

PR #2203 完整报告

radixark/miles

Revert "Reject --disable-weights-backuper for LoRA + colocate + offload-train" (#2077)

合并时间: 2026-08-05 08:50

原文链接: <http://prhub.com.cn/radixark/miles/pull/2203>

执行摘要

- 一句话: 回滚过宽的 backuper 参数断言, 修复 LoRA e2e CI 启动崩溃
- 推荐动作: 值得快速阅读。虽然改动只有 8 行删除, 但 PR body 展现了良好的工程判断: 识别断言覆盖范围与实际行为 (--colocate 隐式开启 offload_train) 的偏差, 并坚持把防御性检查放在真实故障点。建议关注后续 #2204 是否已合入, 确认风险窗口已关闭。

功能与动机

PR body 明确指出: #2077 引入的断言过于宽泛, 破坏了 main 上所有 LoRA e2e CI 套件。原因有三:

1) parse_args 中 --colocate 会强制 offload_train=True; 2) 5 个 LoRA e2e CI 脚本 (如 tests/e2e/megatron/model_scripts/test_glm5_1_744b_a40b_6layer_lora_ci.py、test_inkling_small_4layer_lora_ci.py 等) 故意组合传递 --disable-weights-backuper 与 --colocate; 3) 这些套件在参数解析时死于 AssertionError。CI 首个失败见 actions run 30958354996, 全部 stage-c GPU 任务在 #2077 合入后立即报错。

实现拆解

1. 定位冲突点: 在 miles/utils/arguments.py 的 miles_validate_args() 中, 找到 #2077 插入的 8 行断言, 其位于 offload_train 分支强制设置 disable_param_buffers_cpu_backup 之后、offload_train_target == disk 断言之前。
2. 删除断言及解释性注释: 该断言拒绝 not args.enable_weights_backuper and args.offload_train and args.colocate and is_lora_enabled(args) 的组合, 本 PR 将这一整块 (含 3 行注释) 全部移除, 恢复 #2077 之前的状态。
3. 保留真正必要的断言: --offload-train-target=disk 分支中的既有断言 (要求 enable_weights_backuper) 未被触碰, 因为该组合在 disk offload 场景下确实无法恢复, 必须继续拒绝。
4. 配套修复安排: PR body 声明后续 PR #2204 会在 miles/backends/megatron_utils/update_weight/common.py 的 _maybe_get_cpu_backup() 中直接抛错, 替代这条过宽的启动期防御, 把故障暴露在真实故障点。
5. 测试配套: 本 PR 不新增测试; 其验证方式依赖 LoRA e2e CI 套件恢复通过, 属于回滚类变更的常规做法。

关键文件:

- `miles/utils/arguments.py` (模块 参数解析; 类别 `source`; 类型 `core-logic`; 符号 `miles_validate_args`): 唯一变更文件, 删除 #2077 引入的 8 行启动断言, 恢复 LoRA e2e CI 可用性, 并保留 `--offload-train-target=disk` 的必要断言。

关键符号: `miles_validate_args`

关键源码片段

`miles/utils/arguments.py`

唯一变更文件, 删除 #2077 引入的 8 行启动断言, 恢复 LoRA e2e CI 可用性, 并保留 `--offload-train-target=disk` 的必要断言。

```
# miles/utils/arguments.py —— miles_validate_args() 中 offload_train 相关分支

if args.offload_train:
    args.disable_grad_buffers_cpu_backup = True
    args.disable_param_buffers_cpu_backup = args.enable_weights_backuper

# #2077 曾在这里插入断言, 拒绝
# --disable-weights-backuper + LoRA + --colocate + offload-train 的组合;
# 本 PR (#2203) 将其删除, 原因是 --colocate 会在参数解析中强制
# offload_train=True, 而 LoRA e2e CI 脚本普遍同时传入
# --colocate 与 --disable-weights-backuper, 导致启动即失败。
# 真正不可恢复的组合仍由下方保留的断言拦截:

if args.offload_train_target == "disk":
    assert args.offload_train, "--offload-train-target=disk requires --offload-train"
    assert args.enable_weights_backuper, (
        "--offload-train-target=disk requires the weights backuper (do not pass "
        "--disable-weights-backuper): disk-offloaded weights are read from GPU after resume, "
        "not from a CPU backup."
    )
```

评论区精华

本 PR 没有任何 review 评论或讨论线程, 仅 `yueming-yuan` 给予 APPROVED 空评论。PR body 本身就是完整的论证: 它明确指出被删除断言所防御的故障 (weight sync 读取暂停 GPU 张量导致 30 分钟 gloo 超时) 更适合放在实际故障点处理, 并由 #2204 落地, 这一设计判断是本次回滚的核心依据。

- 暂无高价值评论线程

风险与影响

- 风险:
 1. 回滚后, 用户组合 `--colocate + LoRA + --disable-weights-backuper` (实际 `offload_train` 被强制开启) 时, 可能重新暴露 #2077 试图防御的问题: `named_params_and_buffers()` 将暂停的 GPU 张量交给 `update_weights`, 表现为 30

分钟 gloo barrier 超时。缓解依赖 #2204 的 fail-fast 是否已合入，若 #2204 尚未落地，则存在一段风险窗口。

2. 由于只删除了一条启动期断言，不引入新代码路径，参数解析逻辑的回归风险很低；而 `--offload-train-target=disk` 的硬性限制仍被保留，不会出现真正不可恢复的配置漏网。
3. 无新增测试，CI 恢复与否是唯一验证手段，若未来重构 LoRA e2e 脚本变更参数组合，可能再次触发类似冲突。 - 影响：影响范围集中在 LoRA + colocate + offload-train 场景：5 个 LoRA e2e CI 套件（glm5_1、glm5_2、gpt_oss_20b_moe、qwen3_5_35b_a3b、inkling_small 的 LoRA 变体）将从启动崩溃恢复为正常执行；用户侧该参数组合重新可用，但 `--offload-train-target=disk` 无备份的不可恢复组合仍被拒绝。对团队而言，需要跟进 #2204 合入，确保真正的故障点有 fail-fast 保护。 - 风险标记：回滚防御性断言，依赖后续修复 #2204，无新增测试覆盖

关联脉络

- PR #2077 Reject `--disable-weights-backuper` for LoRA + colocate + offload-train: 本 PR 直接回滚该提交（7546154083e209e71292ae665fdebec181491f9f），需要结合其原始意图评估风险。
- PR #2204 add fail-fast in `_maybe_get_cpu_backup()` (title inferred from PR body): PR body 中声明的后续修复，在 `miles/backends/megatron_utils/update_weight/common.py` 的 `_maybe_get_cpu_backup()` 中直接抛错，替代被删除的启动期断言。