PR #39575 完整报告

vllm-project/vllm

Add Jina Embeddings v5 model support (fixes #38633)

合并时间：2026-04-16 14:37

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/39575

执行摘要

本 PR 新增了对 Jina Embeddings v5 模型（jinaai/jina-embeddings-v5-text-small）的支持，通过实现一个基于 Qwen3-0.6B-Base 并合并任务特定 LoRA 适配器的模型类，解决了 Issue #38633 中报告的不支持问题。变更覆盖了模型实现、注册表、测试和文档，使用户能够通过 vLLM 直接部署该模型进行嵌入任务。

功能与动机

Issue #38633 报告用户尝试部署 jinaai/jina-embeddings-v5-text-small 模型时，出现 Model architectures ['JinaEmbeddingsV5Model'] are not supported 错误，并提示需要 peft 包。PR body 明确说明目的是添加该模型支持以修复此问题，使用户能够通过 vLLM 服务该模型，支持检索、文本匹配、分类和聚类四种任务。

实现拆解

1. 新增模型实现：在 vllm/model_executor/models/jina.py 中新增 JinaEmbeddingsV5Model 类，继承自 Qwen3ForCausalLM，并实现 load_weights 方法。添加辅助函数 _load_adapter 和 _build_lora_pairs，用于从 Hugging Face 仓库加载 LoRA 适配器配置和权重，并在加载时合并到基础权重中，避免运行时依赖 peft。关键代码如下：

```
def _load_adapter(
    model: str,
    task: str,
    revision: str | None,
) -> tuple[dict, dict[str, torch.Tensor]] | None:
    """从本地路径或HF仓库加载适配器配置和权重。

    返回 (adapter_config, adapter_weights) 或 None（如果未找到）。
    """
    config_bytes = get_hf_file_bytes(
        f"adapters/{task}/adapter_config.json", # 从HF仓库获取适配器配置文件
        model,
        revision,
    )
    if config_bytes is None:
        return None # 适配器不存在时返回None
```

```
    adapter_config = json.loads(config_bytes) # 解析JSON配置

    weights_bytes = get_hf_file_bytes(
        f"adapters/{task}/adapter_model.safetensors", # 获取适配器权重文件
        model,
        revision,
    )
    if weights_bytes is None:
        return None # 权重文件不存在时返回None

    adapter_weights = safetensors_load(weights_bytes) # 加载安全张量格式的权重
    return adapter_config, adapter_weights
```

1. 注册模型：在 vllm/model_executor/models/registry.py 的 _EMBEDDING_MODELS 字典中添加 "JinaEmbeddingsV5Model": ("jina", "JinaEmbeddingsV5Model") 条目，使系统能够识别该模型架构。
2. 更新测试套件：在 tests/models/language/pooling_mteb_test/test_jina.py 中添加新模型的测试配置，包括 MTEB 分数和架构属性；修改 tests/models/language/pooling_mteb_test/mteb_embed_utils.py，新增 HfMtebEncoder 类并扩展 VllmMtebEncoder 以支持提示前缀。测试配置示例如下：

```
EMBEDDING_MODELS = [
    EmbedModelInfo(
        "jinaai/jina-embeddings-v3",
        mteb_score=0.824413164,
        architecture="XLMRobertaModel",
        is_matryoshka=True,
        seq_pooling_type="MEAN",
        attn_type="encoder_only",
        is_prefix_caching_supported=False,
        is_chunked_prefill_supported=False,
    ),
    EmbedModelInfo( # 新增Jina v5模型测试配置
        "jinaai/jina-embeddings-v5-text-small",
        mteb_score=0.794535707854956,
        architecture="JinaEmbeddingsV5Model",
        seq_pooling_type="LAST",
        attn_type="decoder",
        is_prefix_caching_supported=True,
        is_chunked_prefill_supported=True,
    ),
]
```

1. 配套调整：更新 docs/models/pooling_models/embed.md 文档，添加模型说明；修改 tests/conftest.py 中的 HfRunner 类，添加 revision 参数以支持模型版本控制。

关键源码片段

vllm/model_executor/models/jina.py

核心模型实现文件，新增了 JinaEmbeddingsV5Model 类和适配器加载逻辑，是实现变更的入口。

```
def _load_adapter(
    model: str,
    task: str,
    revision: str | None,
) -> tuple[dict, dict[str, torch.Tensor]] | None:
    """从本地路径或HF仓库加载适配器配置和权重。

    返回 (adapter_config, adapter_weights) 或 None（如果未找到）。
    """
    config_bytes = get_hf_file_bytes(
        f"adapters/{task}/adapter_config.json", # 从HF仓库获取适配器配置文件
        model,
        revision,
    )
    if config_bytes is None:
        return None # 适配器不存在时返回None

    adapter_config = json.loads(config_bytes) # 解析JSON配置

    weights_bytes = get_hf_file_bytes(
        f"adapters/{task}/adapter_model.safetensors", # 获取适配器权重文件
        model,
        revision,
    )
    if weights_bytes is None:
        return None # 权重文件不存在时返回None

    adapter_weights = safetensors_load(weights_bytes) # 加载安全张量格式的权重
    return adapter_config, adapter_weights
```

评论区精华

- 权重加载缺陷：gemini-code-assist[bot] 指出 load_weights 实现可能跳过非 self.model 属性中的权重（如 lm_head），建议使用 super().load_weights 并优化 LoRA 合并方法。作者最终采纳建议，修复了实现。

> gemini-code-assist[bot]: “The load_weights implementation has several critical issues... By calling self.model.load_weights instead of super().load_weights, any weights defined in the JinaEmbeddingsV5Model... are skipped.” - 文件合并优化：noooop 建议将新增的 jina_embeddings_v5.py 和 test_jina_v5.py 合并到现有文件中，以减少文件数量。作者响应并执行了合并，提升了代码结构清晰度。
>
> noooop: “Please merge this file with jina.py. There are already too many files in this folder; having one less file is always better.”

风险与影响

- 技术风险：权重加载逻辑可能遗漏外部权重，导致模型参数不完整；LoRA 适配器合并时可能引入数值精度误差；测试覆盖依赖于预设的 MTEB 分数，模型更新时可能需要调整。
- 影响范围：对用户，现在可以直接部署 Jina Embeddings v5 模型进行嵌入任务；对系统，扩展了模型支持范围；对团队，代码结构优化减少了维护复杂度。

关联脉络

从近期历史 PR 看，本 PR 与 #39747“Update registry for Nemotron-v3 VL Nano/Super”类似，都属于新增模型支持的功能扩展，反映了 vLLM 持续集成新模型的演进趋势。此外，Issue #38633 直接关联，显示用户需求驱动了此变更。