

PR #45536 完整报告

vllm-project/vllm

Fix docs build on `main`

合并时间: 2026-06-13 23:53

原文链接: <http://prhub.com.cn/vllm-project/vllm/pull/45536>

执行摘要

- 一句话: 修复 docs 构建失败问题
- 推荐动作: 建议快速合并, 属于基础设施修复, 无需深入审查。

功能与动机

由于已停止在每个 PR 上构建文档, 需要在主分支上定期修复文档构建问题。

实现拆解

1. 修复 docstring 格式: 在 `vllm/model_executor/layers/fused_moe/experts/cpu_moe.py` 中将返回类型描述中的多行元组格式合并为一行, 以符合文档构建解析要求。
2. 添加空文件: 新增 `vllm/model_executor/layers/fusion/__init__.py` 文件 (空内容), 修复模块导入时的文档生成问题。

关键文件:

- `vllm/model_executor/layers/fused_moe/experts/cpu_moe.py` (模块 模型执行器; 类别 source; 类型 data-contract; 符号 `prepare_int4_moe_layer_for_cpu`): 修复了 `prepare_int4_moe_layer_for_cpu` 函数的 docstring 格式, 将返回元组从多行改为单行, 消除了文档构建的解析错误。
- `vllm/model_executor/layers/fusion/__init__.py` (模块 模型执行器; 类别 source; 类型 data-contract): 新建空文件, 修复 `vllm.model_executor.layers.fusion` 模块的文档生成, 避免因缺少 `__init__.py` 导致导入错误。

关键符号: `prepare_int4_moe_layer_for_cpu`

关键源码片段

`vllm/model_executor/layers/fused_moe/experts/cpu_moe.py`

修复了 `prepare_int4_moe_layer_for_cpu` 函数的 docstring 格式, 将返回元组从多行改为单行, 消除了文档构建的解析错误。

Returns:

```
(blocked_w13, blocked_w2, blocked_s13, blocked_s2, blocked_z13, blocked_z2)
"""
E = w13_packed.size(0)
```

```

# No qzeros are available in compressed-tensors symmetric checkpoints.
# The GPTQ unpack kernel (unpack_4bit_to_32bit_signed) adds +1 to stored zeros,
# so we store 7 per nibble: 0x77777777 -> +1 -> 8.
if w13_zeros is None:
    num_groups_w13 = w13_scale.size(1)
    N_w13 = w13_scale.size(2) # 2*I
    _zp = 0x77777777
    w13_zeros = torch.full(
        (E, num_groups_w13, N_w13 // 8),
        _zp,
        dtype=torch.int32,
    )

if w2_zeros is None:
    num_groups_w2 = w2_scale.size(1)
    N_w2 = w2_scale.size(2) # K
    _zp = 0x77777777
    w2_zeros = torch.full(
        (E, num_groups_w2, N_w2 // 8),
        _zp,
        dtype=torch.int32,
    )

blocked_w13, blocked_z13, blocked_s13 = convert_weight_packed_scale_zp(
    w13_packed, w13_zeros, w13_scale, quant_algo
)
blocked_w2, blocked_z2, blocked_s2 = convert_weight_packed_scale_zp(
    w2_packed, w2_zeros, w2_scale, quant_algo
)

```

评论区精华

无 review 评论。

- 暂无高价值评论线程

风险与影响

- 风险：风险极低：变更仅为 docstring 格式调整和新增空文件，不影响运行逻辑。
- 影响：直接影响：修复了主分支的文档构建管道，使开发者文档保持可用。间接影响：防止新 PR 因文档构建失败而被阻塞。
- 风险标记：暂无

关联脉络

- PR #44260 Add the QuantizedActivation linear-kernel contract: 可能引入了 `vllm/model_executor/layers/fusion/__init__.py` 模块，但缺少该文件导致文档构建失败。