PR #23682 完整报告

sgl-project/sglang

Add fused moe triton config for Qwen3.5-397B-A17B-FP8

合并时间：2026-04-25 09:35

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/23682

执行摘要

为 Qwen3.5-397B-A17B-FP8 模型在 H100 GPU 上新增 Fused MoE Triton 内核调优配置，通过自动化脚本生成最优参数组合，在解码阶段实现最高约 11% 的吞吐性能提升。

功能与动机

Qwen3.5-397B-A17B-FP8 在 MoE 内核上未经过针对性调优，导致 H100 上的性能受限。本 PR 通过运行 SGLang 官方调优脚本 生成专用于该模型和硬件的配置，释放后端内核潜力。

实现拆解

1. 明确硬件与模型参数：模型 Qwen3.5-397B-A17B-FP8，TP 8，H100 GPU，FP8 w8a8。
2. 执行内核调优：使用 tuning_fused_moe_triton.py 脚本进行自动化参数搜索。
3. 收集最优配置：得到从 batch size 1 到 4096 共 20 个档位的参数组合。
4. 生成 JSON 配置文件：按 SGLang 命名规约存入 configs/ 目录，运行时自动加载。
5. 性能验证：调优后 batch size 16 时 decode 吞吐从 1615.32 tok/s 提升至 1792.89 tok/s。

配置示例（batch size 16）：{ // 批次大小 16 时的最优调优参数 "16":{ "BLOCK_SIZE_M":16, "BLOCK_SIZE_N":64, "BLOCK_SIZE_K":128, "GROUP_SIZE_M":64, "num_warps":4, "num_stages":3 } }

python/sglang/srt/layers/moe/moe_runner/triton_utils/configs/triton_3_5_1/E=512,N=128,device_name=NVIDIA_H100_80GB_HBM3,dtype=fp8_w8a8,block_shape=[128, 128].json

新增 Qwen3.5-397B-A17B-FP8 在 H100 上的 Fused MoE Triton 调优参数配置，直接提升解码吞吐量。

```
{
  // 批次大小 16 时的最优调优参数
  "16": {
    "BLOCK_SIZE_M": 16, // M 维度 tile 大小，较小值适配小批次
    "BLOCK_SIZE_N": 64, // N 维度 tile 大小，平衡计算与访存
    "BLOCK_SIZE_K": 128, // K 维度 tile 大小，固定为 128 匹配 FP8 block shape
    "GROUP_SIZE_M": 64, // M 方向分组大小，影响并行度
    "num_warps": 4, // 每个 thread block 包含 4 个 warp
    "num_stages": 3 // 软件流水线阶段数，控制寄存器占用
```

```
  }
}
```

评论区精华

无 review 评论，作者在 Issue 评论中提到该配置是在调查 Issue #23500 时发现的附带优化。

风险与影响

- 风险：极低。配置文件仅对特定模型和硬件生效，不影响其他场景。未来 kernel 更新可能导致配置过期，但易于替换。
- 影响：直接提升 Qwen3.5-397B 用户的推理吞吐，高并发下效果显著；对系统无侵入，维护成本低。

关联脉络

该 PR 本身独立，但作为 SGLang MoE 调优框架的实践，为后续其他模型的调优配置添加提供了参考流程。