

PR #46313 完整报告

vllm-project/vllm

[Bugfix] Reject matryoshka embedding dimensions above hidden size

合并时间: 2026-06-22 18:16

原文链接: <http://prhub.com.cn/vllm-project/vllm/pull/46313>

执行摘要

- 一句话: 拒绝维度大于隐藏层大小的 Matryoshka 请求
- 推荐动作: 值得精读, 展示了如何为 Matryoshka 嵌入参数添加上界检查, 设计清晰, 测试完善。

功能与动机

对于没有 `matryoshka_dimensions` 列表的 Matryoshka 嵌入模型, `PoolingParams._set_default_parameters` 仅检查 `dimensions >= 1`; 大于隐藏层大小的 `dimensions` 会被静默用于切片 (`[..., :d]`), 返回一个 `hidden_size` 长度的向量而不是拒绝请求。PR 新增上界检查, 参考了 `sglang` 的 `_validate_for_matryoshka_dim`。

实现拆解

1. 新增上界检查: 在 `vllm/pooling_params.py` 的 `_set_default_parameters` 方法中, 当 `mds is None` 且 `dimensions >= 1` 时, 增加 `elif self.dimensions > model_config.embedding_size`: 分支, 抛出 `ValueError`。`embedding_size` 属性是 `review` 中建议使用的, 它优先使用显式 `embedding_size` 并兜底到 `get_hidden_size()`。
2. 添加测试辅助类: 在 `tests/test_pooling_params.py` 新增 `MockMatryoshkaModelConfig` 数据类, 包含 `embedding_size` 字段, 用于模拟没有 `matryoshka_dimensions` 列表的 Matryoshka 模型。
3. 编写边界测试: 新增 `test_embed_dimensions_matryoshka_without_list_upper_bound` 函数, 验证 `dimensions=16` (合法) 正常通过, `dimensions=64` (超过 `embedding_size=32`) 触发 `ValueError`。
4. 无其他配置或部署配套变更。

关键文件:

- `vllm/pooling_params.py` (模块 前端; 类别 `source`; 类型 `core-logic`; 符号 `_set_default_parameters`): 核心逻辑文件: 在 `_set_default_parameters` 方法中新增了 `dimensions` 上界检查, 防止静默截断。
- `tests/test_pooling_params.py` (模块 测试; 类别 `test`; 类型 `test-coverage`; 符号 `MockMatryoshkaModelConfig`, `test_embed_dimensions_matryoshka_without_list_upper_bound`): 测试文件: 新增 `MockMatryoshkaModelConfig` 和 `test_embed_dimensions_matryoshka_without_list_upper_bound` 测试函数, 覆盖上界检

查。

关键符号: `_set_default_parameters`, `test_embed_dimensions_matryoshka_without_list_upper_bound`

关键源码片段

vllm/pooling_params.py

核心逻辑文件: 在 `_set_default_parameters` 方法中新增了 `dimensions` 上界检查, 防止静默截断。

```
# vllm/pooling_params.py _set_default_parameters 方法片段
# 当模型是 Matryoshka 且没有显式 matryoshka_dimensions 列表时,
# 原有代码只检查 dimensions >= 1, 现在额外检查上界。
if mds is not None:
    if self.dimensions not in mds:
        raise ValueError(
            f"Model {model_config.served_model_name!r} "
            f"only supports {str(mds)} matryoshka dimensions, "
            f"use other output dimensions will "
            f"lead to poor results.")
    elif self.dimensions < 1:
        raise ValueError("Dimensions must be greater than 0")
    elif self.dimensions > model_config.embedding_size:
        # 使用 embedding_size (优先显式设置, 否则兜底 get_hidden_size())
        raise ValueError(
            "Dimensions must be less than or equal to the model's "
            f"embedding size ({model_config.embedding_size})")
```

tests/test_pooling_params.py

测试文件: 新增 `MockMatryoshkaModelConfig` 和 `test_embed_dimensions_matryoshka_without_list_upper_bound` 测试函数, 覆盖上界检查。

```
# tests/test_pooling_params.py 新增测试代码
# 模拟没有 matryoshka_dimensions 列表的 Matryoshka 模型
@dataclass()
class MockMatryoshkaModelConfig:
    pooler_config: PoolerConfig
    is_matryoshka: bool = True
    matryoshka_dimensions: list[int] | None = None
    served_model_name: str = "mock-matryoshka-model"
    embedding_size: int = 32 # 模拟 hidden size
```

```
def test_embed_dimensions_matryoshka_without_list_upper_bound():
    task = "embed"
    model_config = MockMatryoshkaModelConfig(
        pooler_config=PoolerConfig(seq_pooling_type="CLS"),
        matryoshka_dimensions=None,
        embedding_size=32,
```

```
)
# 合法维度 16 应通过
PoolingParams(task=task, dimensions=16).verify(model_config)
# 超过 embedding_size 的维度应抛出 ValueError
with pytest.raises(ValueError):
    PoolingParams(task=task, dimensions=64).verify(model_config)
```

评论区精华

reviewer [noooooop](#) 在评论中指出应使用 `model_config.embedding_size` 而非 `model_config.get_hidden_size()`，并引用了 `vllm` 源码中 `embedding_size` 属性的定义（优先使用显式 `embedding_size` 并兜底到 `get_hidden_size()`）。作者 [EazyReal](#) 接受建议并更新了代码和错误消息。

- 建议使用 `embedding_size` 而非 `get_hidden_size (design)`: 作者接受建议，将 `model_config.get_hidden_size()` 替换为 `model_config.embedding_size`，并更新了错误消息和测试 mock。

风险与影响

- 风险：风险极低：变更仅在原有 `elif self.dimensions < 1` 分支后增加了一个额外的 `elif` 条件，不影响已有逻辑路径。测试覆盖了边界情况。回归风险小，但需确保所有 Matryoshka 模型都正确设置了 `embedding_size` 属性（通常从 `get_hidden_size()` 兜底，兼容性良好）。
- 影响：影响范围小：仅影响用户对 Matryoshka 嵌入模型（无显式 `matryoshka_dimensions` 列表）指定超过隐藏层大小的 `dimensions` 参数的情况。此前静默截断的行为现在会被 `ValueError` 拒绝，更早暴露错误，提升用户调试体验。
- 风险标记：无显著风险

关联脉络

- 暂无明显关联 PR