

PR #3162 完整报告

radixark/miles

ci: pin the DSv4 bshd/thd parity test to the miles impl

合并时间: 2026-09-10 06:59

原文链接: <http://prhub.com.cn/radixark/miles/pull/3162>

执行摘要

- 一句话: 修复 DeepSeek-V4 精度对等测试因默认实现变更导致的 CI 失败。
- 推荐动作: 这是一个小而必要的 bugfix PR, 值得精读, 但主要价值在于理解其修复的两个上游变更的联动影响。对于开发者, 关注点应在于: 1) 测试如何与特定实现 (dsv4_impl) 耦合; 2) 当核心系统行为 (如数据转储策略) 变更时, 配套测试需要如何演进。对于管理者, 此 PR 展示了维护跨 PR 依赖 (如 PR#3141 与本测试) 的 CI 实践。变更本身无复杂设计, 但清晰展示了如何响应上游变更来保持测试有效性。

功能与动机

根据 PR body, 该测试在 nightly CI 中失败了两次, 原因已明确:

1) DSv4 启动器默认实现变更为 **megatron** (TP=1, DP=4), 而测试为 **miles** 实现 (TP=4, DP=1) 编写, 其 **eval_global_batch_size** 参数触发了 Megatron 的验证错误。2) 训练数据转储逻辑变更 (PR#3141) 导致输出文件数量变化, 测试仍预期每个 rank 一个文件。为确保测试有效并修复 CI, 需要同时解决这两个问题。

实现拆解

1. 锁定测试实现: 在测试文件 `tests/e2e/precision/test_dsv4_bshd_thd_rollout_parity.py` 的 `_prepare_args` 和 `_run_args` 函数中, 为 `ScriptArgs` 构造显式添加 `dsv4_impl="miles"` 参数。这确保了模型准备 (checkpoint 转换) 和训练运行都使用相同的、测试所针对的 Miles 实现 (TP=4, DP=1), 避免了与默认 Megatron 实现的参数不兼容问题。
2. 适配数据转储策略变更: 将函数 `_load_rank_zero_train_data` 重命名为更准确的 `_load_train_data_shard`。重写其内部逻辑: 不再按 rank 预期多个文件 (`{rollout_id}_{rank}.pt`), 而是使用 glob 模式 `{rollout_id}*.pt` 加载, 并断言恰好找到一个文件 (因为 DP=1, CP=1)。同时移除了对 `payload["rank"]` 的检查, 因为分片逻辑不再依赖 rank 信息。这个改动与 PR#3141 引入的“每个 DP/CP 分片写一次”策略对齐。
3. 验证与回归: 作者通过 `/rerun-test` 命令在分支上两次成功运行了该测试 (链接在 PR body 中), 确认修复有效。本 PR 仅修改了一个测试文件, 无源码主路径或配置变更。

关键文件:

- `tests/e2e/precision/test_dsv4_bshd_thd_rollout_parity.py` (模块 精度测试; 类别 test; 类型 test-coverage; 符号 `_load_rank_zero_train_data`, `_load_train_data_shard`): 这是

本次 PR 唯一修改的文件，包含了所有修复逻辑。它针对 DeepSeek-V4 模型的 BSHD 与 THD 注意力格式进行 rollout 精度对等测试，修复后能适应默认实现变更和训练数据转储策略优化。

关键符号：_load_train_data_shard, _load_rank_zero_train_data, _run_args, _prepare_args

关键源码片段

tests/e2e/precision/test_dsv4_bshd_thd_rollout_parity.py

这是本次 PR 唯一修改的文件，包含了所有修复逻辑。它针对 DeepSeek-V4 模型的 BSHD 与 THD 注意力格式进行 rollout 精度对等测试，修复后能适应默认实现变更和训练数据转储策略优化。

修复后的数据加载函数，适应“每个 DP/CP 分片写一次”的转储策略。

函数名从 _load_rank_zero_train_data 更改为 _load_train_data_shard，语义更准确。

```
def _load_train_data_shard(directory: Path) -> dict:
```

```
    # Train data is dumped once per (dp, cp) shard; this run has dp=1, cp=1.
```

```
    # 使用 glob 模式匹配 rollout_id 前缀的 .pt 文件，而非按 rank 预期固定列表。
```

```
    dump_files = sorted(directory.glob(f"{_ROLLOUT_ID}*.pt"))
```

```
    # 断言恰好找到一个分片文件，因为本测试运行配置为 DP=1, CP=1。
```

```
    assert len(dump_files) == 1, [path.name for path in dump_files]
```

```
    payload = torch.load(dump_files[0], map_location="cpu", weights_only=False)
```

```
    assert payload["rollout_id"] == 0
```

```
    # 移除了对 payload["rank"] == 0 的检查，因为分片逻辑不再依赖 rank。
```

```
    rollout_data = payload["rollout_data"]
```

```
    assert rollout_data["sample_indices"] == list(range(_NUM_SAMPLES))
```

```
    return rollout_data
```

在 _run_args 函数中，显式指定 dsv4_impl="miles"，确保测试与 Miles 实现 (TP=4, DP=1) 对齐。

```
def _run_args(
```

```
    *,
```

```
    run_id: str,
```

```
    debug_root: Path,
```

```
    hf_checkpoint: str,
```

```
    extra_args: str,
```

```
    debug_train: bool,
```

```
) -> ScriptArgs:
```

```
    return ScriptArgs(
```

```
        run_id=run_id,
```

```
        model_name=_MODEL_NAME,
```

```
        dsv4_impl="miles", # 锁定实现，避免默认 megatron 实现的参数不兼容。
```

```
        task="gsm8k",
```

```
        # ... 其余参数保持不变 ...
```

```
)
```

评论区精华

审查过程直接且无争议。

- Claude Bot指出这是一个 fork PR，自动化审查被禁用，需维护者手动触发。
- yueming-yuan直接批准，无额外评论。这表明变更范围小且清晰，符合预期。
- 讨论的核心在于 Issue 评论中的失败和成功 rerun 记录，直接验证了修复的有效性。
- fork PR 自动化审查禁用 (other): 不影响 PR 的评审和合并流程，仅为流程提醒。
- 测试失败根因分析与验证 (correctness): 测试失败问题已通过本次修复得到解决，并通过 CI 验证。

风险与影响

- 风险：风险较低且集中。 1. 测试可靠性风险：修复后的测试逻辑 (`_load_train_data_shard`) 更泛化，但与 `dsv4_impl="miles"` 强绑定。如果未来 Miles 实现的 DP/CP 策略再次改变（例如引入 $CP>1$ ），此测试可能需要再次调整。目前风险可控。 2. 无源码风险：变更仅限于单个端到端测试文件，不影响训练、推理或核心库的生产逻辑。唯一风险是测试本身误报或漏报，但通过两次 CI 验证已降低。 3. 兼容性风险：此修复是针对上游变更（PR#3141 和 DSv4 默认实现切换）的配套调整，属于必要的跟进，不引入新的兼容性问题。
- 影响：直接影响：
 - CI 稳定性：修复了 nightly CI 中一个具体测试的失败，恢复了 DeepSeek-V4 BSHD/THD 精度对等验证的 CI 绿灯状态。
 - 测试维护：测试代码变得更健壮，能适应新的训练数据转储策略。

间接影响：

- 团队效率：减少了开发人员处理 CI 失败的干扰。
- 功能验证：确保了 DeepSeek-V4 模型在两种注意力格式（BSHD 和 THD）下的行为一致性得到持续监控，这对模型质量至关重要。

范围：影响仅限于 CI 测试套件和相关开发工作流，无用户面向功能变更。

- 风险标记：测试与特定实现耦合，依赖上游行为变更

关联脉络

- PR #3141 Dump debug train data once per DP/CP shard instead of on every rank: 本 PR 修复了因 PR#3141 引入的训练数据转储策略变更（从每个 rank 写一个文件改为每个 DP/CP 分片写一个文件）所导致的测试断言失败。
- PR #2038 [feat] Enable THD packed-sequence training for DeepSeek-V4-Flash, with context parallelism: 本 PR 修复的测试 (`test_dsv4_bshd_thd_rollout_parity.py`) 正是为了验证 DeepSeek-V4 模型在 THD 格式（由 PR#2038 启用）与 BSHD 格式下的行为一致性。