

PR #53589 完整报告

vllm-project/vllm

[ROCm][CI] Skip ModernBERT FP8 MTEB test when no FP8 ScaledMM kernel exists

合并时间: 2026-08-25 16:07

原文链接: <http://prhub.com.cn/vllm-project/vllm/pull/53589>

执行摘要

- 一句话: 跳过无 ScaledMM 内核的 ModernBERT FP8 测试, 并将 MTEB 任务迁移到 MI300
- 推荐动作: 值得关注此 PR 中的跳过逻辑实现, 特别是如何复用现有 `is_quant_method_supported` 并增加平台特定检查。

功能与动机

在 ROCm 平台 (特别是 MI250 gfx90a) 上, 运行 ModernBERT FP8 测试会抛出 `ValueError: Failed to find a kernel that can implement the ScaledMM linear layer`。这是因为 `is_quant_method_supported("fp8")` 在 gfx90a 上返回 True, 但实际 per-tensor FP8 linear 层要求 CDNA3+ (gfx942/gfx950) 或 RDNA4 (gfx12x)。该 PR 旨在让测试跳过而不是失败, 并确保 CI 在支持硬件的硬件上运行相关测试。

实现拆解

1. 修改测试跳过逻辑: 在 `tests/models/language/pooling_mteb_test/test_modernbert_fp8.py` 中, 新增 `_fp8_scaled_mm_unsupported()` 函数, 它首先复用 `is_quant_method_supported("fp8")` 检查, 然后额外判断 ROCm 平台是否具有相应的 ScaledMM 内核 (通过 `get_cdna_version()` 和 `on_gfx12x()`), 只有明确支持时才继续运行测试。
2. 调整 CI 任务: 将 `.buildkite/test-amd.yaml` 中的 Language Models (MTEB) 任务从 MI250 迁移到 MI300, 这样整个任务组都运行在支持 ScaledMM 内核的硬件上 (MI300 为 CDNA3 架构)。
3. 测试验证: PR 中提供了修改前后的测试结果对比, 修改后测试在 MI250 上正确跳过, 且运行时间从 10.48s 缩短至 2.43s。

关键文件:

- `tests/models/language/pooling_mteb_test/test_modernbert_fp8.py` (模块 测试; 类别 test; 类型 test-coverage; 符号 `_fp8_scaled_mm_unsupported`): 修改了测试的跳过条件, 使其在 ROCm 平台上更精确地判断是否支持 ScaledMM 内核。
- `.buildkite/test-amd.yaml` (模块 CI 配置; 类别 config; 类型 configuration): 将 Language Models (MTEB) CI 任务从 MI250 迁移到 MI300, 确保在支持 ScaledMM 内核的硬件上运行。

关键符号: `_fp8_scaled_mm_unsupported`

关键源码片段

[tests/models/language/pooling_mteb_test/test_modernbert_fp8.py](#)

修改了测试的跳过条件，使其在 ROCm 平台上更精确地判断是否支持 ScaledMM 内核。

```
# tests/models/language/pooling_mteb_test/test_modernbert_fp8.py

import pytest

from tests.models.utils import EmbedModelInfo
from tests.quantization.utils import is_quant_method_supported
from vllm.platforms import current_platform

from .mteb_embed_utils import mteb_test_embed_models

def _fp8_scaled_mm_unsupported() -> bool:
    """判断 ModernBERT 在线 FP8 测试是否应被跳过。

    ``is_quant_method_supported("fp8")`` 在 MI250 (gfx90a) 上返回 True,
    但在 ROCm 上运行 per-tensor FP8 linear 层需要 CDNA3+ 或 RDNA4,
    因此这里将其视为不受支持。
    """
    # 如果 fp8 本身不受支持，则跳过
    if not is_quant_method_supported("fp8"):
        return True
    # 在 ROCm 平台上，需要检查是否有 ScaledMM 内核支持
    if current_platform.is_rocm():
        from vllm.platforms.rocm import get_cdna_version, on_gfx12x
        # 若 CDNA 版本 <= 2 且不是 gfx12x，则说明没有所需内核，跳过
        return get_cdna_version() <= 2 and not on_gfx12x()
    # 非 ROCm 平台，正常执行
    return False

MODEL_INFO = EmbedModelInfo(
    "Alibaba-NLP/gte-modernbert-base",
    mteb_score=0.748193353,
    architecture="ModernBertModel",
    seq_pooling_type="CLS",
    attn_type="encoder_only",
    is_prefix_caching_supported=False,
    is_chunked_prefill_supported=False,
    enable_test=True,
)

@pytest.mark.skipif(
```

```

    _fp8_scaled_mm_unsupported(),
    reason="No FP8 ScaledMM kernel is available on this GPU type.",
)
def test_modernbert_online_fp8_mteb(hf_runner, vllm_runner, monkeypatch) -> None:
    monkeypatch.setenv("VLLM_ALLOW_INSECURE_SERIALIZATION", "1")

def check_online_fp8(vllm_model) -> None:
    vllm_model.apply_model(_assert_modernbert_online_fp8)

mteb_test_embed_models(
    hf_runner,
    vllm_runner,
    MODEL_INFO,
    vllm_extra_kwargs={"quantization": "fp8_per_tensor"},
    # Account for Online FP8 kernel variance across GPU architectures.
    atol=2e-3,
    vllm_model_callback=check_online_fp8,
)

```

评论区精华

由于 PR 来自 fork，自动 review 被禁用，且没有任何代码评审评论。除了 CI 触发之外，讨论很有限。

- 暂无高价值评论线程

风险与影响

- 风险：主要风险在于 MI300 上的运行结果可能与 MI250 存在差异，可能引入新的测试失败。但该 PR 包含了容差设置（ $\text{atol}=2e-3$ ），并且由于 MI300 是支持内核的硬件，预计运行会成功。
- 影响：用户影响：修复了 CI 测试失败，避免了误报。系统影响：将 MTEB 任务从 MI250 迁移到 MI300，可能会增加 MI300 的负载，但 MI250 的测试覆盖减少。团队影响：确保在支持硬件上运行相关测试，减少排查无谓失败的时间。
- 风险标记：CI 迁移可能引入 MI300 上的新失败，测试跳过逻辑可能过于严格或宽松

关联脉络

- PR #53530 Exclude the cpu backend from vLLM's active-Triton-driver count (#53530): 同样涉及 ROCm 平台的支持判断，都可能影响 CI 测试的跳过逻辑。
- PR #53646 [CI][The Rock] Increase flex attention abs tol: 同样涉及 ROCm CI 测试的调整，可能共享类似的 CI 配置或测试跳过策略。