

PR #46851 完整报告

vllm-project/vllm

[ROCm][CI] Fix rlhf_nccl.py on ROCm

合并时间: 2026-06-27 04:41

原文链接: <http://prhub.com.cn/vllm-project/vllm/pull/46851>

执行摘要

- 一句话: 修复 ROCm 上 RLHF 示例的 RCCL 运行时 bug
- 推荐动作: 此 PR 是一次有针对性的 bug 修复, 代码量小且逻辑清晰。值得关注的是其采用的临时方案设计——通过清除环境变量绕过后端 bug, 而不是修改核心库。作为 ROCm 平台兼容性修复的示例, 可供参考。

功能与动机

在 ROCm 平台上运行时, `rlhf_nccl.py` 因 RCCL 处理 `HIP_VISIBLE_DEVICES` 和 `CUDA_VISIBLE_DEVICES` 环境变量的运行时 bug 导致崩溃。此 PR 提供了一个临时工作区以绕过该问题, 直到 RCCL 上游修复。

实现拆解

实现分为三步:

1. 导入新的依赖: 在 `examples/rl/rlhf_nccl.py` 中添加了 `import torch` 和 `from vllm.platforms import current_platform`, 用于平台检测和设备设置。
2. 新增 `get_assigned_gpu()` 函数: 该函数首先检查当前平台是否为 ROCm; 如果不是, 直接返回 0 (保持原有行为)。如果是 ROCm, 则通过 `ray.get_gpu_ids()[0]` 获取实际分配的 GPU ID, 然后清除可能干扰 RCCL 的 `CUDA_VISIBLE_DEVICES` 和 `HIP_VISIBLE_DEVICES` 环境变量, 最后调用 `torch.accelerator.set_device_idx(assigned_gpu)` 设置正确的设备索引。
3. 修改 `TrainModel.__init__`: 将原来硬编码的 `.to("cuda:0")` 替换为调用 `get_assigned_gpu()` 获取实际 GPU ID 并构造 `.to(f"cuda:{assigned_gpu}")`, 确保 Trainer 模型加载到正确的 GPU 上。

关键文件:

- `examples/rl/rlhf_nccl.py` (模块 示例脚本; 类别 `source`; 类型 `core-logic`; 符号 `get_assigned_gpu`): 唯一的变更文件, 新增 `get_assigned_gpu()` 函数并修改 `TrainModel.__init__` 使用动态 GPU ID。

关键符号: `get_assigned_gpu`

关键源码片段

examples/rl/rlhf_nccl.py

唯一的变更文件，新增 `get_assigned_gpu()` 函数并修改 `TrainModel.__init__` 使用动态 GPU ID。

```
# examples/rl/rlhf_nccl.py

def get_assigned_gpu():
    """
    这是一个临时工作区，用于绕过 ROCm 上 RCCL 处理 VISIBLE_DEVICES
    环境变量时的运行时 bug。
    """
    # 非 ROCm 平台保持原行为，直接返回 GPU 0
    if not current_platform.is_rocm():
        return 0

    # ROCm 平台：通过 Ray 获取实际分配的 GPU ID
    assigned_gpu = int(ray.get_gpu_ids()[0])

    # 清除可能导致 RCCL 混乱的环境变量
    os.environ.pop("CUDA_VISIBLE_DEVICES", None)
    os.environ.pop("HIP_VISIBLE_DEVICES", None)

    # 显式设置当前设备索引，确保后续操作使用正确的 GPU
    torch.accelerator.set_device_idx(assigned_gpu)

    return assigned_gpu
```

评论区精华

审核者 micah-wil 询问是否有向 RCCL 提交 bug 工单 (line 51)。AndreasKaratzas 回应称没有公开工单，但 @charlifufu 已与 RCCL 团队的工程师沟通。micah-wil 指出该问题可能关联其之前提交的 ROCm 问题 #5756。charlifufu 回复称 RCCL 团队表示修复运行时 bug 需要很长时间，因此需要此临时方案。

- RCCL bug 上游修复计划 (question): 确认 RCCL 上游修复短期内不可用，接受此临时方案。

风险与影响

- 风险：风险较低。该变更仅在 ROCm 平台上执行额外逻辑（清除环境变量和设置设备索引），非 ROCm 平台行为不变。清除环境变量可能影响其他依赖这些变量的组件，但在 RLHF 示例的上下文中影响可控。缺少直接针对新函数的单元测试，但合并者已批准。
- 影响：影响范围仅限于 ROCm 平台上运行 RLHF 示例的用户。修复了 RLHF 工作流在 ROCm 上的崩溃问题，使 AMD GPU 用户能够正常使用 RLHF 功能。非 ROCm 平台无影响。
- 风险标记：平台特定修复，临时方案，缺少测试覆盖

关联脉络

- 暂无明显关联 PR