

PR #6346 完整报告

verl-project/verl

[trainer] feat: add set_expandable_segments support for npu

合并时间: 2026-05-18 11:43

原文链接: <http://prhub.com.cn/verl-project/verl/pull/6346>

执行摘要

- 一句话: NPU 内存 allocator 设置重构: 从环境变量迁移到统一函数
- 推荐动作: 建议精读此 PR, 并关注其中引入的回归问题。虽然 PR 本身方向正确 (统一路径), 但缺少对 `TrainingWorker.__init__` 中调用 `set_expandable_segments(True)` 的补充, 可能导致 NPU 上功能退化。推荐在合并后补充该调用。

功能与动机

之前 NPU 的 `expandable segments` 是通过设置环境变量 `PYTORCH_NPU_ALLOC_CONF` 启用的, 但这种做法不够优雅, 且与 CUDA 路径使用的 `set_expandable_segments` 函数不统一。PR 旨在将 NPU 路径也纳入到统一的 `set_expandable_segments` 函数中, 实现代码复用和统一管理。

实现拆解

1. 在 `verl/utils/device.py` 的 `set_expandable_segments` 函数中新增 NPU 支持: 在原有的 CUDA 条件分支之后, 添加 `elif is_npu_available`: 分支, 尝试调用 `torch.npu.memory._set_allocator_settings`, 如果失败 (如 `torch-npu` 版本过低) 则记录警告日志, 提示用户升级。
2. 从 `verl/workers/engine_workers.py` 中删除旧的 NPU 配置: 移除 `TrainingWorker.__init__` 中通过环境变量 `PYTORCH_NPU_ALLOC_CONF` 启用 `expandable segments` 的代码块以及相关的 TODO 注释。
3. 清理导入: 删除 `engine_workers.py` 中对 `is_npu_available` 的导入, 因为该变量不再被使用。

关键文件:

- `verl/utils/device.py` (模块 工具函数; 类别 `source`; 类型 `core-logic`; 符号 `set_expandable_segments`): 核心逻辑变更: 在 `set_expandable_segments` 函数中新增对 NPU 的支持, 包括异常处理, 是本次 PR 的主要功能实现。
- `verl/workers/engine_workers.py` (模块 工作节点; 类别 `source`; 类型 `dependency-wiring`): 删除了旧的 NPU 环境变量配置和 TODO 注释, 并移除了 `is_npu_available` 的导入, 是本次 PR 的清理目标。

关键符号: `set_expandable_segments`

关键源码片段

verl/utils/device.py

核心逻辑变更：在 `set_expandable_segments` 函数中新增对 NPU 的支持，包括异常处理，是本次 PR 的主要功能实现。

```
def set_expandable_segments(enable: bool) -> None:
    """CUDA / NPU 内存分配器的 expandable segments 设置。`
    # 注释：该函数现在同时支持 CUDA 和 NPU,
    # 通过 try-except 兼容旧版本 torch-npu。
    """

    if is_cuda_available:
        torch.cuda.memory._set_allocator_settings(f"expandable_segments:{enable}")
    elif is_npu_available:
        try:
            torch.npu.memory._set_allocator_settings(f"expandable_segments:{enable}")
        except Exception:
            logger.warning(
                "Current version of torch-npu does not support "
                "`_set_allocator_settings`, "
                "please upgrade torch-npu to 2.9.0 or later"
            )
```

评论区精华

在 review 中，[gemini-code-assist\[bot\]](#) 指出，删除环境变量设置后，`TrainingWorker` 的 `__init__` 中没有显式调用 `set_expandable_segments(True)`，导致 NPU 上该功能默认关闭，属于回归。建议在 `__init__` 中调用该函数以维持之前的行为。该评论未被回复或解决，但 PR 最终由 [wuxibin89](#) 批准合并。

- 删除环境变量后未调用新函数导致回归 (correctness): 未明确解决，但 PR 被批准合并。

风险与影响

- 风险：存在回归风险：`TrainingWorker` 及其子类在 NPU 上初始化时，`set_expandable_segments` 可能未被调用，导致 expandable segments 处于禁用状态，可能增加 OOM 风险。此问题已被审核者指出但未修复，需要后续处理。
- 影响：直接影响 NPU 上的所有训练工作负载。如果 `set_expandable_segments` 未在合适位置被调用，NPU 上的内存分配行为将不同于之前，可能导致性能下降或 OOM。影响范围限于 NPU 用户，CUDA 用户不受影响。
- 风险标记：回归风险，核心路径变更，未解决审查意见

关联脉络

- 暂无明显关联 PR