

PR #54847 完整报告

vllm-project/vllm

[Bugfix] Fix ColQwen3.5 pooler projector initialization

合并时间: 2026-09-02 10:35

原文链接: <http://prhub.com.cn/vllm-project/vllm/pull/54847>

执行摘要

- 一句话: 修复 ColQwen3.5 池化器投影器初始化错误, 解决权重加载失败和形状不匹配问题。
- 推荐动作: 建议精读此 PR, 因为它展示了池化器投影器初始化的正确模式 (匹配 ColPali), 对于理解 vLLM 中多模态模型的池化器集成有参考价值。

功能与动机

PR body 描述 ColQwen3.5 模型加载失败, 出现缺失权重错误 (pooler.head.projector.0.weight/bias) 和运行时形状不匹配, 原因是模型创建了 `custom_text_proj` 但传递 `projector=None` 给 `pooler_for_token_embed()`, 导致 `_load_st_projector()` 创建未初始化投影器, 引发双重投影和维度不兼容。

实现拆解

1. 修改池化器初始化: 在 `vllm/model_executor/models/colqwen3_5.py` 的 `__init__` 方法中, 将 `pooler_for_token_embed()` 的 `projector` 参数从 `None` 改为 `self.custom_text_proj`, 使池化器负责投影, 匹配 ColPali 模式。
2. 简化前向传播: 在 `forward` 方法中移除投影逻辑 (包括类型检查和投影调用), 直接返回 `super().forward()` 的结果, 避免双重投影, 确保返回原始隐藏状态供池化器处理。
3. 调整权重加载: 在 `load_weights` 方法中, 为投影层权重添加额外的加载路径 `pooler.head.projector.*`, 确保权重正确映射到池化器投影层, 同时保持原有 `custom_text_proj.*` 路径。
4. 文档更新: 更新类文档字符串, 明确投影和归一化由池化器处理, 并说明 `custom_text_proj` 传递给池化器。本次变更仅涉及源码文件, 测试通过但未包含测试修改。

关键文件:

- `vllm/model_executor/models/colqwen3_5.py` (模块 模型执行器; 类别 source; 类型 data-contract): 包含 ColQwen3.5 模型的初始化、前向传播和权重加载逻辑, 是本次修复的核心文件, 涉及数据契约变更。

关键符号: `init`, `forward`, `load_weights`

关键源码片段

`vllm/model_executor/models/colqwen3_5.py`

包含 ColQwen3.5 模型的初始化、前向传播和权重加载逻辑，是本次修复的核心文件，涉及数据契约变更。

修复后的 `__init__` 和 `forward` 方法关键部分，展示投影器传递和逻辑简化：

```
def __init__(self, *, vllm_config: VllmConfig, prefix: str = ""):
    super().__init__(vllm_config=vllm_config, prefix=prefix)
    # ... 其他初始化代码 ...
    self.custom_text_proj = nn.Linear(
        hidden_size, self.embed_dim, bias=True, dtype=head_dtype,
    )
    nn.init.zeros_(self.custom_text_proj.bias)
    pooler_config = vllm_config.model_config.pooler_config
    assert pooler_config is not None
    self.pooler = pooler_for_token_embed(
        pooler_config, projector=self.custom_text_proj,
        # 修复：传递投影器给池化器，避免未初始化
    )

def forward(
    self, input_ids: torch.Tensor | None, positions: torch.Tensor,
    intermediate_tensors=None, inputs_embeds: torch.Tensor | None = None,
    **kwargs: object,
) -> torch.Tensor:
    """Run forward pass returning hidden states for pooler."""
    # 修复：移除投影逻辑，直接返回隐藏状态，让池化器处理投影和归一化
    return super().forward(
        input_ids=input_ids, positions=positions,
        intermediate_tensors=intermediate_tensors,
        inputs_embeds=inputs_embeds,
        **kwargs,
    )
```

评论区精华

无实质性技术讨论；review 仅有自动化提示和批准评论（njhill: "Thanks!"）。修复基于 ColPali 模式的对齐，未引发争议。

- 暂无高价值评论线程

风险与影响

- 风险：风险较低。变更集中于 ColQwen3_5Model 类的初始化路径，修复了明显的双重投影 bug，测试覆盖相关场景。潜在回归风险在于池化器行为变化，但已通过现有测试验证。
- 影响：影响范围限于 ColQwen3.5 多模态模型，修复后模型能正确加载权重并执行 token 嵌入任务，提升检索 / 重排序功能的可靠性。对用户而言，解决了模型加载失败和推理错误。
- 风险标记：模型加载路径变更，潜在回归

关联脉络

- 暂无明显关联 PR