

PR #46160 完整报告

vllm-project/vllm

[CI][ROCm] Skip unsupported test cases on ROCm

合并时间: 2026-06-24 00:35

原文链接: <http://prhub.com.cn/vllm-project/vllm/pull/46160>

执行摘要

- 一句话: 跳过 ROCm 不支持的 NVFP4 测试用例
- 推荐动作: 值得合并, 修复 CI 噪声。保持后续关注, 待 ROCm 支持 NVFP4 后应移除跳过逻辑。

功能与动机

Buildkite CI 显示 `test_compressed_tensors_nvfp4[args0]` 和 `test_modelopt_mixed_precision_dispatches_w4a16_layer[W4A16_NVFP4-...]` 在 ROCm 上运行失败, 因为这些混合精度用例在 ROCm 上不支持。PR 目的是暂时跳过这些测试, 以稳定 CI。

实现拆解

1. `test_compressed_tensors.py`: 引入 `contextmanager` 和 `current_platform` (已有), 新增 `_nvfp4_marlin_error_context` 上下文管理器, 在 ROCm 且模型为 NVFP4A16 时捕获 `RuntimeError` 并验证错误消息; 否则正常执行。修改 `test_compressed_tensors_nvfp4` 函数, 使其接收 `capfd` 参数, 并在 `with` 语句中嵌套该上下文管理器。
2. `test_modelopt.py`: 导入 `current_platform`。在 `test_modelopt_mixed_precision_dispatches_w4a16_layer` 函数中, 当预期 `LinearMethod` 为 `ModelOptNvFp4W4A16LinearMethod` 且平台是 ROCm 时, 直接 `pytest.skip`。

关键文件:

- `tests/quantization/test_compressed_tensors.py` (模块 量化测试; 类别 test; 类型 test-coverage; 符号 `_nvfp4_marlin_error_context`, `test_compressed_tensors_nvfp4`): 核心变更文件, 新增 NVFP4 错误上下文管理器并调整测试函数。
- `tests/quantization/test_modelopt.py` (模块 量化测试; 类别 test; 类型 test-coverage): 次要变更, 添加 ROCm 跳过逻辑。

关键符号: `_nvfp4_marlin_error_context`, `test_compressed_tensors_nvfp4`

关键源码片段

`tests/quantization/test_compressed_tensors.py`

核心变更文件，新增 NVFP4 错误上下文管理器并调整测试函数。

```
# 新增的上下文管理器，用于在 ROCm 上正确处理 NVFP4A16 测试
@contextmanager
def _nvfp4_marlin_error_context(model, capfd):
    # 判断是否为 ROCm 且模型为 NVFP4A16（该组合不受支持）
    is_rocm_and_unsupported = (
        model == "nm-testing/TinyLlama-1.1B-Chat-v1.0-NVFP4A16"
        and current_platform.is_rocm()
    )

    if is_rocm_and_unsupported:
        # 期望的错误消息
        expected_error = (
            "ValueError: Forced NVFP4 kernel MarlinNvFp4LinearKernel is not "
            "supported: Marlin FP4 not available"
        )
        # 断言引擎初始化失败，并捕获输出
        with pytest.raises(RuntimeError, match="Engine core initialization failed"):
            yield

        captured = capfd.readouterr()
        assert expected_error in captured.out + captured.err
    else:
        # 其他情况（CUDA 或 NVFP4）正常执行
        yield

# 修改后的测试函数，使用上下文管理器
@pytest.mark.parametrize(
    "args",
    [
        ("nm-testing/TinyLlama-1.1B-Chat-v1.0-NVFP4A16", True),
        ("nm-testing/TinyLlama-1.1B-Chat-v1.0-NVFP4", False),
    ],
)
def test_compressed_tensors_nvfp4(vllm_runner, args, capfd):
    model, use_a16 = args
    with (
        _nvfp4_marlin_error_context(model, capfd),
        vllm_runner(model, enforce_eager=True) as llm,
    ):
        def check_model(model):
            layer = model.model.layers[0]
            qkv_proj = layer.self_attn.qkv_proj
            assert isinstance(qkv_proj.quant_method, CompressedTensorsLinearMethod)
            assert isinstance(qkv_proj.scheme, CompressedTensorsW4A4Fp4)
            assert qkv_proj.scheme.use_a16 == use_a16
            assert qkv_proj.scheme.group_size == 16
```

```
llm.apply_model(check_model)
output = llm.generate_greedy(["Hello my name is"], max_tokens=4)
assert output
```

评论区精华

AndreasKaratzas 建议将条件语句改为更可读的形式（如 `is_unsupported_nvfp4`），作者采纳并修改。

- 代码可读性建议 (style): 作者采纳并修改了代码，使用 `is_rocm_and_unsupported` 变量提高可读性。

风险与影响

- 风险：低风险。变更仅限于测试文件，使用 `pytest.skip` 和上下文管理器，不影响生产代码。但需要注意，跳过测试意味着 ROCm 上的 NVFP4 功能不会被测试覆盖，未来如果支持了需移除跳过逻辑。
- 影响：影响范围小：仅影响 ROCm CI 的量化测试组，使两个测试在 ROCm 上不再失败，而 CUDA 平台行为不变。
- 风险标记：跳过测试可能遗漏未来回归

关联脉络

- 暂无明显关联 PR