

PR #7722 完整报告

verl-project/verl

[perf, tool] fix: Avoid overlapping rollout and training profilers in Trainer V1

合并时间: 2026-09-08 16:13

原文链接: <http://prhub.com.cn/verl-project/verl/pull/7722>

执行摘要

- 一句话: 分离 rollout 和训练 profiler 窗口, 修复 NPU/GPU 并发采集冲突导致数据丢失。
- 推荐动作: 本 PR 是一个高质量、有针对性的 bugfix, 解决了实际硬件平台上的关键问题。它不仅修复了核心逻辑, 还显著增强了测试和验证基础设施。强烈建议精读其实现, 尤其是三种 trainer 模式下 profiler 生命周期控制的新设计, 这对理解 V1 架构的扩展性和健壮性有重要参考价值。实现逻辑清晰, 测试充分。

功能与动机

在 colocated 工作负载中, rollout 和训练 profiler 可能重叠, 导致 NPU profiler 数据丢失和 GPU 上的 Nsight Systems 指标冲突。PR body 指出: “Previously, rollout profiling could remain active when training profiling started, causing multiple processes on the same device to collect profiling data simultaneously”。这影响了性能分析的准确性。

实现拆解

1. 基类抽象 (trainer_base.py): 在 PPOTrainer 基类中新增 _start_rollout_profiling() 和 _stop_rollout_profiling() 方法, 封装对 _rollout_server_managers() 的调用。原有 _start_profiling() 现在调用 _start_rollout_profiling(), 而 _stop_profiling() 中移除了 rollout 停止逻辑, 将控制权交给子类。
2. Sync/ColocateAsync 模式调整 (trainer_sync.py, trainer_colocate_async.py): 在 on_sample_end() 方法中, 于 sleep_replicas() 之后增加条件调用 self._stop_rollout_profiling()。这确保了在同步和协同异步模式下, rollout profiler 在采样结束后、训练开始前立即停止。
3. SeparateAsync 模式调整 (trainer_separate_async.py): 在 on_step_end() 方法开头, 根据 prev_step_profile 标志调用 self._stop_rollout_profiling()。因为该模式下 rollout 和训练可能在不同资源上异步进行, 需要在整个步骤结束时统一停止 rollout profiler。
4. 测试覆盖增强 (test_rollout_profiling_on_cpu.py): 重写并大幅扩展了单元测试, 新增针对 PPOTrainerSync、PPOTrainerColocateAsync 和 PPOTrainerSeparateAsync 的测试用例 (如 test_sync_stops_rollout_after_sampling), 验证新的 profiler 生命周期顺序 (如 rollout stop 在第一个训练 stage 之前)。引入了辅助函数 _prepare_separate_step_end 和更精细的调用顺序断言 (call mock)。

5. Profiler 输出验证工具升级 (test_check_profiler_output.py, CI 脚本): 重构了 ProfilerChecker 工具, 使其支持更灵活的 stage 模式匹配和 GPU JSON/NPU 目录验证, 并新增 --stage 参数允许 CI 脚本指定待验证的阶段。相应更新了 run_ppo_trainer_veomni.sh 和 .github/workflows/e2e_ascend.yml 中的调用命令, 以验证分离后的 profiler 输出文件。

关键文件:

- verl/trainer/ppo/v1/trainer_base.py (模块 训练器基类; 类别 source; 类型 core-logic; 符号 _start_rollout_profiling, _stop_rollout_profiling): 核心抽象层, 引入了 _start_rollout_profiling 和 _stop_rollout_profiling 方法, 是 profiler 生命周期控制权分离的关键。
- verl/trainer/ppo/v1/trainer_sync.py (模块 同步训练器; 类别 source; 类型 core-logic): 同步训练模式实现, 在 on_sample_end 中添加了 rollout profiler 停止调用, 是解决时序冲突的核心改动点之一。
- tests/trainer/ppo/v1/test_rollout_profiling_on_cpu.py (模块 训练器测试; 类别 test; 类型 test-coverage; 符号 test_sync_stops_rollout_after_sampling, test_colocate_async_stops_rollout_only_after_replicas_sleep, test_separate_async_closes_hybrid_and_standalone_rollout_once, test_rollout_stops_before_the_first_training_stage): 全面重写的测试文件, 是验证新 profiler 生命周期正确性的关键, 覆盖了三种训练模式的核心场景。

关键符号: _start_rollout_profiling, _stop_rollout_profiling, _rollout_server_managers, on_sample_end, on_step_end

评论区精华

1. 核心讨论 (Issue 评论 by tardis-key): 提出了两个关键考量点: a) 新流程必须覆盖所有场景, 确保 GPU 行为不受影响; b) NPU CI 需要能检测“单卡多进程”导致的数据丢失问题。作者在回复中确认, 已在 GPU/NPU 的 sync、colocate_async、separate_async 模式下通过端到端测试, 并增强了 profiler 输出文件的验证。
 2. 决策结论: PR 通过作者的解释和验证 (包括提交 trace 截图) 消除了 reviewer 的疑虑。核心设计——将 rollout profiler 停止点从步骤结束提前到采样结束 (或在 separate_async 中在步骤结束统一停止) ——被接受。
 3. 未解决疑虑: 无。讨论已闭环。
- 变更覆盖范围与 GPU 行为保持 (correctness): 作者确认已在 GPU/NPU 平台的 sync、colocate_async、separate_async 三种模式下通过端到端 profiling 测试, 并提供了 trace 截图作为证据。
 - NPU CI 对“单卡多进程” profiler 问题的检测能力 (testing): 作者增强了 GPU 和 NPU 的 profiling 测试, 现在会验证生成的 profiling 产物 (如 JSON 文件、目录结构), 以检测数据缺失。

风险与影响

- 风险:

1. 回归风险: 变更涉及 Trainer V1 的三种核心运行模式 (sync、colocate_async、separate_async) 的 profiler 控制流。虽然测试覆盖全面, 但 profiler 逻辑与主训练循环紧密耦合, 任何遗漏都可能影响 profiling 数据的完整性或导致意外的行为变化 (如 profiler 窗口过早关闭)。
2. 兼容性风险: 新的 profiler 停止时机改变了输出文件的生成模式。CI 脚本和验证工具必须同步更新 (已在本 PR 中完成), 否则会导致 CI 失败。使用旧版验证工具的用户可能无法正确检查新格式的输出生。
3. 性能风险: 无。主要是控制流调整, 无性能敏感路径变更。

- 影响:

1. 对用户: 直接影响使用性能分析 (profiling) 功能的用户。修复后, 生成的性能 trace 文件将更准确、完整, 不再包含重叠的空闲或冲突数据, 提升了性能分析的可靠性。
2. 对系统: 提高了 Trainer V1 框架的健壮性, 特别是针对 NPU 硬件的兼容性和 GPU 上与第三方工具 (Nsight Systems) 的集成。增强了自动化测试基础设施对 profiler 输出的验证能力。
3. 对团队: 为性能分析功能建立了一个更清晰、更可维护的控制模型 (通过基类方法抽象)。增加了团队在 CI 中捕获 profiler 问题的能力, 减少了未来调试相关问题的成本。 - 风险标记: 核心路径变更, 需要跨模式验证

关联脉络

- 暂无明显关联 PR