PR #20089 完整报告

sgl-project/sglang

feat: [1/2] [DeepEP] Fuse shared expert into MoE dispatch under EP

合并时间：2026-04-09 16:48

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/20089

执行摘要

此 PR 将 DeepSeek V3/R1 模型中的共享专家融合到 DeepEP 的 MoE 分发路径，作为每个 rank 的本地附加专家处理，消除了独立计算的开销，为后续水印负载平衡功能铺路，影响专家并行下的推理效率。

功能与动机

当前在 DeepSeek V3/R1 使用专家并行（EP）时，每个 rank 独立计算共享专家，而不是作为 DeepEP 分发管道的一部分，导致重复计算和效率低下。此变更旨在将共享专家融合到 MoE 路径中，为后续的水印负载平衡功能（请求自 #19290）做准备。

实现拆解

- 新模块 deepep_shared_expert_fusion.py: 提供 expand_topk_with_shared_expert 函数，用于在 TopK 选择后扩展共享专家列。
- 模型层 deepseek_v2.py: 调整 num_experts 和 top_k 以容纳扩展布局（例如，EP=16 时从 256 路由专家增加到 272 总专家），并跳过独立的共享专家前向传播。
- MoE 层 fused_moe_triton/layer.py: 修改专家 ID 重映射逻辑，处理 DeepEP 下的共享专家槽位。
- TopK 计算 topk.py: 更新 biased_grouped_topk_gpu 函数，支持共享专家的单独追加。
- 服务器参数 server_args.py: 添加 --enable-deepep-waterfill CLI 标志来控制功能启用。

评论区精华

- 潜在 bug: gemini-code-assist[bot] 指出 get_moe_weights 中 num_local 计算可能不正确，影响 EPLB 功能。

"The calculation of num_local seems incorrect when _enable_deepep_waterfill is true."

- 设计讨论: ch-wan 质疑 biased_grouped_topk_gpu 修改的必要性，询问原始实现是否支持融合。

"Could you confirm if the original implementation of biased_grouped_topk_gpu supports shared expert fusion?"

• 兼容性: ch-wan 强调在 DeepEP 结合 TBO 或 SBO 时应禁用融合。

"We should disable fusion for deepep + TBO / SBO."

风险与影响

风险:

1. get_moe_weights 中的计算错误可能导致专家并行负载平衡功能中断。
2. 修改 topk 逻辑可能破坏标准 TP 模式下的兼容性。
3. 新融合路径增加代码复杂性，需确保充分测试以避免回归。

影响:

• 用户需通过 CLI 标志启用功能，默认行为不变。
• 系统优化了 DeepEP 下的资源利用，减少重复计算。
• 团队为后续负载平衡功能奠定基础，但维护成本增加。

关联脉络

此 PR 是水印负载平衡功能的第一部分，关联 #19290。历史 PR 中，PR #21822 涉及 MoE bugfix，PR #22389 涉及调度重构，均与此 PR 的 MoE 和调度修改相关，反映了团队在优化专家并行推理方面的持续努力。