PR #23922 完整报告

sgl-project/sglang

transformers v5 adapt HFRunner

合并时间：2026-05-19 17:07

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/23922

执行摘要

- 一句话：适配 transformers v5 的 HFRunner 变更
- 推荐动作：该 PR 属于紧急兼容性修复，但 review 中提出的两个问题尚未解决。建议作者确认 transformers v5 中 Qwen2VLForConditionalGeneration 的 vision tower 输出格式，若确实需要 pooler_output 则需调整赋值逻辑；同时避免使用 **kwargs 或明确过滤参数。在修复前不宜合并到 main。

功能与动机

PR body 指出，在 transformers 5.3 下，使用 HFRunner 进行图像嵌入时会触发 AttributeError: 'Qwen2VLForConditionalGeneration' object has no attribute 'visual'。变更旨在修复此兼容性问题。

实现拆解

1. 在 _forward_gme_qwen2_vl 中，将 self.model.embed_tokens 替换为 self.model.model.get_input_embeddings()，以兼容新的模型封装。
2. 将 self.model.visual 替换为 self.model.model.visual，确保通过外层模型访问 vision 子模块。
3. 视觉编码结果添加 .pooler_output，以适配 transformers v5 的返回值格式。
4. 在模型 forward 调用中添加 **kwargs，确保额外参数不会导致错误。
5. 变更仅影响 python/sglang/test/runners.py 文件中的测试辅助函数。

关键文件：

- python/sglang/test/runners.py（模块 测试运行器；类别 test；类型 test-coverage；符号 _forward_gme_qwen2_vl）：唯一变更文件，适配 transformers v5 的模型接口变化，修复 AttributeError。

关键符号：_forward_gme_qwen2_vl

关键源码片段

python/sglang/test/runners.py

唯一变更文件，适配 transformers v5 的模型接口变化，修复 AttributeError。

```
def _forward_gme_qwen2_vl(
    self,
    input_ids: Optional[torch.LongTensor] = None,
    attention_mask: Optional[torch.Tensor] = None,
    position_ids: Optional[torch.LongTensor] = None,
    past_key_values: Optional[List[torch.FloatTensor]] = None,
    inputs_embeds: Optional[torch.FloatTensor] = None,
    pixel_values: Optional[torch.Tensor] = None,
    image_grid_thw: Optional[torch.LongTensor] = None,
    pooling_mask: Optional[torch.LongTensor] = None,
    **kwargs,
) -> torch.Tensor:
    if inputs_embeds is None:
        # transformers v5: embed_tokens 改为通过 get_input_embeddings 获取
        inputs_embeds = self.model.model.get_input_embeddings()(input_ids)
        if pixel_values is not None:
            pixel_values = pixel_values.type(self.model.model.visual.get_dtype())
            # 注意：.pooler_output 可能不是正确的字段，视觉编码器通常返回 last_hidden_state
            image_embeds = self.model.model.visual(
                pixel_values, grid_thw=image_grid_thw
            ).pooler_output.to(inputs_embeds.device)
            image_mask = input_ids == self.model.config.image_token_id
            inputs_embeds[image_mask] = image_embeds
        if attention_mask is not None:
            attention_mask = attention_mask.to(inputs_embeds.device)

    outputs = self.model(
        input_ids=input_ids,
        position_ids=position_ids,
        attention_mask=attention_mask,
        past_key_values=past_key_values,
        output_hidden_states=True,
        return_dict=True,
        inputs_embeds=inputs_embeds,
        image_grid_thw=image_grid_thw,
        **kwargs, # 传递额外参数，但需确保模型 forward 支持
    )

    embeddings = outputs.hidden_states[-1][:, -1]
    embeddings = torch.nn.functional.normalize(embeddings, p=2, dim=1)
    return embeddings.contiguous()
```

评论区精华

review 中 gemini-code-assist[bot] 提出了两个问题：

1. 使用 .pooler_output 可能不正确，因为 Qwen2-VL 的视觉编码器返回 patch embeddings（通常是 last_hidden_state），而 pooler_output 是单个池化向量，赋值给多个 image tokens 会导致 shape 不匹配。

2. 添加 **kwargs 可能引发 TypeError，因为 Qwen2VLForConditionalGeneration.forward 的签名可能不支持未预料的参数。 这两个问题均未得到回应或解决。sglang-npu-bot 批准了 PR。

- 使用 pooler_output 可能导致形状不匹配 (correctness): 未解决。作者未回应，PR 被批准但问题依然存在。
- 传递 **kwargs 可能引发 TypeError (correctness): 未解决。作者未回应，PR 被批准但问题依然存在。

风险与影响

- 风险:
  1. (高).pooler_output 的使用可能导致视觉嵌入形状错误，进而使图像 token 对应的 embeddings 维度不匹配，引起运行时错误或静默错误（如输出异常）。
  2. (中)**kwargs 的传递可能因模型 forward 签名限制而抛出 TypeError，但实际运行时 kwargs 可能为空，暂时安全。
  3. (低)测试仅涉及一个文件，影响范围有限，但若该测试辅助函数被用于 CI 验证，可能导致 CI 失败或误报。 - 影响：影响范围限于 python/sglang/test/runners.py 中的 _forward_gme_qwen2_vl 方法，该方法用于 HFRunner 的 GME Qwen2-VL 模型推理。直接用户不受影响，但依赖该函数的 CI 测试（如 test_gme_qwen_models.py）可能因上述风险而失败。 - 风险标记：测试覆盖调整，配置键调整

关联脉络

- 暂无明显关联 PR