PR #38680 完整报告

vllm-project/vllm

[CI][ROCm] Remove unsupported cases in test_fusion.py

合并时间：2026-05-15 05:37

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/38680

执行摘要

- 一句话：移除 ROCm 不支持的 FP8 测试用例并修复 normalize 条件
- 推荐动作：此 PR 为维护性清理，不值得精读。但可以关注 ROCm FP8 AITER 的支持边界以及 fp8_utils 中 normalize 条件的改进思路。

功能与动机

ROCm 上 AITER 仅支持 group 量化，不支持 per-tensor 量化融合；另外 test_fuse_act_padding 存在已知精度问题（见 ROCm/aiter#2614）。

实现拆解

1. 在 vllm/model_executor/layers/quantization/utils/fp8_utils.py 中的三个处理函数（process_fp8_weight_tensor_strategy、process_fp8_weight_channel_strategy、process_fp8_weight_block_strategy）的条件判断中增加 weight.dtype == torch.float8_e4m3fn 检查，确保只有在权重类型为 float8_e4m3fn 时才调用 normalize_e4m3fn_to_e4m3fnuz，避免重复或错误转换。
2. 在 tests/compile/passes/test_fuse_act_padding.py 中，将 @pytest.mark.skipif 替换为 @pytest.mark.skip，并关联上游 issue，永久跳过该测试；同时移除 is_aiter_found_and_supported 导入和 outputs_unfused = model(x) 行。
3. 在 tests/compile/passes/test_fusion.py 中，从 AITER_KERNEL_GROUPSHAPE_COMBINATIONS 列表中移除 (ROCmFP8ScaledMMLinearKernel, GroupShape.PER_TENSOR, False) 条目，因为 AITER 不支持 per-tensor 量化融合。

关键文件：

- vllm/model_executor/layers/quantization/utils/fp8_utils.py（模块 量化工具；类别 source；类型 data-contract；符号 process_fp8_weight_tensor_strategy, process_fp8_weight_channel_strategy, process_fp8_weight_block_strategy）：核心源码变更，在三个 FP8 weight 处理函数中添加了类型检查条件，防止不必要的 normalize 操作。
- tests/compile/passes/test_fuse_act_padding.py（模块 融合填充测试；类别 test；类型 test-coverage）：由于已知精度问题，整个测试被永久跳过；同时清理了相关导入和未使用代码。

- tests/compile/passes/test_fusion.py（模块 FP8 融合测试；类别 test；类型 test-coverage）：移除 AITER 不支持的 per-tensor 量化融合测试组合

关键符号：process_fp8_weight_tensor_strategy, process_fp8_weight_channel_strategy, process_fp8_weight_block_strategy

关键源码片段

vllm/model_executor/layers/quantization/utils/fp8_utils.py

核心源码变更，在三个 FP8 weight 处理函数中添加了类型检查条件，防止不必要的 normalize 操作。

```
def process_fp8_weight_tensor_strategy(
    weight: torch.Tensor,
    weight_scale: torch.Tensor,
    logical_widths: list[int],
    input_scale: torch.Tensor | None = None,
) -> tuple[torch.Tensor, torch.Tensor, torch.Tensor | None]:
    """Process weights for tensor-wise quantization strategy."""
    from vllm.model_executor.layers.quantization.utils.w8a8_utils import (
        normalize_e4m3fn_to_e4m3fnuz,
        requantize_with_max_scale,
    )

    # 仅当平台是 FP8 fnuz 且权重类型为 float8_e4m3fn 时，
    # 才执行 normalize 操作，避免重复转换。
    if current_platform.is_fp8_fnuz() and weight.dtype == torch.float8_e4m3fn:
        weight, weight_scale, input_scale = normalize_e4m3fn_to_e4m3fnuz(
            weight=weight, weight_scale=weight_scale, input_scale=input_scale
        )

    # Requantize with max scale
    weight_scale, weight = requantize_with_max_scale(
        weight=weight,
        weight_scale=weight_scale,
        logical_widths=logical_widths,
    )

    weight = _maybe_pad_fp8_weight(weight)
    return weight, weight_scale, input_scale
```

评论区精华

review 中无实质争议。维护者 yewentao256 与 AndreasKaratzas 均批准更改。yewentao256 表示问题已在主分支修复。charlifu 在评论中说明因精度问题跳过了 test_fuse_act_padding，并引用了上游 issue。

- 暂无高价值评论线程

风险与影响

- 风险：主要风险是测试覆盖减少：test_fuse_act_padding 被完全跳过，可能导致未来相关融合 pass 的回归未被捕获。此外，fp8_utils.py 的条件增强虽降低了错误转换风险，但可能掩藏其他类型权重的问题。整体风险低。
- 影响：影响范围限定于 ROCm 平台的 FP8 量化功能。测试变更使 AITER 测试更准确反映实际支持情况；源码修改避免了不必要的类型转换，对 CUDA 平台无影响。团队需注意测试覆盖率下降，并跟踪上游 issue 的修复进展。
- 风险标记：测试覆盖减少，上游 issue 依赖

关联脉络

- 暂无明显关联 PR