PR #26744 完整报告

sgl-project/sglang

[RL] Forward Kimi K2.5 weight hooks to language model

合并时间：2026-05-30 06:08

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/26744

执行摘要

- 一句话：Kimi K2.5 模型权重钩子转发
- 推荐动作：建议精读，这是一个典型的模型包装类设计问题，展示了如何在多模态模型中正确转发内部组件的方法以保持接口统一。值得 RL 训练和模型开发团队关注。

功能与动机

Kimi K2.5 的 RL 权重同步代码操作在顶层模型包装器上，但实际的语言模型权重加载元数据（如 stacked_params_mapping、expert_params_mapping）和权重后处理逻辑（post_load_weights）位于内部 language model 上。若不转发这些方法，融合权重和专家权重将无法正确映射，且权重更新后的准备工作无法执行，导致 RL 训练失败。

实现拆解

1. 在 KimiK25ForConditionalGeneration 类中新增 post_load_weights 方法：当 language_model 不为空时，直接调用内部语言模型的同名方法，用于权重更新后的准备工作（如解包 MLA 压缩张量）。
2. 添加 stacked_params_mapping 和 expert_params_mapping 属性：通过 getattr 安全地返回内部语言模型的对应属性，默认为空列表，方便外层代码访问映射配置。
3. 添加 mutate_weight_preload 和 custom_scale_remap 方法：直接委托给内部语言模型，用于加载权重前的名称变异和自定义缩放重映射。
4. 所有变更集中在 python/sglang/srt/models/kimi_k25.py 文件中，新增 18 行代码，无删除，保持向后兼容。

关键文件：

- python/sglang/srt/models/kimi_k25.py（模块 模型层；类别 source；类型 data-contract；符号 post_load_weights, stacked_params_mapping, expert_params_mapping, mutate_weight_preload）：核心变更文件，在 KimiK25ForConditionalGeneration 中转发 5 个关键方法 / 属性到内部语言模型，支持 RL 权重更新。

关键符号：post_load_weights, mutate_weight_preload, custom_scale_remap, stacked_params_mapping, expert_params_mapping

关键源码片段

python/sglang/srt/models/kimi_k25.py

核心变更文件，在 KimiK25ForConditionalGeneration 中转发 5 个关键方法 / 属性到内部语言模型，支持 RL 权重更新。

```
# python/sglang/srt/models/kimi_k25.py
class KimiK25ForConditionalGeneration:
    # ... 已有代码 ...

    def post_load_weights(self):
        """Forward to inner language model so RL weight sync and
        CUDA graph capture can trigger post-load preparation (e.g., MLA decompression)."""
        if self.language_model is not None:
            self.language_model.post_load_weights()

    @property
    def stacked_params_mapping(self):
        """Provide fused weight mapping metadata for RL weight sync."""
        return getattr(self.language_model, "stacked_params_mapping", [])

    @property
    def expert_params_mapping(self):
        """Provide expert weight mapping metadata for RL weight sync."""
        return getattr(self.language_model, "expert_params_mapping", [])

    def mutate_weight_preload(self, name):
        """Delegate name mutation hook for weight loading."""
        return self.language_model.mutate_weight_preload(name)

    def custom_scale_remap(self, name):
        """Delegate custom scale remapping hook for weight loading."""
        return self.language_model.custom_scale_remap(name)
```

评论区精华

该 PR 未触发 review 讨论或评论。作者 ByronHsu 在 PR body 中解释了动机，并引用了关联 PR #23339 作为类似问题的背景。

- 暂无高价值评论线程

风险与影响

- 风险：风险较低。变更仅在 Kimi K2.5 模型的包装类中新增转发方法，不修改任何已有逻辑。若内部语言模型不支持相应方法，getattr 会安全返回默认值或抛出 AttributeError，但现有所有语言模型都应支持这些接口。需确保 RL 训练流程已正确适配这些转发，否则权重更新仍可能失败。
- 影响：直接影响 Kimi K2.5 模型的 RL 在线权重更新流程，使 RL 训练能够正确执行权重映射和加载后处理。对推理无影响，因为推理不调用这些方法。影响范围限定在 Kimi K2.5 模型及相关 RL 训练代码。

- 风险标记：缺少测试覆盖

关联脉络

- PR #23339 Kimi K2.5 checkpoint-engine update: PR body 提到该 PR 处理了类似问题，也转发了 post_load_weights 到内部语言模型，本 PR 是类似场景的延续。