PR #43803 完整报告

vllm-project/vllm

[Perf] remove seqlen from Mamba SSD chunk kernels

合并时间：2026-05-28 23:40

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/43803

执行摘要

- 一句话：移除 Mamba SSD 内核死参数，TTFT 降低 17%
- 推荐动作：建议仔细审查以确保所有内核的 seqlen 参数均已移除，并考虑在类似内核中检查其他可能引发重编译的整型参数。

功能与动机

seqlen 是五个 Mamba 内核中的无效参数，Triton 会为每个不同的整数值触发重新编译，导致 TTFT 显著增加。PR body 中提及："Triton specializes plain int args by default, so every prefill with a new total-token-count was triggering a fresh JIT compile."

实现拆解

1. 识别死参数: 在 _chunk_cumsum_fwd_kernel、_chunk_state_fwd_kernel、_chunk_scan_fwd_kernel、_bmm_chunk_fwd_kernel 和 _causal_conv1d_fwd_kernel 中，seqlen 参数从未被读取或被立即覆盖。
2. 移除参数: 分别在内核函数签名和宿主函数调用处移除 seqlen 参数，涉及 4 个文件：
   - vllm/model_executor/layers/mamba/ops/ssd_chunk_state.py
   - vllm/model_executor/layers/mamba/ops/causal_conv1d.py
   - vllm/model_executor/layers/mamba/ops/ssd_bmm.py
   - vllm/model_executor/layers/mamba/ops/ssd_chunk_scan.py
3. 性能验证: 清空 Triton 缓存后，使用 ShareGPT 数据集进行冷启动基准测试，确认 TTFT 显著下降。

关键文件：

- vllm/model_executor/layers/mamba/ops/ssd_chunk_state.py（模块 内核；类别 source；类型 core-logic；符号 _chunk_cumsum_fwd_kernel, _chunk_state_fwd_kernel, _chunk_cumsum_fwd, _chunk_state_fwd）：移除了 _chunk_cumsum_fwd_kernel 和 _chunk_state_fwd_kernel 中的 seqlen 参数，以及对应的宿主函数调用。
- vllm/model_executor/layers/mamba/ops/causal_conv1d.py（模块 内核；类别 source；类型 core-logic；符号 _causal_conv1d_fwd_kernel）：移除了 _causal_conv1d_fwd_kernel 中的 seqlen 参数，该参数在函数体内被立即覆盖。

- vllm/model_executor/layers/mamba/ops/ssd_bmm.py（模块 内核；类别 source；类型 core-logic；符号 _bmm_chunk_fwd_kernel, _bmm_chunk_fwd）：移除了 _bmm_chunk_fwd_kernel 中的 seqlen 参数。
- vllm/model_executor/layers/mamba/ops/ssd_chunk_scan.py（模块 内核；类别 source；类型 core-logic；符号 _chunk_scan_fwd_kernel, _chunk_scan_fwd）：移除了 _chunk_scan_fwd_kernel 中的 seqlen 参数。

关键符号：_chunk_cumsum_fwd_kernel, _chunk_state_fwd_kernel, _chunk_scan_fwd_kernel, _bmm_chunk_fwd_kernel, _causal_conv1d_fwd_kernel, _chunk_cumsum_fwd, _chunk_state_fwd, _bmm_chunk_fwd, _chunk_scan_fwd

关键源码片段

vllm/model_executor/layers/mamba/ops/ssd_chunk_state.py

移除了 _chunk_cumsum_fwd_kernel 和 _chunk_state_fwd_kernel 中的 seqlen 参数，以及对应的宿主函数调用。

```
# Before: kernel had unused seqlen param
# @triton.jit
# def _chunk_cumsum_fwd_kernel(dt_out_ptr, dA_cumsum_ptr, cu_chunk_seqlens_ptr,
# seqlen, nheads: tl.constexpr, ...):

# After: seqlen removed
@triton.jit
def _chunk_cumsum_fwd_kernel(
    dt_out_ptr, dA_cumsum_ptr, cu_chunk_seqlens_ptr,
    nheads: tl.constexpr, # seqlen was never read; removed
    chunk_size: tl.constexpr,
    dt_min: tl.constexpr,
    ...
):
    # ... kernel body uses cu_chunk_seqlens_ptr for bounds, not seqlen
```

评论区精华

审核者 tomeras91 建议直接将 seqlen 参数全部移除，而非使用 do_not_specialize 装饰器，认为更简洁。作者采纳建议并重新测试，确认性能增益保持一致。

- 移除 seqlen 参数 vs. 使用 do_not_specialize (design): 全部移除 seqlen 参数，保留 do_not_specialize 方案仅作为后备。

风险与影响

- 风险：变更仅删除未使用的参数，不改变内核逻辑。风险极低，但需确认所有调用路径已更新，无遗漏。当前修改覆盖了内核定义和宿主函数调用，未涉及其他文件。
- 影响：对 Nemotron-H-8B 等使用 Mamba SSD 内核的模型，冷启动 TTFT 显著改善（平均 -17%，P99 -46%）。TPOT 和吞吐量无退化。非 Mamba 模型不受影响。

- 风险标记：无测试覆盖（代码变更安全）

关联脉络

- 暂无明显关联 PR