

PR #7610 完整报告

verl-project/verl

[ckpt, fsdp] fix: FSDP model merger concatenates replicated buffers

合并时间: 2026-08-31 14:45

原文链接: <http://prhub.com.cn/verl-project/verl/pull/7610>

执行摘要

- 一句话: 修复 FSDP 合并器误拼接 replicated buffer 崩溃与 shape 损坏
- 推荐动作: 值得精读。这是一个小而完整的数据契约修复: 核心决策是识别 FSDP 下“非 DTensor 即完整复制”的语义, 并用 fail-fast 替代静默损坏。可借鉴的点包括: 错误信息携带 key 名与 rank 号、用 torch.equal 做全量一致性校验、配套聚焦的 CPU 单测直接锁定回归。

功能与动机

PR body 明确指出 FSDP2 never shards buffers, so each rank holds an identical copy, 而 FSDPModelMerger 把所有非 DTensor 条目都当作行分片执行 torch.cat(dim=0), 导致两类错误: 0-d buffer 抛出 RuntimeError: zero-dimensional tensor (at position 0) cannot be concatenated, ≥ 1 -d buffer 被静默拼接成 world_size 倍并产生 shape 错误的 HF checkpoint。该问题由作者在转换 google/gemma-4-E4B-it 的 FSDP2 checkpoint 时实际触发, 属于分布式 checkpoint 转 HF 格式链路中的数据契约缺陷。

实现拆解

1. 新增合并辅助函数: 在 verl/model_merger/fsdp_model_merger.py 顶部新增模块级函数 merge_non_dtensor_shards(key, shards)。它遍历 shards[1:] 与 shards[0] 比较 shape 和数值 (torch.equal), 不一致时抛出带 key 名和 rank 号的 ValueError; 一致则直接返回 shards[0]。这样既避免 concat 崩溃, 又能对畸形 checkpoint 显式报错。
2. 替换合并分支: 在 _load_and_merge_state_dicts 的 else 分支 (非 DTensor 且不在 param_placements 中的条目) 将 torch.cat(state_dict[key], dim=0) 替换为 merge_non_dtensor_shards(key, state_dict[key])。该分支覆盖所有被 FSDP 完整复制的 buffer 和未包装参数。
3. 新增 CPU 单测: 新建 tests/model_merger/test_fsdp_merge_non_dtensor_shards_on_cpu.py, 包含 4 个测试: 标量 buffer 保持原样、1-d buffer 不被拼接、单 rank checkpoint 原样返回、各 rank 不一致时抛出含 key 名的错误 (覆盖同 shape 异值和异 shape 两类情况), 直接锁定新函数的数据契约。
4. 配置与部署配套: 无配置文件、schema 或部署脚本改动; 变更仅涉及源码与测试, 纯 CPU 可验证。

关键文件:

- verl/model_merger/fsdp_model_merger.py (模块 模型合并; 类别 source; 类型 data-contract; 符号 merge_non_dtensor_shards, _load_and_merge_state_dicts) : 核心修复文件: 新增 merge_non_dtensor_shards 替代 torch.cat, 消除 0-d buffer 崩溃和 ≥ 1 -d buffer 静默膨胀问题, 并加入跨 rank 一致性检测。
- tests/model_merger/test_fsdp_merge_non_dtensor_shards_on_cpu.py (模块 模型合并; 类别 test; 类型 test-coverage; 符号 test_replicated_scalar_buffer_is_kept_as_is, test_replicated_1d_buffer_is_not_concatenated, test_single_rank_checkpoint_returns_the_tensor, test_ranks_disagreeing_is_an_error_naming_the_key) : 新增 4 个 CPU 单测, 覆盖标量 buffer、1-d buffer、单 rank checkpoint 及异常路径, 锁定新函数的数据契约。

关键符号: merge_non_dtensor_shards, _load_and_merge_state_dicts

关键源码片段

verl/model_merger/fsdp_model_merger.py

核心修复文件: 新增 merge_non_dtensor_shards 替代 torch.cat, 消除 0-d buffer 崩溃和 ≥ 1 -d buffer 静默膨胀问题, 并加入跨 rank 一致性检测。

```
def merge_non_dtensor_shards(key: str, shards: list[torch.Tensor]) -> torch.Tensor:
    """Return the single copy of a plain (non-DTensor) checkpoint entry.

    FSDP only produces DTensors for what it actually shards; plain tensors are
    buffers and unwrapped parameters, replicated in full on every rank (or the
    whole tensor when world_size == 1).
    """
    for rank, shard in enumerate(shards[1:], start=1):
        # 每个 rank 的副本都必须与 rank 0 形状、数值完全一致, 否则说明 checkpoint 已损坏
        if shard.shape != shards[0].shape or not torch.equal(shard, shards[0]):
            raise ValueError(
                f"Non-DTensor entry {key!r} differs between rank 0 and rank {rank}; "
                "expected a replicated buffer/parameter identical on every rank"
            )
    # 直接保留 rank 0 的副本, 避免 torch.cat 把 buffer 错误拼接成 world_size 倍
    return shards[0]
```

tests/model_merger/test_fsdp_merge_non_dtensor_shards_on_cpu.py

新增 4 个 CPU 单测, 覆盖标量 buffer、1-d buffer、单 rank checkpoint 及异常路径, 锁定新函数的数据契约。

```
def test_replicated_scalar_buffer_is_kept_as_is():
    # 0-d buffer 被 FSDP2 复制到每个 rank, 旧实现 torch.cat 会直接抛错
    shards = [torch.tensor(-6.375) for _ in range(8)]
    merged = merge_non_dtensor_shards("clamp_min", shards)
    assert merged.shape == ()
    assert merged.item() == -6.375

def test_ranks_disagreeing_is_an_error_naming_the_key():
```

```
# 各 rank 数值不一致时，应抛出带 key 名的 ValueError，便于定位损坏条目
same_shape = [torch.full((2, 3), float(rank)) for rank in range(4)]
with pytest.raises(ValueError, match="'suspicious'.*rank 1"):
    merge_non_dtensor_shards("suspicious", same_shape)
```

评论区精华

PR 没有 reviewer 评论，Issue 评论仅为 CLA 签署确认和作者询问加入 Slack 的非技术内容。从提交历史可观察到作者第一笔提交实现基本逻辑后，第二笔提交 "Made check strictier" 将校验从仅比较 shape 升级为同时比较数值 ([torch.equal](#))，使各 rank 值不一致的损坏 checkpoint 也能被识别，属于作者自发的加固。

- 无 review 技术讨论，提交历史显示校验加固 (other): 作者自行加固校验后由维护者 wuxibin89 合并。

风险与影响

- 风险：行为假设风险：新逻辑依赖“非 DTensor 条目即 replicated buffer”的 FSDP 语义，若存在非标准用法导致非 DTensor 条目实际是分片，合并结果将错误（只保留 rank 0 副本）。性能风险：torch.equal 对每个 buffer 做全量比对，在超大 buffer 上有额外 O(N) 开销，但相对整个 checkpoint 的加载与合并成本可忽略。兼容性风险：对之前能“成功”合并但实际 shape 已损坏的 checkpoint，现在会直接抛 ValueError 中止，属于有意的 fail-fast，但可能阻塞依赖旧行为的用户。测试覆盖缺口：当前仅有单进程 CPU 单测，缺少真实多 rank 下 FSDP2 checkpoint 的端到端合并测试。
- 影响：影响用户：仅影响使用 FSDPModelMerger 将 FSDP2 checkpoint 转换为 HF 格式的用户，尤其模型含 persistent buffer（如 google/gemma-4-E4B-it）时，修复后不再崩溃或产出错误 shape。影响系统：合并链路对畸形 checkpoint 从静默容忍转为显式报错，提升产物正确性。影响团队：无配置或接口变更，其他后端（Megatron 等）不受影响，FSDP 从训练到 checkpoint 转换的整条链路更可信。
- 风险标记：核心合并路径变更，修复静默 shape 损坏，缺少多 rank 集成测试，依赖非 DTensor 即复制的假设

关联脉络

- PR #7458 [fsdp] fix: make deferred gradient sync configurable: 同属 FSDP 功能线，改动 FSDP 训练侧行为；本 PR 修正 FSDP checkpoint 合并侧，两者共同完善 FSDP 从训练到产物转换的全链路。
- PR #7513 [trainer, ckpt, cfg] feat: add config-driven checkpoint callback hook: 同属 checkpoint 工具链，前者在训练侧控制保存时机，本 PR 修复保存后合并产物的正确性，属于同一功能方向的不同环节。