PR #23525 完整报告

sgl-project/sglang

Upgrade transformers from 5.5.4 to 5.6.0

合并时间：2026-04-27 13:33

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/23525

执行摘要

- 一句话：升级 transformers 5.6.0 并修复权重键映射
- 推荐动作：建议开发者在升级后密切关注多模态模型的权重加载日志，确认无 KeyError。维护者可以考虑为权重键映射编写单元测试（如检查已知 checkpoint 的键转换正确性），防止未来回归。本 PR 的设计决策——在自定义加载器中手动声明键映射——是处理上游 breaking change 的实用模式，值得类似场景借鉴。

功能与动机

Transformers 5.6.0 引入了不兼容变更（huggingface/transformers#44431），移除了 CLIPVisionModel、SiglipVisionModel 等模型中的 vision_model 中间包装，导致 sglang 自定义加载器中的权重键不匹配。PR body 指出：'The real break for us is huggingface/transformers#44431 — "Refactor CLIP-like models": the vision_model intermediate wrapper was removed from every CLIP-like class'，以及 'sglang's LLaVA loader walks params_dict by hand and the Gemma3 VLM uses a sglang-local SiglipVisionModel reimpl, so both bypass the HF path; we have to declare the renaming ourselves.'

实现拆解

1. 依赖版本升级：在 python/pyproject.toml、_cpu、_npu、_xpu、_other 共 5 个配置文件中将 transformers==5.5.4 改为 ==5.6.0。
2. LLaVA 家族权重键映射：在 llava.py、llavavid.py、yivl.py 的 projector_weights 字典中加入 "vision_tower.vision_model.": "vision_tower." 映射项，使上游扁平化的键通过字符串替换还原为模型期望的嵌套结构。
3. Gemma3 VLM 权重键双向映射：在 gemma_3.py 的 Gemma3ForConditionalGeneration 类中添加类属性 param_names_mapping 和 reverse_param_names_mapping，使用正则将 vision_tower.(embeddings|encoder|post_layernorm|head).* 与 vision_tower.vision_model.(...).* 相互映射，实现加载和保存的双向兼容。
4. 测试与 CI：未新增单元测试，但通过多次手动触发 CI 重跑（13 条评论中包含多个 /rerun-* 命令）确保 LLaVA 和 Gemma3 的相关集成测试通过。

关键文件：

- python/sglang/multimodal_gen/runtime/models/encoders/gemma_3.py（模块 多模态编码器；类别 source；类型 data-contract；符号 Gemma3ForConditionalGeneration, param_names_mapping, reverse_param_names_mapping）：核心修复 Gemma3 VLM 的权重键映射，通过类属性实现双向兼容，是整个 PR 中修改最复杂的部分。
- python/sglang/srt/models/llava.py（模块 视觉模型；类别 source；类型 data-contract；符号 projector_weights）：为 LLaVA 模型添加 vision_tower.vision_model 键映射，修复权重加载时的 KeyError。
- python/sglang/srt/models/llavavid.py（模块 视觉模型；类别 source；类型 data-contract；符号 projector_weights）：与 llava.py 相同的键映射修复，针对 LLaVA-Vid 模型。
- python/sglang/srt/models/yivl.py（模块 视觉模型；类别 source；类型 data-contract；符号 projector_weights）：与 llava.py 相同的键映射修复，针对 Yi-VL 模型。
- python/pyproject.toml（模块 依赖配置；类别 config；类型 configuration）：主依赖配置，升级 transformers 版本号，触发整个变更。
- python/pyproject_cpu.toml（模块 依赖配置；类别 config；类型 configuration）：CPU 构建依赖配置，同步升级 transformers 版本。
- python/pyproject_npu.toml（模块 依赖配置；类别 config；类型 configuration）：NPU 构建依赖配置，同步升级 transformers 版本。
- python/pyproject_other.toml（模块 依赖配置；类别 config；类型 configuration）：其他平台构建依赖配置，同步升级 transformers 版本。
- python/pyproject_xpu.toml（模块 依赖配置；类别 config；类型 configuration）：XPU 构建依赖配置，同步升级 transformers 版本。

关键符号：load_weights, Gemma3ForConditionalGeneration

关键源码片段

python/sglang/multimodal_gen/runtime/models/encoders/gemma_3.py

核心修复 Gemma3 VLM 的权重键映射，通过类属性实现双向兼容，是整个 PR 中修改最复杂的部分。

```
# gemma_3.py: 适配 transformers 5.6.0 的 SiglipVisionModel 键映射
class Gemma3ForConditionalGeneration(nn.Module):
    # transformers 5.6.0 扁平化了 SiglipVisionModel，移除了 vision_model 中间包装。
    # 我们的 reimpl 保留包装，所以需要将 HF 上游的扁平键映射回嵌套结构。
    param_names_mapping = {
        r"^(vision_tower\.)(embeddings|encoder|post_layernorm|head)\.": r"\1vision_model.\2.",
    }
    reverse_param_names_mapping = {
        r"^(vision_tower\.)vision_model\.(embeddings|encoder|post_layernorm|head)\.": r"\1\2.",
    }
```

python/sglang/srt/models/llava.py

为 LLaVA 模型添加 vision_tower.vision_model 键映射，修复权重加载时的 KeyError。

```
# llava.py: 适应 transformers 5.6.0 的权重键映射
```

```
projector_weights = {
    "model.mm_projector.0": "multi_modal_projector.linear_1",
    "model.mm_projector.2": "multi_modal_projector.linear_2",
    "model.vision_tower.vision_tower": "vision_tower",
    # transformers 5.6.0 移除了 vision_model 中间层，
    # 所以将上游的 vision_tower.vision_model.* 映射到 vision_tower.*
    "vision_tower.vision_model.": "vision_tower.",
    "model.image_newline": "language_model.model.image_newline",
}
```

评论区精华

PR body 中详细分析了上游变更的上下文，指出 release notes 只标记了 rotary_fn，实际 break 来自 CLIP 重构。贡献者 JustinTong0323 和审核者 Kangyan-Zhou 通过多次 CI 重跑确保稳定性，没有其他实质性的 review 讨论。

- 暂无高价值评论线程

风险与影响

- 风险：
  1. 上游潜在不兼容：Transformers 5.6.0 除 rotary_fn 和 CLIP 重构外，可能还存在其他未发现的 breaking changes，虽经 grep 确认无影响，但边缘路径可能暴露。
  2. 字符串替换脆弱性：LLaVA 家族的映射采用简单的 str.replace，如果未来 HF 进一步调整命名格式（如嵌套深度变化），可能再次失效。
  3. 测试覆盖不足：新增的键映射逻辑缺乏专门的单元测试，仅依赖 CI 集成测试，无法覆盖所有可能的 checkpoint 变体。
  4. 非全量覆盖：AltCLIP、XCLIP 等变体未处理，若未来启用可能遇到相同问题。 - 影响：影响所有使用 transformers 5.6.0 的部署，尤其是涉及多模态模型（LLaVA、Yi-VL、Gemma3）的用户。由于依赖版本被 pin 死，任何新安装或重建都会自动应用新版本。正向影响：获得 transformers 5.6.0 的 bugfix 和新功能。负面影响：若用户手动降级至 5.5.4 或使用其他构建方式，权重加载可能因键不匹配而失败。影响程度中等，变更集中且已通过 CI 验证。 - 风险标记：上游依赖变更，权重映射回归风险，缺少测试覆盖

关联脉络

- 暂无明显关联 PR