PR #6170 完整报告

verl-project/verl

[vllm] chore: fix sleep_level in Ascend

合并时间：2026-04-27 18:58

原文链接：http://prhub.com.cn/verl-project/verl/pull/6170

执行摘要

- 一句话：修复 Ascend NPU 上 vLLM sleep_level 导致精度问题
- 推荐动作：建议阅读 review 评论了解潜在不完修复风险。如果团队使用 Ascend NPU 且开启 EP，此变更必要但可能不充分，需进一步验证 ServerAdapter 行为。

功能与动机

修复 Ascend NPU 上训练时启用 Expert Parallelism (EP) 导致精度异常的问题。
vllm_ascend 不支持 sleep_level=2，EP 开启时会引发精度问题。

实现拆解

1. 修改条件表达式：在 verl/workers/rollout/vllm_rollout/vllm_async_server.py 的 _sleep_hybrid 方法中，将 if self.lora_as_adapter: 改为 if self.lora_as_adapter or is_torch_npu_available(check_device=False):。
2. 增加注释说明：在条件前添加注释 "vllm_ascend not support sleep_level now. Enabling EP during training may lead to accuracy issues."，解释变更动机。
3. 影响：Ascend NPU 设备（包括 Lora 和全权重场景）均使用 sleep_level=1，避免 EP 导致的精度异常。

关键文件：

- verl/workers/rollout/vllm_rollout/vllm_async_server.py（模块 rollout；类别 source；类型 core-logic；符号 _sleep_hybrid）：核心修改文件，_sleep_hybrid 方法增加了 NPU 判断，强制使用 sleep_level=1 避免 EP 精度问题。

关键符号：_sleep_hybrid

关键源码片段

verl/workers/rollout/vllm_rollout/vllm_async_server.py

核心修改文件，_sleep_hybrid 方法增加了 NPU 判断，强制使用 sleep_level=1 避免 EP 精度问题。

```
async def _sleep_hybrid(self):
    """HYBRID sleep: lora adapters only need level=1; full weights need level=2."""
    # Don't use engine.sleep(level=2) here
```

```
# lora only update adapter weights, so set sleep level to 1
# vllm_ascend not support sleep_level now. Enabling EP during training may lead to accuracy issues.
if self.lora_as_adapter or is_torch_npu_available(check_device=False): # 增加 NPU 判断
    sleep_level = 1
else:
    sleep_level = 2
await self.engine.collective_rpc("sleep", kwargs={"level": sleep_level})
if _VLLM_VERSION >= version.parse("0.17.0"):
    await self.engine.reset_encoder_cache()
```

评论区精华

由 gemini-code-assist[bot] 提出的 review 指出：

- 该修复仅影响 vLLMReplica.sleep()，但标准训练流程（如 PPO）中内存管理由 ServerAdapter.release() 处理，后者在 __init__ 中独立计算 sleep_level 并通过 collective_rpc 直接传递，绕过了 _sleep_hybrid 逻辑。因此本修改可能未覆盖实际生效路径。
- 建议将条件细化到仅针对启用 EP 的情况，以避免在非 EP 的 NPU 配置下强制提高内存占用。
- 当前状态：该 issue 未解决，但 PR 已被审核者 wuxibin89 批准合并。
- 修复是否覆盖了实际生效路径 (correctness): 未解决。PR 已被批准，但 reviewer 的疑虑未被回应。
- 条件是否应该只针对 EP 启用时 (design): 未采纳。当前实现为所有 NPU 启用 sleep_level=1。

风险与影响

- 风险：
  1. 不完修复风险：如 review 指出，ServerAdapter 可能独立管理 sleep_level，本变更可能未覆盖实际生效路径，NPU 上仍有精度问题。
  2. 内存占用增加：强制 NPU 使用 sleep_level=1（保留更多权重内存），对非 EP 配置可能不必要。
  3. 回归风险：低。仅增加一个条件判断，不影响已有逻辑。 - 影响：直接影响 Ascend NPU 上的 vLLM 推理引擎睡眠行为；由于变更极小，对现有其他硬件（如 GPU）无影响。当前精度问题仅在 EP 启用时出现，非 EP 用户可忽略。 - 风险标记：修复可能不完整，未覆盖 ServerAdapter 路径

关联脉络

- 暂无明显关联 PR