PR #21387 完整报告

sgl-project/sglang

[Diffusion] Optimize diffusion Triton rotary embedding by processing multiple heads per token

合并时间：2026-03-26 08:59

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/21387

执行摘要

本 PR 通过重构 Triton rotary embedding 内核，实现一次处理多个注意力头，显著提升扩散模型推理性能。基准测试显示微基准速度提升最高 3.1 倍，端到端提升 1.3%，风险在于 autotune 配置可能未覆盖所有场景，建议关注后续性能监控。

功能与动机

优化旨在解决扩散模型中 Triton rotary embedding 内核的效率瓶颈。PR body 引用“改进 cos/sin 行重用、减少冗余加载和降低启动开销”作为核心动机，使用 AKO4ALL 优化技能处理多个头以提高 GPU 资源利用。

实现拆解

主要改动在文件 python/sglang/jit_kernel/diffusion/triton/rotary.py 的 _rotary_embedding_kernel 函数：

- 启动布局变更：从一维改为二维，使用 bt_idx 和 head_block_idx 作为程序 ID，以同时处理多个头。
- 参数引入：新增 BLOCK_HEADS 参数，与 BLOCK_HS_HALF 一起在 autotune 中调优。
- 内存访问优化：调整指针计算（如 x_row_ptrs）和掩码逻辑（head_mask[:, None] & half_mask[None, :]），减少冗余加载。
- autotune 配置更新：从单一 BLOCK_HS_HALF 调优扩展为联合调优，但最大值从 256 降至 64。

评论区精华

review 中仅有 gemini-code-assist[bot] 的评论：

> "The new autotune configurations have a maximum BLOCK_HS_HALF of 64, while the previous version included values up to 256. For models with a large head_size, this could result in suboptimal performance..."

此评论指出配置覆盖不足的风险，但未在讨论中解决，PR 被直接批准。

风险与影响

- 技术风险：autotune 配置缩减可能对大 head_size 模型性能不利；内核重构可能引入边界错误（如掩码计算）；缺少回归测试覆盖。
- 影响评估：正面影响为性能提升（基准测试证实），但需在真实场景验证；对团队，此优化可加速扩散模型推理，但需警惕配置导致的性能回归。

关联脉络

与近期 PR 关联显示团队在扩散模型内核优化上的持续努力：

- PR 21318：优化 Qwen select01 Triton 调制内核，共享性能优化模式。
- PR 21091：添加扩散模型性能比较 CI job，可用于追踪此优化的长期效果。
- PR 21323：添加 AKO4ALL 优化技能文档，与本 PR 使用的技能相呼应。 这些关联表明 SGLang 项目在扩散模块的性能调优和自动化测试方面有系统演进。