

PR #2367 完整报告

radixark/miles

Add a PPO example for the shared actor/critic setup

合并时间: 2026-08-13 06:20

原文链接: <http://prhub.com.cn/radixark/miles/pull/2367>

执行摘要

- 一句话: 新增单机 Qwen3-4B PPO 可运行示例与文档
- 推荐动作: 值得精读。重点看 `examples/ppo/README.md` 的 `Constraints` 和 `Which numbers here are verified` 两节, 以及 `run_qwen3_4b_ppo.py` 中注释串联的 PPO 参数设计: critic 共置、loss 级 KL、每 GPU 一个 engine 等决策都有明确理由。想在单机上跑 PPO 的团队可以直接以此为起点; 后续接入官方文档时也建议保留这些约束说明。

功能与动机

PR body 指出, miles 早已通过 `--advantage-estimator ppo` 支持 PPO, 但仓库里没有可运行的示例: 唯一可用的 flag 组合藏在 `tests/e2e/megatron/test_qwen3_4B_ppo.py` 这个三层冒烟测试里, 用户不会去那里找训练配方。同时, 该 e2e 测试此前因 PPO placement group has conflict on port 被禁用, 最近在 actor 与 critic 共享 placement group 后才重新启用, 单机 PPO 真正可用且被 CI 覆盖——因此现在是把 PPO 配方文档化的合适时机。

实现拆解

1. 新增启动脚本 `examples/ppo/run_qwen3_4b_ppo.py`: 采用与仓库 Python 化启动器一致的结构 (`ScriptArgs` 继承 `U.ExecuteTrainConfig` + `prepare()` + `execute()` + `@U.dataclass_cli main()`)。 `prepare()` 用 `U.exec_command_cpu` 下载 Qwen3-4B checkpoint 与 zhuzilin/dapo-math-17k 数据集, 并调用 `U.convert_checkpoint` 转成 Megatron torch_dist 格式; 重复运行会跳过转换。 `execute()` 将参数分为 `ckpt_args`、`rollout_args`、`perf_args`、`ppo_args`、`optimizer_args`、`sglang_args`、`misc_args` 组装后交给 `U.execute_train` 执行, 由 `execute_train` 统一负责 ray 启停与 NVLink 检测。
2. PPO 关键参数与约束: `--advantage-estimator ppo` 是唯一触发 critic 构建与 GAE 优势计算的开关; `--critic-lr 1e-5` 与 `--num-critic-only-steps 1` 做值函数预热; 奖励级 KL 被禁用 (`--kl-coef 0.00`), 改用 `--use-kl-loss`, 原因写进了 README: critic 先于 actor 训练且看不到 ref log probs; `--colocate` 体现 critic 与 actor 共置并继承其并行度, 因此 `--offload-train` 被强制开启。
3. 文档配套: `examples/ppo/README.md` 给出 Quick Start、critic flags 表格 (含各 flag 的默认回退)、约束清单、以及“哪些数字是 CI 验证的、哪些不是”的说明; `examples/README.md` 增加一行 ppo 条目索引。
4. 测试与验证: PR 没有新增测试文件, 但作者在 8xH200 单机上原样运行了启动脚本, 连续完成 20 个 rollout 步骤 (0~19) 零报错, 验证了 critic 预热门控、共置放置、双

checkpoint 落盘、value loss 收敛等声明。

关键文件：

- `examples/ppo/run_qwen3_4b_ppo.py` (模块 启动脚本；类别 `source`；类型 `core-logic`；符号 `ScriptArgs`, `prepare`, `execute`, `main`)：示例核心启动脚本：完整给出 PPO 训练所需的全部参数组合，是本次变更的主体；其参数分组与注释直接体现了 critic 共置、KL 禁用、每 GPU 一个 engine 等设计决策。
- `examples/ppo/README.md` (模块 示例文档；类别 `docs`；类型 `documentation`)：示例唯一配套文档，系统解释 PPO 与 GRPO 的差异、critic flags 回退逻辑、参数校验约束及 CI 验证边界，是理解示例设计意图的关键。
- `examples/README.md` (模块 示例索引；类别 `docs`；类型 `documentation`)：在 `examples` 索引中补充 ppo 条目，让示例从导航上可见。

关键符号：ScriptArgs, prepare, execute, main

评论区精华

guapisolo 的 review 推动了启动脚本从 bash (`run-qwen3-4b-ppo.sh`) 改写成 Python，作者在第三提交中完全采纳，与仓库 #2356 的 Python 化启动器形态保持一致；同一次 review 中，guapisolo 指出 4B 模型用每 GPU 一个 engine 更快，最终采用 `--rollout-num-gpus-per-engine 1`。对于 README 放在 `examples` 是否意味着未完全验证的疑问，Shi-Dong 明确回复示例已完全验证（20 步 rollout 零报错），并将在后续单独 PR 中同步进 Miles 官方文档。此外，guapisolo 补充确认了 critic 共置并共享并行这一关键约束，README 据此将约束集中到 Constraints 一节。

- 将 bash 启动脚本改为 Python (design): 采纳，第三提交将 launcher 转为 Python，匹配现有 `dataclass + typer + prepare/execute` 形态，`execute_train` 接管 ray 启停与 NVLink 检测。
- 每 GPU 一个 rollout engine (performance): 采纳，`--rollout-num-gpus-per-engine 1`，README 说明与 CI 的差异。
- 示例是否因未完全验证才放在 `examples` (question): Shi-Dong 回复 'The example is fully verified. We will add this example to the Miles docs in a separate PR.'
- critic 共置与并行继承约束 (design): README 将共置约束与 offload 强制开启写进 Constraints 节。
- README 结构调整 (docs): 最终 README 把约束集中到 Constraints 一节；具体是否完全按建议调整无进一步证据。

风险与影响

- 风险：
 - 配置漂移风险：示例中的 `--eps-clip 0.2`、`--num-rollout 300`、`--rollout-num-gpus-per-engine 1` 与 CI 冒烟测试（`4e-4`、3 步、TP=2）不同，这些值是作者意图用于真实训练的，但只经过单次 20 步验证，长期收敛质量尚未观察。

- 文档时效性风险：README 对 `--critic-*` 回退规则、`--kl-coef` 禁用、`critic` 共置继承的说明依赖 `miles/utils/arguments.py` 与训练框架的当前实现；未来参数校验或并行逻辑变化时，该文档可能过时且无测试强制同步。
- 运行环境风险：启动脚本硬编码 `/root/datasets`、`/root/models`、`/root/Megatron-LM` 等路径，并依赖 HuggingFace 下载；在没有网络或非标准目录的集群上不能直接运行。
- 回归风险低：本次只新增示例与文档，未改动任何核心训练、rollout 或 dashboard 代码。
- 影响：
 - 用户侧：此前只有 `test_qwen3_4B_ppo.py` 一处可用的 PPO 配置，现在用户可以直接 `python examples/ppo/run_qwen3_4b_ppo.py` 跑起单机 PPO，明显降低上手门槛。
 - 知识侧：README 是首个系统说明 `miles` 中 `critic` 共置 / 继承并行、`--offload-train` 强制开启、奖励级 KL 被禁等设计约束的文档，对 debug 和二次开发都有价值。
 - 团队侧：为后续把 PPO 示例接入正式 docs（如 PR body 提到的单独文档 PR）铺路；也延续了 #2356 启动脚本 Python 化的方向。
 - 系统侧：无核心代码影响，训练、rollout、dashboard 均不受影响。
 - 风险标记：示例参数与 CI 配置不同，真机验证仅 20 步，依赖外部模型 / 数据集下载，文档依赖参数校验实现

关联脉络

- PR #2356 Replace all the .sh launch scripts with .py launch script: 本 PR 的启动脚本最终从 bash 改写成 Python，采用与 #2356 相同的 `dataclass + typer + prepare/execute` 结构，两者共同推动启动脚本形态的统一。