PR #21019 完整报告

sgl-project/sglang

[Qwen3.5] Fuse split/reshape/cat ops in GDN projection with Triton kernel

合并时间：2026-03-23 23:17

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/21019

执行摘要

本 PR 优化了 Qwen3.5 模型的 Gated Delta Net 投影层，通过引入 Triton 核融合 split/reshape/cat 操作，减少内核启动和内存分配，提升推理性能。尽管在小模型上存在性能讨论，但整体改进显著，需注意 FP8 量化兼容性和准确性验证。

功能与动机

作为 PR #19321 的后续，针对 Qwen3.5 检查点布局与 Qwen3-Next 不同的特点，将多个投影操作融合以优化性能。PR body 中明确指出：'In PR https://github.com/sgl-project/sglang/pull/19321 we fused Qwen3-Next GDN's qkvz_proj and ba_proj. This PR is a follow up.' 目标是减少内核启动和中间张量分配，提升预填充和解码阶段的效率。

实现拆解

- 新增 Triton 核：在 python/sglang/jit_kernel/triton/gdn_fused_proj.py 中新增 fused_qkvzba_split_reshape_cat_contiguous 核函数，处理连续输入格式。
- 重构投影层：修改 python/sglang/srt/models/qwen3_5.py，将原先的 in_proj_qkv、in_proj_z、in_proj_b、in_proj_a 合并为 in_proj_qkvz 和 in_proj_ba 两个融合层。
- 增强权重加载器：实现 _make_packed_weight_loader 方法，支持融合和拆分检查点格式的权重加载，确保参数初始化正确。
- 清理冗余代码：修改 python/sglang/srt/models/qwen3_next.py，移除旧 Triton 核，复用新核以保持代码一致性。

评论区精华

- 性能争议：jasperjiaguo 报告小模型性能下降，但作者 yuan-luo 验证后显示改进，引发对优化效果的讨论。引用 yuan-luo：'I'll verify the small model's performance and do refactor to avoid the corresponding impact.'
- FP8 修复：yuan-luo 在 issue 评论中修复 FP8 量化兼容性问题：'FP8 problem fixed.' 强调权重加载器需处理不同参数类型。
- 准确性检查：cs-cat 提到可能影响准确性：'This PR does indeed bring significant performance improvements, but it seems to affect the accuracy of the model? Please refer to #21696.' 需要进一步验证。

风险与影响

- 技术风险：小模型性能可能出现波动；FP8 量化模型需特殊处理权重加载器；变更可能引入准确性偏差，需加强测试。
- 影响范围：对用户而言，大模型推理性能提升，但小模型需监控；系统层面减少内核启动和内存开销；团队需适应新的融合架构，并关注兼容性测试。

关联脉络

- 与 PR #19321 直接关联，延续了投影层融合的优化路线，展示了跨模型版本的一致性优化策略。
- 涉及 jit-kernel 和量化主题，与仓库中其他性能优化 PR（如 #21657 的 AMD 优化）共享技术思路，反映了团队在核融合和量化适配上的持续演进。