial_extend_with_prefix）：新增 GQA partial extend 回归测试，覆盖修复场景。

关键符号：flash_attn_softmax::apply

关键源码片段

sgl-kernel/csrc/cpu/flash_attn.h

核心修复文件：修改 softmax 输出转换，仅复制 brgemm 实际消费的列数，保持行优先布局。

```
// 位于 sgl-kernel/csrc/cpu/flash_attn.h
// 修复前：copy_stub<scalar_t, BLOCK_N> 复制整个 BLOCK_N 块，
// 超出 padded_n_size 的列会被错误地写入 s_delta2，
// 导致后续 brgemm(P @ V) 读取到无效的 softmax 概率。
// 修复后：指定复制长度为 padded_n_size，仅保留 brgemm 所需列，
// 保持行优先的行列式概率布局正确。
fill_stub(s_delta + row * BLOCK_N + n_size, 0.f, padded_n_size - n_size);
copy_stub<scalar_t>(s_delta2 + row * BLOCK_N, s_delta + row * BLOCK_N, 1.f, padded_n_size);
```

test/srt/cpu/test_extend.py

新增 GQA partial extend 回归测试，覆盖修复场景。

```
# test/srt/cpu/test_extend.py
# 新增 GQA partial extend 测试，模拟非零前缀 + 部分扩展的典型场景
# 修复前该场景会因概率布局错误导致输出异常（如重复 '!' token）
def test_extend_attention_gqa_partial_extend_with_prefix(self):
    self._test_extend_attention_once(
        B=1,
        N_CTX=256,
        H_Q=16, # 16 个 query head
        H_KV=4, # 4 个 KV head (GQA=4)
        D=128,
        DV=96,
        b_seq_len_prefix=[97], # 前缀 97 tokens
        b_seq_len_extend=[37], # 扩展 37 tokens
    )
```

评论区精华

无实质性 review 讨论，mingfeima 直接批准了 PR。

- 暂无高价值评论线程

风险与影响

- 风险：风险低：变更仅影响 AMX CPU 上的 extend attention 路径，且逻辑由复制整个 BLOCK_N 改为仅复制有效列，不会影响其他路径或硬件后端。新增的回归测试覆盖了修复场景。
- 影响：影响范围小，仅影响 Intel AMX CPU 上使用 GQA extend attention 的场景（如 DeepSeek V2/V3 等 GQA 模型）。修复后模型输出恢复正常，无其他影响。
- 风险标记：暂无

关联脉络

- 暂无明显关联 PR