

PR #38447 完整报告

vllm-project/vllm

[Transformers v5] Vendor HCXVisionConfig for compatibility

合并时间: 2026-05-04 12:19

原文链接: <http://prhub.com.cn/vllm-project/vllm/pull/38447>

执行摘要

- 一句话: 自带 HCXVisionConfig 修复 Transformers v5 兼容性
- 推荐动作: 推荐阅读。该 PR 展示了如何通过自带 (vendor) 上游配置类来修复 Transformers v5 兼容性问题, 并涉及 AutoConfig 注册优先级的协作 (关联 Transformers PR #45093), 对理解 vLLM 的配置加载机制以及跨仓库兼容性修复有参考价值。

功能与动机

Transformers v5 引入的 `@strict` 配置验证要求所有属性在初始化时明确设置。上游远程的 `HCXVisionConfig` 在 `__init__` 中仅在 `text_config is not None` 时才设置 `self.text_config`, 当 Transformers v5 创建空配置进行验证或比较时, `get_text_config()` 因访问不存在属性而抛出 `AttributeError`。本 PR 通过自带 (vendor) 修复的配置类解决此问题, 详见关联 Issue #38387 和 #45093。

实现拆解

分四步实现:

1. 在 `vllm/transformers_utils/configs/hyperclovax.py` 中新增 `HCXVisionConfig` 类: 继承自 `PretrainedConfig`, 固定设置 `model_type = "hyperclovax_vlm"`。 `__init__` 方法接受 `text_config` 和 `vision_config` 参数, 先对 `text_config` 中的属性名进行映射 (如 `n_embd`→`hidden_size`), 然后始终将 `self.text_config` 初始化为 `None`, 若传入非空则通过 `AutoConfig.for_model` 生成实际配置对象。 `get_text_config` 方法在 `self.text_config` 为 `None` 时回退返回 `self`, 确保兼容 Transformers v5 的空校验。
2. 在 `vllm/transformers_utils/config.py` 的 `_CONFIG_REGISTRY` 中添加映射: 将 `hyperclovax_vlm` 映射到 `HCXVisionConfig` 类, 确保 vLLM 的配置加载路径能识别该模型类型; 同时优化了注册注释, 说明注册后可被 `AutoTokenizer` 和 `AutoProcessor` 复用。
3. 在 `vllm/transformers_utils/configs/__init__.py` 中注册类: 将 `HCXVisionConfig` 加入 `_CLASS_TO_MODULE` 字典和 `__all__` 导出列表, 使其可通过模块路径访问。
4. 在 `tests/models/registry.py` 中移除版本限制: 删除此前为 `HCXVisionForCausalLM` 设置的 `max_transformers_version` 和版本理由, 因为自带的配置类已兼容 Transformers v5, 同时保留 `trust_remote_code=True` 以保证其他依赖远程代码的模型不受影响。

关键文件:

- vllm/transformers_utils/configs/hyperclova.py (模块 配置层; 类别 source; 类型 core-logic; 符号 HCXVisionConfig, init, get_text_config) : 核心文件, 新增 HCXVisionConfig 类, 修复 text_config 空初始化问题, 是实现兼容性修复的主要修改。
- vllm/transformers_utils/config.py (模块 配置层; 类别 source; 类型 core-logic) : 在 _CONFIG_REGISTRY 中添加 hyperclova_vlm 映射, 使模型类型能被 vLLM 配置加载识别, 同时优化注册注释以说明 AutoTokenizer 复用。
- vllm/transformers_utils/configs/__init__.py (模块 配置层; 类别 source; 类型 core-logic) : 将 HCXVisionConfig 加入模块映射和公开导出列表, 确保可通过 import 访问。
- tests/models/registry.py (模块 测试注册; 类别 test; 类型 test-coverage) : 移除 HCXVisionForCausalLM 的 max_transformers_version 限制, 表明修复后测试可在 Transformers v5 通过。

关键符号: HCXVisionConfig.init, HCXVisionConfig.get_text_config

关键源码片段

vllm/transformers_utils/configs/hyperclova.py

核心文件, 新增 HCXVisionConfig 类, 修复 text_config 空初始化问题, 是实现兼容性修复的主要修改。

文件: vllm/transformers_utils/configs/hyperclova.py

```
from transformers import AutoConfig
from transformers.configuration_utils import PretrainedConfig
```

```
class HCXVisionConfig(PretrainedConfig):
    """自带的 HyperCLOVAX 视觉配置类, 修复 Transformers v5 兼容性。
```

```
    上游远程配置在 `text_config=None` 时不初始化 `self.text_config`,
    导致 Transformers v5 的 `@strict` 验证在 `get_text_config()` 中
    抛出 `AttributeError`。本类始终初始化 `self.text_config = None`,
    并提供安全的回退逻辑。
    """
```

```
    model_type = "hyperclova_vlm"
    keys_to_ignore_at_inference = ["past_key_values"]
```

```
    text_config_attribute_map = {
        "n_embd": "hidden_size",
        "n_positions": "max_position_embeddings",
        "n_head": "num_attention_heads",
        "n_layer": "num_hidden_layers",
    }
```

```
    def __init__(
        self,
```

```

text_config=None,
vision_config=None,
use_nth_layer=-2,
img_start_id=100009,
decoder_max_length=4096,
anyres=False,
unpad=False,
max_num_grids=-1,
num_queries_vis_abstractor=-1,
ignore_index=-100,
proj_pos_emb=True,
proj_prenorm=False,
use_1x1_grid=False,
**kwargs,
):
    # 将 text_config 中的旧属性名映射为新属性名
    for key, val in self.text_config_attribute_map.items():
        if text_config is not None and key in text_config:
            text_config[val] = text_config.pop(key)

    # 始终初始化 text_config 为 None，确保 @strict 验证不崩溃
    self.text_config = None
    if text_config is not None:
        _text_config = AutoConfig.for_model(text_config["model_type"])
        self.text_config = _text_config.from_dict(text_config)
        # 直接从已实例化的对象获取 hidden_size，避免映射键被移除后取不到值
        self.hidden_size = self.text_config.hidden_size

    self.vision_config = None
    if vision_config is not None:
        _vision_config = AutoConfig.for_model(vision_config["model_type"])
        self.vision_config = _vision_config.from_dict(vision_config)

    self.use_nth_layer = use_nth_layer
    self.decoder_max_length = decoder_max_length
    self.anyres = anyres
    self.unpad = unpad
    self.max_num_grids = max_num_grids
    self.num_queries_vis_abstractor = num_queries_vis_abstractor
    self.img_start_id = img_start_id
    self.ignore_index = ignore_index
    self.proj_pos_emb = proj_pos_emb
    self.proj_prenorm = proj_prenorm
    self.use_1x1_grid = use_1x1_grid
    super().__init__(**kwargs)

def get_text_config(self, decoder=False):
    """返回文本子配置；若未设置则回退到视觉配置自身。"""
    if self.text_config is not None:

```

```
        return self.text_config
    return self
```

评论区精华

Gemini Code Assist 机器人在 Review 中指出 `__init__` 中 `self.hidden_size` 的 fallback 逻辑存在缺陷：属性映射循环已从 `text_config` 字典中移除了 `n_embd` 键，后续的 `text_config.get("n_embd")` 将始终返回 `None`，导致 fallback 无法生效。建议直接从已实例化的 `self.text_config.hidden_size` 获取值。该建议被采纳，最终代码中 `self.hidden_size = self.text_config.hidden_size`。

- hidden_size fallback 逻辑缺陷 (correctness): 开发者采纳了建议，最终代码使用 `self.hidden_size = self.text_config.hidden_size`。

风险与影响

- 风险：风险较低。主要风险包括：
 - 未来冲突：若 Transformers 官方将来合并 HyperCLOVAX 配置（见 upstream PR #44956），vLLM 自带的版本可能与官方版本不一致，需及时移除本次修改（PR 已标注 TODO）。
 - 注册覆盖：通过 `AutoConfig.register` 注册自定义配置类可能影响其他依赖同一 `model_type` 的加载路径，但 `exist_ok=True` 确保了覆盖安全性。
 - 测试回归：测试仅验证了模型可以初始化，未验证完整推理流程，可能存在更深层兼容问题。
 - 影响：仅影响使用 `HXCXVisionForCausalLM` 模型（HyperCLOVAX 视觉系列）的用户。该修复使模型能在 Transformers v5 下正常加载，不再受版本限制。对 vLLM 其他模型无影响，但展示了通用的配置 vendoring 模式，可作为后续处理类似上游兼容问题的参考。
- 风险标记：依赖上游修复，配置冲突风险，迁移待办

关联脉络

- 暂无明显关联 PR