PR #24717 完整报告

sgl-project/sglang

LFM2: pass has_initial_state to causal_conv1d_fn for prefill

合并时间：2026-05-14 12:57

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/24717

执行摘要

- 一句话：修复 LFM2 短卷积预填充状态污染
- 推荐动作：值得精读，展示了状态管理类 bug 的典型修复思路。可关注与 #23975 的关联，理解完整的修复链条。

功能与动机

修复高并发场景下 Mamba 槽位状态复用导致的多请求状态污染问题。PR body 指出，当 Mamba 槽位被释放并立即分配给新请求时，旧卷积状态残留会导致生成内容偏离（如产生无关段落）。

实现拆解

1. 计算 has_initial_state：在 Lfm2ShortConv.forward 和 Lfm2MoeShortConv.forward 的预填充分支中，根据 forward_batch.extend_prefix_lens > 0 计算布尔张量，指示每个请求是否具有有效的初始卷积状态。
2. 区分多序列与单序列：当存在多个序列（extend_start_loc 长度 > 1）时，对全部序列计算 has_initial_state；单序列时通过 [:1] 切片适配。
3. 传递给内核：将计算得到的 has_initial_state 传递给 causal_conv1d_fn，覆盖原来 None 的默认行为，使内核能正确处理状态清零或保留。

关键文件：

- python/sglang/srt/models/lfm2.py（模块 模型层；类别 source；类型 data-contract）：LFM2 短卷积前向路径，修复 has_initial_state 传递逻辑
- python/sglang/srt/models/lfm2_moe.py（模块 模型层；类别 source；类型 data-contract）：LFM2-MoE 短卷积前向路径，与 lfm2.py 对称修复

关键符号：未识别

关键源码片段

python/sglang/srt/models/lfm2.py

LFM2 短卷积前向路径，修复 has_initial_state 传递逻辑

```
# python/sglang/srt/models/lfm2.py (lines 304-319)
# 多序列分支：对所有批量序列计算 has_initial_state
```

```
    if extend_start_loc is not None and len(extend_start_loc) > 1:
        query_start_loc = torch.cat(
            [extend_start_loc, torch.tensor([T], dtype=torch.int32, device=hidden_states.device)]
        )
        cache_indices = mamba_indices.to(torch.int32)
        # True 保留先前状态（chunked-prefill 延续）；False 清零避免跨请求泄漏
        has_initial_state = forward_batch.extend_prefix_lens > 0
    else:
        # 单序列分支：只取第一个序列的索引
        query_start_loc = torch.tensor([0, T], dtype=torch.int32, device=hidden_states.device)
        cache_indices = mamba_indices[:1].to(torch.int32)
        has_initial_state = forward_batch.extend_prefix_lens[:1] > 0

    conv_out = causal_conv1d_fn(
        Bx_t,
        self.conv_weight,
        self.conv_bias,
        query_start_loc=query_start_loc,
        cache_indices=cache_indices,
        # 关键修复：传递计算后的布尔张量而非 None
        has_initial_state=has_initial_state,
        conv_states=conv_state,
        activation=None,
    ).transpose(0, 1)
```

python/sglang/srt/models/lfm2_moe.py

LFM2-MoE 短卷积前向路径，与 lfm2.py 对称修复

```
    # python/sglang/srt/models/lfm2_moe.py (lines 356-376)
    # 多序列分支（使用 new_empty 避免 cat 开销）
    if extend_start_loc is not None and len(extend_start_loc) > 1:
        query_start_loc = extend_start_loc.new_empty(len(extend_start_loc) + 1)
        query_start_loc[:-1] = extend_start_loc
        query_start_loc[-1] = T
        cache_indices = mamba_indices.to(torch.int32)
        has_initial_state = forward_batch.extend_prefix_lens > 0
    else:
        # 单序列分支
        query_start_loc = hidden_states.new_tensor([0, T], dtype=torch.int32)
        cache_indices = mamba_indices[:1].to(torch.int32)
        has_initial_state = forward_batch.extend_prefix_lens[:1] > 0

    conv_out = causal_conv1d_fn(
        Bx_t,
        self.conv_weight,
        self.conv_bias,
        query_start_loc=query_start_loc,
        cache_indices=cache_indices,
        has_initial_state=has_initial_state, # 修复：传递正确状态
```

```
        conv_states=conv_state,
        activation=None,
    ).transpose(0, 1)
```

评论区精华

无 review 评论，PR 由维护者 hnyls2002 直接批准。

- 暂无高价值评论线程

风险与影响

- 风险：修改范围极小（仅两处文件，各 3 行），逻辑与现有 Mamba2Mixer 的实现模式一致，回归风险低。但无新增测试覆盖，若未来对 forward_batch.extend_prefix_lens 的语义有变动，可能产生未预期行为。
- 影响：直接影响 LFM2 和 LFM2-MoE 模型在高并发下的输出正确性与确定性，对使用 Mamba 缓存重用场景尤为重要。影响范围限于对应模型的短卷积前向路径，不影响其他模型或解码路径。
- 风险标记：缺少测试覆盖

关联脉络

- PR #23975 Fix LFM2 ShortConv Mamba State Indexing: 此前序修复了 Mamba 状态索引问题，但未处理 has_initial_state，导致状态泄漏残留