PR #6688 完整报告

verl-project/verl

[rollout] fix: clone LoRA weights out of the reused IPC buffer before add_lora

合并时间：2026-06-12 14:21

原文链接：http://prhub.com.cn/verl-project/verl/pull/6688

执行摘要

- 一句话：修复 LoRA 权重在 IPC 缓冲中的视图导致崩溃
- 推荐动作：值得快速合入。该 PR 修复了一个明确的 tensor 生命周期 bug，改动极小且安全。对于理解 verl 中 IPC 缓冲区与 LoRA 权重的交互有参考价值。

功能与动机

Issue #6454 报告：当使用 unmerged LoRA（lora.merge=False）、vLLM rollout 后端和 free_cache_engine=True 时，训练在 step 1 之后每步都因 cudaErrorIllegalAddress 崩溃。合并 LoRA 路径（lora.merge=True）工作正常。根因是 _update_weights 中的 dict(weights) 创建了 IPC 桶缓冲区的视图，add_lora 保留这些视图，而后续桶覆盖或释放缓冲区，导致非法内存访问。

实现拆解

1. 在 verl/workers/rollout/vllm_rollout/utils.py 的 _update_weights 方法中，将 weights = dict(weights) 替换为 weights = {name: tensor.clone() for name, tensor in weights}。
2. 改动仅 1 行，但关键：clone() 为每个权重张量分配独立存储，避免视图依赖于被回收的 IPC 缓冲区。
3. 与标准权重加载路径（已使用 copy_）保证一致的存储所有权语义。
4. 无 API 变更，无测试配套（GPU-only 路径，作者提供了 numpy 机制演示验证）。

关键文件：

- verl/workers/rollout/vllm_rollout/utils.py（模块 rollout；类别 source；类型 core-logic；符号 _update_weights）：核心修改文件：修复 _update_weights 中 LoRA 权重的存储所有权问题。

关键符号：_update_weights

关键源码片段

verl/workers/rollout/vllm_rollout/utils.py

核心修改文件：修复 _update_weights 中 LoRA 权重的存储所有权问题。

```
# verl/workers/rollout/vllm_rollout/utils.py 第 306-319 行 (head 版本 )
def _update_weights(self, weights: list[tuple[str, torch.Tensor]], peft_config: dict, base_sync_
```

```
done: bool):
    if peft_config and base_sync_done:
        # 修复：克隆张量，避免 add_lora 保留 IPC 桶缓冲区的视图
        # 原代码 weights = dict(weights) 只创建视图，缓冲区被覆盖后导致崩溃
        weights = {name: tensor.clone() for name, tensor in weights}
        lora_request = TensorLoRARequest(
            lora_name=VLLM_LORA_NAME,
            lora_int_id=VLLM_LORA_INT_ID,
            lora_path=VLLM_LORA_PATH,
            peft_config=peft_config,
            lora_tensors=weights,
        )
        self.add_lora(lora_request)
        logger.info(f"vLLM load weights, loaded_params: {len(weights)}")
    else:
        # 非 LoRA 路径：使用标准 weight_loader，因为其中会复制数据，所以无此问题
        param_updates, buffer_updates, named_buffers = split_buffer_updates(self.model_runner.
        model, weights)
        # ... 后续逻辑
```

评论区精华

该 PR 无人工 review 评论；gemini-code-assist[bot] 自动审查未提出问题。维护者 wuxibin89 直接批准。

- 暂无高价值评论线程

风险与影响

- 风险：风险极低：仅修改 1 行，将 dict(weights) 替换为显式 clone()，这正是标准权重加载路径的安全做法。clone() 产生深拷贝，不会改变行为语义且不影响性能（LoRA 权重数据量通常较小）。
- 影响：影响范围：修复 unmerged LoRA + vLLM rollout + free_cache_engine 组合下的崩溃，使用该配置的用户将不再遇到 cudaErrorIllegalAddress。合并 LoRA 路径不受影响。
- 风险标记：核心路径变更（权重更新），缺少 GPU 单元测试覆盖

关联脉络

- PR #5599 [megatron] Qwen3.5 LoRA & MTP support: 相关 PR，引入了 LoRA 支持，本 PR 修复了其中未覆盖的 IPC 缓冲区生命周期问题。
- PR #5801 [vllm, fsdp] fix: apply FSDP buffer updates during rollout weight sync: 相关 PR，涉及 rollout 权重更新路径中的 IPC 缓冲区处理，本 PR 补充了 LoRA 场景下的修复。