

PR #30400 完整报告

sgl-project/sglang

[Cherry pick to release/v0.5.15] Fix NVILA weight loading

合并时间: 2026-07-08 04:21

原文链接: <http://prhub.com.cn/sgl-project/sglang/pull/30400>

执行摘要

- 一句话: 修复 NVILA 权重加载 `KeyError`
- 推荐动作: 该 PR 为必要性 bugfix, 值得精读以确保类似的前缀映射模式在其他模型中得到处理。设计上采用的软 fallback 方式 (仅在键不存在时重映射) 是合理的最小修复方案。

功能与动机

PR body 指出 `test/registered/eval/test_vlms_mmmu_eval.py` 出现 `KeyError: 'vision_tower.vision_model.embeddings.patch_embedding.bias'`, 由 PR#29393 引入。需要修复权重加载时的键名不匹配问题。

实现拆解

1. 在 `python/sglang/srt/models/nvila.py` 的 `load_weights` 方法中, 在从 `params_dict` 查找参数之前, 增加判断: 若当前名称不在 `params_dict` 中且以 `"vision_tower.vision_model."` 开头, 则移除多余的 `"vision_model."` 层级, 将名称改为 `"vision_tower."` 开头。
2. 在 `python/sglang/srt/models/nvila_lite.py` 的 `load_weights` 方法中应用完全相同的补丁, 保持两个模型处理逻辑一致。
3. 变更仅涉及条件判断和字符串替换, 无其他逻辑改动, 符合最小修复原则。

关键文件:

- `python/sglang/srt/models/nvila.py` (模块 模型加载; 类别 source; 类型 data-contract; 符号 `load_weights`): 核心修复文件, 修改了 NVILA 模型的 `load_weights` 方法, 添加了权重名称重映射逻辑。
- `python/sglang/srt/models/nvila_lite.py` (模块 模型加载; 类别 source; 类型 data-contract; 符号 `load_weights`): 与 `nvila.py` 相同的修复, 保持 NVILA Lite 模型权重加载一致性。

关键符号: `load_weights`

关键源码片段

`python/sglang/srt/models/nvila.py`

核心修复文件, 修改了 NVILA 模型的 `load_weights` 方法, 添加了权重名称重映射逻辑。

```
def load_weights(self, weights: Iterable[tuple[str, Tensor]]) -> None:
```

```

params_dict = dict(self.named_parameters())

for name, loaded_weight in weights:
    if name.startswith("llm."):
        # 语言模型部分直接传递给 llm
        self.llm.load_weights([(name[len("llm.") :], loaded_weight)])
    else:
        # 修复 KeyError: 当权重名称中有 'vision_tower.vision_model.' 前缀但模型参数中没有时
        ,
        # 移除多余的 'vision_model.' 层级 ( 例如 'vision_tower.vision_model.embeddings...' -
        > 'vision_tower.embeddings...' )
        if name not in params_dict and name.startswith(
            "vision_tower.vision_model."
        ):
            name = "vision_tower." + name[len("vision_tower.vision_model.") :]
            param = params_dict[name]
            weight_loader = getattr(
                param, "weight_loader", weight_utils.default_weight_loader
            )
            weight_loader(param, loaded_weight)

```

python/sglang/srt/models/nvila_lite.py

与 nvila.py 相同的修复，保持 NVILA Lite 模型权重加载一致性。

```

def load_weights(self, weights: Iterable[tuple[str, Tensor]]) -> None:
    params_dict = dict(self.named_parameters())

    for name, loaded_weight in weights:
        if name.startswith("llm."):
            self.llm.load_weights([(name[len("llm.") :], loaded_weight)])
        else:
            # 修复 KeyError: 权重名称中存在 vision_tower.vision_model.
            前缀但模型参数中不包含时 ,
            # 移除多余的 vision_model. 层级 ( 与 nvila.py 完全相同 )
            if name not in params_dict and name.startswith(
                "vision_tower.vision_model."
            ):
                name = "vision_tower." + name[len("vision_tower.vision_model.") :]
                param = params_dict[name]
                weight_loader = getattr(
                    param, "weight_loader", weight_utils.default_weight_loader
                )
                weight_loader(param, loaded_weight)

```

评论区精华

审核人 Fridge003 直接批准，无 review 评论。作者 b8zhong 在 issue 评论中确认 NVILA 部分测试已通过，其他失败是独立问题。

- 暂无高价值评论线程

风险与影响

- 风险：该 PR 风险极低。仅增加了条件与字符串替换逻辑，且只在权重键不匹配时才触发，不会影响正常加载路径。但需确保所有以 `vision_tower.vision_model.` 前缀的权重都被正确映射，若存在其他子模块前缀（如 `vision_tower.vision_model.encoder.`）未覆盖，可能出现新的 `KeyError`。
- 影响：影响范围仅限于 NVILA 和 NVILA Lite 模型的权重加载。修复后，MMMU 评估测试中的 `KeyError` 将消失，模型可以正常加载权重。对系统其他模块无影响。
- 风险标记：仅条件覆盖，可能存在其他未映射前缀，已测试通过

关联脉络

- PR #29393 [Model] Support NVILA 1.5: 该 PR 引入了 NVILA 模型支持，但使用了 `vision_tower.vision_model.` 前缀的权重名称，与模型参数名称不匹配，导致当前 PR 修复的 `KeyError`。