PR #25775 完整报告

sgl-project/sglang

[Perf][Qwen3.5] Add case 512 to topkGatingSoftmaxKernelLauncher,

合并时间：2026-05-25 16:08

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/25775

执行摘要

为 MoE top-k softmax CUDA 内核添加对 512 个专家的直接支持，使 Qwen3.5-397B-A17B 等大专家模型能走融合单内核路径，延迟降低约 1.79x-4.25x，E2E 吞吐在高并发提升 7.24%。改动简洁，风险低。

功能与动机

Qwen3.5-397B-A17B 等模型使用了 512 个专家，但此前 topkGatingSoftmaxKernelLauncher 的 switch-case 只支持到 256 专家，超过 256 会回退到需要额外工作区（workspace）的通用路径，导致性能损失。PR 旨在通过添加 case 512 分支并调整阈值，使 512 专家也能享受融合单内核路径的性能优势。

实现拆解

1. 添加内核特化分支：在 sgl-kernel/csrc/moe/moe_topk_softmax_kernels.cu 的 topkGatingSoftmaxKernelLauncher 函数中新增 case 512: LAUNCH_SOFTMAX(T, 512, WARPS_PER_TB); break;。该宏会展开为模板特化的 top-k softmax 内核调用，避免通用路径的内存分配和数据复制。
2. 扩展工作区阈值：将 needs_workspace 判断条件从 num_experts > 256 改为 num_experts > 512。512 专家现在被认定为“小专家数”，无需预分配 flattened softmax 临场张量。
3. 配套 benchmark 与测试更新：benchmark 脚本 (sgl-kernel/benchmark/bench_moe_topk_softmax.py) 的 topk_range 增加 10，对应 Qwen3.5 实际 topk；单元测试 (sgl-kernel/tests/test_moe_topk_softmax.py) 参数化列表也加入 topk=10，保证正确性验证。

sgl-kernel/csrc/moe/moe_topk_softmax_kernels.cu

核心 CUDA 内核的 switch-case 分支和工作区阈值变更。

```
// 在 topkGatingSoftmaxKernelLauncher 的 switch 中新增 case 512
case 512:
  LAUNCH_SOFTMAX(T, 512, WARPS_PER_TB);
  break;
// 同时将 needs_workspace 阈值从 256 提升至 512，
// 确保 512 专家走融合内核路径而非 fallback 路径
const bool needs_workspace = !is_pow_2 || num_experts > 512;
```

评论区精华

无实质性 review 讨论。

风险与影响

- 回归风险：极低。仅影响 512 专家且 topk<=10 的场景，其他路径完全不变。
- 性能影响：512 专家场景显著提速；其他规模无变化。高并发下 E2E 提升约 7%。
- 兼容性：无 Breaking Change。

关联脉络

无关联其他 PR。此 PR 是内核微调，服务于特定模型（Qwen3.5）的推理加速。