PR #24732 完整报告

sgl-project/sglang

[codex] Optimize LTX2 split rotary kernel

合并时间：2026-05-16 20:58

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/24732

执行摘要

- 一句话：优化 LTX2 分裂 RoPE Triton 内核，合并多个 head 的 launch grid
- 推荐动作：该 PR 值得精读，尤其对需要优化 Triton kernel 以利用 GPU 的开发者。核心设计决策是使用程序块合并多个 head，这是一种常见的 GPU 优化模式（减少 program 数量，增加每个 program 的工作量以更好地隐藏延迟）。自适应 warp 数量的选择也值得参考。建议在合并到主分支前，确认其他 GPU 架构（如 A100）的基准测试结果。

功能与动机

LTX-2 模型使用分裂式旋转位置编码（RoPE），原始的 Triton 内核为每个 token-head 对启动一个独立的 program，导致 GPU 上的 launch grid 碎片化严重，降低了执行效率。特别是当 head 数量较多（如 32 head）时，program 数量激增，影响吞吐量。该 PR 旨在通过合并多个 head 的计算到单个 program 中来减少 grid 大小，提升内核性能。

实现拆解

优化通过核心里面的向量化实现，将原来每个 program 处理一个 head 改为处理一块（最多 16 个）head，从而减少 GPU program launch 数量和 grid 大小。

1. 修改 Triton 内核 _ltx2_split_rotary_kernel 的 grid 和参数：
   - 在 apply_ltx2_split_rotary_emb 函数中，将原来的二维 grid (batch * seq_len, num_heads) 改为 (batch * seq_len, triton.cdiv(num_heads, block_heads))，其中 block_heads = min(16, triton.next_power_of_2(num_heads))。这样每个 program 处理 BLOCK_HEADS 个 head。
   - 新增编译时常量 BLOCK_HEADS 和相应的 num_warps 自适应逻辑（num_warps = min(8, max(1, block_heads))），替代原来固定的 num_warps=1。
2. 修改内核内部实现以支持多 head 的向量化加载 / 存储：
   - 将原来标量的 head 替换为向量 heads = head_block * BLOCK_HEADS + tl.arange(0, BLOCK_HEADS)。
   - 所有地址计算和掩码都扩展为二维：mask = (heads[:, None] < num_heads) & (offsets[None, :] < half_dim)，x_base 通过广播 heads[:, None] 得到 (BLOCK_HEADS, BLOCK_HALF) 形状的地址矩阵。
   - 相应地，tl.load 和 tl.store 的指针和掩码也扩展为二维（offsets[None, :] 以保持维数一致）。

3. 保留 BF16 舍入顺序以匹配 PyTorch 参考路径：
   - 内核中的算术顺序保持不变：先计算 (x_first * cos).to(tl.bfloat16).to(tl.float32)，然后与经过 FP32 累加的 -x_second * sin 相加，确保与 PyTorch 的 BF16 操作顺序一致。
4. 验证：
   - 在 H200 上对两个代表性形状进行了微基准测试，并验证了与 PyTorch BF16 参考路径的数值一致性（max_abs_diff = 0.0）。
   - 使用 2×H200 完整运行 LTX-2 模型推理，生成视频和性能 JSON 文件，确认无功能回归。

关键文件：

- python/sglang/jit_kernel/diffusion/triton/ltx2_rotary.py（模块 JIT 内核；类别 source；类型 core-logic；符号 _ltx2_split_rotary_kernel, apply_ltx2_split_rotary_emb）：唯一修改的文件，包含所有核心优化逻辑：grid 重划分、多 head 向量化加载 / 存储、自适应 num_warps。

关键符号：_ltx2_split_rotary_kernel, apply_ltx2_split_rotary_emb

评论区精华

该 PR 没有产生 review 讨论（0 条 review 评论），只有一个审核员 mickqian 批准。PR 作者 BBuf 在描述中提供了详细的基准测试数据和模型验证结果，未引起争议。

- 暂无高价值评论线程

风险与影响

- 风险：
  - 影响：该 PR 直接影响 LTX-2 扩散模型的分裂 RoPE 计算性能。在 H200 上，短序列加速 1.5x，长序列加速 1.94x，对扩散模型的端到端生成时间有可量化的提升（总耗时约 37.9 秒，优化后预计减少 1-3 秒）。由于仅修改一个内核函数，对其他模型没有影响。团队需确保其他 GPU 架构（如 A100）上的性能不会倒退。
  - 风险标记：缺少测试覆盖

关联脉络

- 暂无明显关联 PR