

PR #7354 完整报告

verl-project/verl

[ci] chore: fix npu ci

合并时间: 2026-08-11 21:05

原文链接: <http://prhub.com.cn/verl-project/verl/pull/7354>

执行摘要

- 一句话: 适配 Megatron v0.16.0 保存参数并修复 NPU CI
- 推荐动作: 不值得精读, 扫一眼即可。值得记住的模式是: 当上游依赖 (如 Megatron core) 调整构造函数默认值时, 把关键参数显式化是低成本且安全的修复方式; 但建议后续补一个直接覆盖 `save_dist_checkpointing` 的参数化单元测试, 避免这类适配再次回归。

功能与动机

PR body 原文: `fix megatron TorchDistSaveShardedStrategy version and backend default value in megatron core v0.16.0`。Megatron core 升级后, 保存策略无参构造的默认行为不再可靠, 需要显式指定 `backend` 与 `version`; 同时 `preprocess_thd_engine` 新增 `local_cp_size` 参数使既有测试断言失配。两者都表现为 NPU CI 红, 因此打包为一次 "fix npu ci" 提交。

实现拆解

1. 保存策略参数显式化: 在 `verl/utils/megatron/dist_checkpointing.py` 的 `save_dist_checkpointing` 中, 将 `TorchDistSaveShardedStrategy()` 改为 `TorchDistSaveShardedStrategy(backend='torch_dist', version=1)`。原因是 Megatron core v0.16.0 修改了该构造函数的默认参数语义, `backend` 默认值会让 `dist_checkpointing.save` 行为偏移; 显式传参可保证保存格式与旧版本一致, `version=1` 也是对 checkpoint 数据格式版本的固定。该改动影响所有走 `save_dist_checkpointing` 的 Megatron 后端保存路径 (包括 NPU 与 GPU)。load 路径未变, 因为本次问题只出现在保存策略。
2. 适配 fused forward 测试断言: `tests/models/test_model_forward_fused.py` 中两处 `preprocess_calls` 期望值增加 `local_cp_size: None` 键。这是对 `verl/models/mcore/model_forward_fused.py` 中 `preprocess_thd_engine` 新增 `local_cp_size` 参数 (本次或近期其他改动引入) 的同步, 否则断言失败造成 CI 红。
3. 提交与合并模式: 5 个 commit 均为 "fix ci" 迭代, 无独立 review 评论, 最终由 wuxibin89 批准合并。测试改动与源码修复在功能上无直接关联, 说明这是一个打包修复 CI 红点的维护型 PR。

关键文件:

- verl/utils/megatron/dist_checkpointing.py (模块 检查点保存; 类别 source; 类型 core-logic; 符号 save_dist_checkpointing) : 核心修复文件: 为 TorchDistSaveShardedStrategy 显式传入 backend 与 version, 适配 Megatron core v0.16.0 默认值变更, 恢复 checkpoint 保存行为。
- tests/models/test_model_forward_fused.py (模块 前向测试; 类别 test; 类型 test-coverage) : 同步适配 preprocess_thd_engine 新增 local_cp_size 参数导致的断言失败, 属于 CI 修复的一部分; 与源码改动无直接功能关联。

关键符号: save_dist_checkpointing

关键源码片段

verl/utils/megatron/dist_checkpointing.py

核心修复文件: 为 TorchDistSaveShardedStrategy 显式传入 backend 与 version, 适配 Megatron core v0.16.0 默认值变更, 恢复 checkpoint 保存行为。

```
def save_dist_checkpointing(
    sharded_state_dict,
    ckpt_path,
    async_save=False,
    content_metadata=None,
):
    """保存分布式分片 checkpoint。
```

Megatron core v0.16.0 调整了 TorchDistSaveShardedStrategy 的默认参数语义, 这里显式指定 backend 与 version, 保证保存行为与旧版本一致。

```
"""
validate_sharding_integrity = True
# 显式传参以适配 mcore v0.16.0 的 API 变更;
# backend="torch_dist" 指定后端实现, version=1 固定 checkpoint 格式版本
save_strategy = TorchDistSaveShardedStrategy(backend="torch_dist", version=1)
save_strategy = FullyParallelSaveStrategyWrapper(
    save_strategy, mpu.get_data_parallel_group(with_context_parallel=True)
)
# 参考 Megatron-LM core_v0.14.0 中 distrib_optimizer.py 的保存路径
mcore_ge_014 = version.parse(megatron.core.__version__) >= version.parse("0.14.0")
save_kwargs = dict(
    sharded_strategy=save_strategy,
    async_sharded_save=async_save,
    validate_access_integrity=validate_sharding_integrity,
)
if content_metadata is not None and mcore_ge_014:
    save_kwargs["content_metadata"] = content_metadata
return dist_checkpointing.save(sharded_state_dict, ckpt_path, **save_kwargs)
```

评论区精华

该 PR 没有任何 review 评论或讨论线程 (comments_count = 0、review_comments_count = 0)。wuxibin89 先后给出 DISMISSED 与 APPROVED 两次 review，最终直接合并，没有可提炼的交锋。

- 暂无高价值评论线程

风险与影响

- 风险：
 - checkpoint 格式版本硬编码：version=1 是显式固定值，若未来 Megatron 再次调整格式版本或 verl 需要迁移到新版本格式，此处需同步更新，且旧 checkpoint 的兼容性依赖 Megatron 侧行为。
 - load 路径未同步验证：本次只改 save_dist_checkpointing，load_dist_checkpointing 仍用无参 TorchDistLoadShardedStrategy()；如果 v0.16.0 对 load 侧默认值也有变更，可能在加载新保存的 checkpoint 时报错，但本次 NPU CI 未暴露该问题。
 - 测试与源码无直接对应关系：test_model_forward_fused.py 的改动与 dist_checkpointing.py 无功能关联，说明并无针对保存策略修复的专门测试；该测试是否在 NPU CI 覆盖到也不明确。
 - 影响：影响范围集中在 Megatron 后端的分片 checkpoint 保存路径 (dist_checkpointing.save)，涉及使用 Megatron core v0.16.0 的所有用户 (NPU 与 GPU)。由于改动只是把隐式默认值变显式，对已在新版本下失败的用户是修复，对旧版本则是行为保持。影响面窄、程度低，不涉及训练主链路。
- 风险标记：依赖版本适配，checkpoint 格式硬编码，保存修复缺专项测试

关联脉络

- PR #7264 [ckpt, megatron] fix: megatron save checkpoints with strict false when vanilla_bridge is false: 同属 Megatron checkpoint 保存路径的修复，改动相邻模块 (megatron_checkpoint_manager.py)。
- PR #7348 [megatron] feat: cache the Megatron-Bridge HF export plan across weight updates: Megatron 后端近期改动，与本 PR 同处 verl/utils/megatron 与 worker 引擎链路。
- PR #7337 [ci] chore: add three baselines for npu's nightly ci: 同为 NPU CI 维护，反映 Ascend 侧 CI 持续演进背景。