PR #6029 完整报告

verl-project/verl

[fully_async] fix: replace routed_experts on partial rollout resume i···

合并时间：2026-04-17 10:42

原文链接：http://prhub.com.cn/verl-project/verl/pull/6029

执行摘要

- 一句话：修复完全异步策略中部分 rollout 恢复时 MoE 路由专家重复拼接导致的训练不稳定问题。
- 推荐动作：该 PR 值得精读，它揭示了在异步训练中处理路由专家数据时的关键设计决策：直接替换而非拼接，以确保路由与权重版本一致。关注作者与 reviewer 关于路由切片方案的讨论，以及实验结果的权衡。

功能与动机

根据 PR body 描述，sglang 引擎在每次调用时返回的 routed_experts 覆盖整个序列（提示词 + 所有生成 token），而部分 rollout 恢复时输入为 提示词+已生成token，sglang 会重新处理整个输入并返回完整路由数据。原代码使用 torch.cat 拼接新旧数据，导致路由专家重复（如 提示词+A B C + 提示词+A B C + D E），进而破坏 MoE 专家重放，引发 actor/ppo_kl 指标异常。相关 Issue 包括 #4348（部分 rollout RFC）、#4101（R3 路由器重放）和 #5344（完全异步中的 R3）。

实现拆解

1. 移除冗余导入：在 verl/experimental/fully_async_policy/agent_loop/agent_loop.py 中删除 import torch，因为路由专家处理不再需要拼接操作。
2. 修改路由专家合并逻辑：将 FullyAsyncLLMServerManager.generate() 方法中处理 routed_experts 的代码从条件拼接改为直接赋值。原逻辑为：如果 output.routed_experts 不为 None，则检查 final_output.routed_experts 是否为 None，若是则赋值，否则使用 torch.cat 沿第 0 维拼接。新逻辑为：如果 output.routed_experts 不为 None，则直接 final_output.routed_experts = output.routed_experts。
3. 添加注释说明：在修改处添加注释，解释 sglang 返回完整序列路由专家数据，因此在部分 rollout 恢复时新输出已覆盖所有位置，无需拼接。
4. 无测试配套改动：由于这是分布式异步训练中的 bug，涉及多节点 sglang+Megatron 设置和 MoE 模型，难以在 CI 中复现，因此未添加单元或端到端测试。

关键文件：

- verl/experimental/fully_async_policy/agent_loop/agent_loop.py（模块 异步策略；类别 source；类型 core-logic；符号 FullyAsyncLLMServerManager.generate）：唯一变更文件，包含完全异步策略中 LLM 服务器管理器的核心生成逻辑，修复了路由专家处理 bug。

关键符号：FullyAsyncLLMServerManager.generate

关键源码片段

verl/experimental/fully_async_policy/agent_loop/agent_loop.py

唯一变更文件，包含完全异步策略中 LLM 服务器管理器的核心生成逻辑，修复了路由专家处理 bug。

```
async def generate(
    self,
    request_id: str,
    prompt_ids: List[int],
    sampling_params: Dict[str, Any],
    image_data: Optional[List[ImageType]] = None,
    video_data: Optional[List[VideoType]] = None,
) -> TokenOutput:
    """
    生成token的异步方法，支持部分rollout恢复。
    """
    # ... 初始化final_output等代码 ...
    while True:
        # 1. 生成tokens
        output = await super().generate(
            request_id=request_id,
            prompt_ids=prompt_ids + final_output.token_ids, # 部分rollout恢复时，输入为提示词+
            已生成token
            sampling_params=sampling_params,
            image_data=image_data,
            video_data=video_data,
        )

        # 2. 将output合并到final_output
        final_output.token_ids.extend(output.token_ids)
        if output.log_probs is not None:
            final_output.log_probs.extend(output.log_probs)
        # sglang返回routed_experts用于完整序列（提示词+所有token），
        # 因此在部分rollout恢复时，新输出已覆盖所有位置。
        if output.routed_experts is not None:
            final_output.routed_experts = output.routed_experts # 直接替换，而非拼接
        if output.num_preempted is not None:
            final_output.num_preempted += output.num_preempted
        final_output.stop_reason = output.stop_reason
        # ... 后续更新global_steps和检查停止条件的代码 ...
```

评论区精华

reviewer wuxibin89 提出替代方案：建议提取与 output.token_ids 对应的路由专家切片进行拼接，以反映不同模型版本。作者 NoonePauseferg 回应：直接替换更优，因为 actor 更新始

终使用最新权重（v2），如果使用 v1 的路由与 v2 的权重重放，计算出的 logprob 既不是 pi_v1 也不是 pi_v2，而是不一致的混合；直接替换可确保 v2 路由 +v2 权重 = 干净的 pi_v2，使重要性比率定义明确。作者进一步通过实验比较，显示记住路由的版本训练结果更稳定，并创建了 PR #6046 进行后续探索。最终结论：先合并本 PR 作为基线，后续再实验每迭代切片方案。

- 路由专家处理策略：替换 vs 切片拼接 (design): 先合并直接替换方案作为基线，后续通过 PR #6046 实验切片方案。

风险与影响

- 风险：1. 回归风险：修改后路由专家数据不再累积历史，如果后续逻辑依赖完整历史路由（如某些调试或分析工具），可能引发问题。但根据讨论，这符合当前训练设计。 2. 性能风险：无显著性能影响，仅改变数据赋值方式。 3. 兼容性风险：与 sglang v0.5.9 及更高版本行为兼容，但若 sglang 未来更改 routed_experts 返回逻辑（如只返回新增 token 的路由），则需重新适配。 4. 测试覆盖不足：由于分布式环境复杂性，缺少自动化测试，可能隐藏边缘情况。
- 影响：1. 对用户影响：修复了使用完全异步策略、部分 rollout 和 MoE 模型时可能出现的训练不稳定问题（actor/ppo_kl 异常），提升训练可靠性。 2. 对系统影响：仅影响 FullyAsyncLLMServerManager.generate() 方法中的路由专家处理逻辑，不改变其他模块。 3. 对团队影响：提供了更清晰的路由专家处理模式，为后续优化（如 PR #6046）奠定基础。
- 风险标记：核心路径变更，缺少测试覆盖

关联脉络

- PR #6046 [fully_async] experiment: per-iteration routed_experts slicing for partial rollout resume: 由本 PR 讨论引发，进一步探索路由专家切片方案，以比较训练稳定性。
- PR #5997 [trainer,algo] feat: Support On-Policy Distillation in main_ppo_sync: 涉及在线策略蒸馏和路由专家重放（R3），与本 PR 的 MoE 专家重放相关。
- PR #5989 [megatron] fix: add missing FP8 padding for router replay: 修复 Megatron 路由器重放路径问题，与本 PR 的路由专家处理同属路由重放领域。