

PR #6360 完整报告

verl-project/verl

Revert "[fsdp] fix: emit distillation outputs in use_remove_padding=False path (#6293)"

合并时间: 2026-05-15 10:38

原文链接: <http://prhub.com.cn/verl-project/verl/pull/6360>

执行摘要

- 一句话: 回退 distillation 在 use_remove_padding=False 的修复
- 推荐动作: 此 PR 值得关注其背后的设计权衡: 在分布式蒸馏输出修复中, 如何同时保证两个 padding 分支的对称性与序列并行兼容性。建议阅读 #6350 和相关讨论以理解完整上下文。短期内, 受影响用户可切换至 use_remove_padding=True 作为临时规避。

功能与动机

原修复 #6350 在 use_remove_padding=False 路径添加蒸馏输出后, 可能引入了序列并行 (Ulysses SP) 不兼容或其他运行时错误, 且该路径在当前配置下并不启用 Ulysses SP, 因此合并者决定回退以保持稳定。

实现拆解

1. 移除蒸馏输出逻辑: 在 verl/workers/engine/fsdp/transformer_impl.py 的 prepare_model_outputs 方法中, 删除了位于 pad_mode == DatasetPadMode.NO_PADDING 分支内、log_probs = logprobs_from_logits(...) 之后的一段代码块 (1188-1201 行)。该块用于在 distillation_use_topk=True 时调用 logits_processor_func 并将结果以嵌套张量形式写入 model_output。
2. 删除回归测试: 整体移除了 tests/workers/test_distillation_topk_symmetry_on_cpu.py 文件, 该文件是 #6350 新增的 CPU 单元测试, 用于验证两个 padding 模式下蒸馏输出均能被正确发出。
3. 无其他配套改动: 回退后引擎逻辑恢复到修复前的状态, 即只有 use_remove_padding=True 分支包含蒸馏处理, False 分支仍缺失该逻辑。

关键文件:

- tests/workers/test_distillation_topk_symmetry_on_cpu.py (模块 蒸馏测试; 类别 test; 类型 deletion; 符号 _make_engine_stub, _EngineCfg, _make_logits_processor, _proc): 被整体删除。该文件是 #6350 新增的 CPU 回归测试, 用于验证两个 padding 模式下的蒸馏输出正确性。删除后该回归路径失去防护。
- verl/workers/engine/fsdp/transformer_impl.py (模块 FSDP 引擎; 类别 source; 类型 core-logic): 核心引擎文件, 删除了 use_remove_padding=False 路径下的蒸馏输出调用块, 是回退的实际作用点。

关键符号: FSDPEngineWithLMHead.prepare_model_outputs

关键源码片段

tests/workers/test_distillation_topk_symmetry_on_cpu.py

被整体删除。该文件是 #6350 新增的 CPU 回归测试，用于验证两个 padding 模式下的蒸馏输出正确性。删除后该回归路径失去防护。

```
# -*- coding: utf-8 -*-
# Copyright ...
"""
被 PR#6360 整体删除的回归测试文件（原为 #6350 新增）。
以下为被删除的核心测试函数，用于验证蒸馏输出在两种 padding 模式下均能被发出。
"""

import os
os.environ.setdefault("KMP_DUPLICATE_LIB_OK", "TRUE")
import pytest
import torch
from tensordict import TensorDict
from verl.utils import tensordict_utils as tu
from verl.utils.dataset.dataset_utils import DatasetPadMode
from verl.workers.engine.fsdpy.transformer_impl import FSDPEngineWithLMHead

_VOCAB_SIZE = 8
_DISTILLATION_KEYS = ("distillation_losses", "student_mass")

def _make_engine_stub():
    """绕过 __init__，仅设置测试所需属性。"""
    eng = object.__new__(FSDPEngineWithLMHead)
    eng.use_ulysses_sp = False
    class _EngineCfg:
        entropy_checkpointing = False
    eng.engine_config = _EngineCfg()
    return eng

@pytest.mark.parametrize("use_remove_padding", [True, False])
def test_distillation_outputs_emitted_in_both_padding_modes(use_remove_padding):
    """验证 distillation_use_topk=True 时 model_output 包含蒸馏键。"""
    # 构造输入，模拟 prepare_model_outputs 的两种 padding 路径
    ... # 略去具体构造细节，该测试共 141 行，被完全删除
```

verl/workers/engine/fsdp/transformer_impl.py

核心引擎文件，删除了 `use_remove_padding=False` 路径下的蒸馏输出调用块，是回退的实际作用点。

```
# 以下为 prepare_model_outputs 方法中 NO_PADDING 分支的关键片段（head 版本）。
# 被删除的块标注在注释中。
if pad_mode == DatasetPadMode.NO_PADDING:
    cu_seqlens = input_ids.offsets()
    ... # 此处省去 logits 转换等前序逻辑
    log_probs = logprobs_from_logits(logits=logits_rmpad, labels=input_ids_rmpad_rolled)
```

```
# (deleted block) 原 #1188-1201 行:
# if distillation_use_topk:
#     outputs = logits_processor_func(student_logits=logits_rmpad.unsqueeze(0), data=
micro_batch)
# for k, v in outputs.items():
#     v = v.squeeze(0)
#     assert v.shape == log_probs.shape, ...
#     model_output[k] = torch.nested.nested_tensor_from_jagged(v, cu_seqlens)
# 该块在 revert 中被完全移除，导致蒸馏键不再写入 model_output。

# (bsz, j1), 将 log_probs 转换为嵌套张量
log_probs = torch.nested.nested_tensor_from_jagged(log_probs, cu_seqlens)
... # 后续正常逻辑
```

评论区精华

仅有一条来自 [gemini-code-assist\[bot\]](#) 的 review 评论，指出回退会重新引入 #6293 描述的回归：[distillation_losses](#) 等蒸馏键将从模型输出中缺失，导致 loss 计算时 [KeyError](#)。评论建议先修复底层序列并行逻辑，而不是直接删除该分支的处理。但该评论未被 resolved，PR 仍被合并。

- 回退后蒸馏输出回归的风险 (correctness): 评论未获得回应或 resolved，PR 仍被合并，表明维护者接受此风险暂缓处理。

风险与影响

- 风险：回归风险：回退后 `use_remove_padding=False` 路径下的蒸馏输出再次缺失，违反 #6293 的修复目标，使用 `distillation_use_topk=True` 且未启用 `use_remove_padding` 的用户将遭遇 `KeyError`。兼容性风险：若已有迁移至原修复的代码，回退可能导致行为不一致。测试缺口：对应的回归测试被完全删除，该路径未来再出现断裂时不易察觉。
- 影响：用户影响：依赖蒸馏功能且使用 `NO_PADDING` 模式 (`use_remove_padding=False`) 的用户将无法正常使用，loss 计算会抛出异常。系统影响：无性能退化或安全风险。团队影响：该决策表明团队倾向于在出现不确定性时快速回退以保证主干稳定，并计划后续用更完善的方案重新实现。
- 风险标记：回归蒸馏输出缺失，测试覆盖删除，未解决评论

关联脉络

- PR #6350 [fsdp] fix: emit distillation outputs in `use_remove_padding=False` path (#6293): 被回退的 PR，本 PR 是其内容反转。
- PR #6293 distillation top-k `KeyError` when `use_remove_padding=False`: 原 issue，描述蒸馏输出缺失的 bug，被 #6350 修复，现被本 PR 重新引入。