

PR #47607 完整报告

vllm-project/vllm

[bugfix] fix MOSS-Audio deepstack_input_embeds initialization in PP

合并时间: 2026-07-06 17:15

原文链接: <http://prhub.com.cn/vllm-project/vllm/pull/47607>

执行摘要

- 一句话: 修复 MOSS-Audio PP 下 deepstack_input_embeds 初始化缺失
- 推荐动作: 值得合入, 改动虽小但正确性关键, 修复了 PP 下的崩溃。建议同步检查其他是否有其他类似初始化未分片的模型, 可作为学习 PP 模式下 intermediate_tensors 分片处理的范例。

功能与动机

PR #44443 将 MOSS-Audio 切换到 ModelRunnerV2 后, 在 PP 模式下出现 KeyError 崩溃 (具体见 PR body 中的错误栈)。根本原因是 deepstack_input_embeds 的初始化未按 PP 分片裁剪, 非首 rank 创建了过多 key, 导致后续访问时找不到预期的 key。

实现拆解

1. 在 vllm/model_executor/models/moss_audio.py 的 make_empty_intermediate_tensors 方法中, 遍历 self.deepstack_inject_layer_indices 时增加判断: if layer_idx < self.model.start_layer: continue。
2. 这样非首 PP rank 只会为其负责的层创建 deepstack_input_embeds_{layer_idx} 张量, 与 intermediate_tensors 的分片逻辑对齐。
3. 无其他文件修改, 改动仅 4 行, 不涉及测试或配置变更。

关键文件:

- vllm/model_executor/models/moss_audio.py (模块 模型实现; 类别 source; 类型 data-contract; 符号 make_empty_intermediate_tensors): 唯一修改的文件, 在 make_empty_intermediate_tensors 中增加 start_layer 判断以修复 PP 下 KeyError。

关键符号: make_empty_intermediate_tensors

关键源码片段

[vllm/model_executor/models/moss_audio.py](#)

唯一修改的文件, 在 make_empty_intermediate_tensors 中增加 start_layer 判断以修复 PP 下 KeyError。

```
# vllm/model_executor/models/moss_audio.py
def make_empty_intermediate_tensors(
```

```

self,
batch_size: int,
dtype: torch.dtype,
device: torch.device,
) -> IntermediateTensors:
    intermediate_tensors = self.model.make_empty_intermediate_tensors(
        batch_size, dtype, device
    )
    for layer_idx in self.deepstack_inject_layer_indices:
        # Non-first PP ranks only receive DeepStack payloads for layers
        # at or after their local start layer.
        if layer_idx < self.model.start_layer:
            continue
        intermediate_tensors[f"deepstack_input_embeds_{layer_idx}"] = torch.zeros(
            (batch_size, self.config.hidden_size),
            dtype=dtype,
            device=device,
        )
    return intermediate_tensors

```

评论区精华

没有实质性 review 讨论，仅有一个 claude[bot] 的自动评论提示不可见，以及三位 reviewer 的 approve。

- 暂无高价值评论线程

风险与影响

- 风险：风险极低。改动仅增加一个条件判断跳过无关层，且该逻辑是其他模型已有的标准做法（如 deepseek 系列）。回归风险限于 MOSS-Audio 的 PP 执行路径，单一文件内单一方法。
- 影响：影响范围仅限 MOSS-Audio 模型在 PP 模式下的运行。修复后非首 rank 不再因 KeyError 崩溃，模型可正常推理。对于非 PP 用户无影响。
- 风险标记：核心路径变更

关联脉络

- PR #44443 [Core] Use ModelRunnerV2 as default for all dense models: 引入本 bug 的根因 PR，切换了默认 model runner 导致 MOSS-Audio 的 PP 初始化分支触发崩溃。