

PR #45308 完整报告

vllm-project/vllm

[Bugfix][Model] Pass revision by name in Run:ai and bitsandbytes index downloads

合并时间: 2026-06-12 11:21

原文链接: <http://prhub.com.cn/vllm-project/vllm/pull/45308>

执行摘要

- 一句话: 修复两个模型加载器中 revision 参数错传到 subfolder 的 bug
- 推荐动作: 建议检查项目中所有调用 `download_safetensors_index_file_from_hf` 的地方, 确保均使用关键字参数传递避免类似错误。PR 修改简洁清晰, 适合作为参数传递最佳实践的参考案例。

功能与动机

PR 描述指出: `RunaiModelStreamerLoader._prepare_weights` 和 `BitsAndBytesModelLoader._prepare_weights` 调用

`download_safetensors_index_file_from_hf` 时将 `revision` 作为第四个位置参数传递。但函数签名为 `(model_name_or_path, index_file, cache_dir, subfolder=None, revision=None)`, 导致 `revision` 被绑定到 `subfolder`, 而真实的 `revision` 参数保持 `None`。当请求非默认 `revision` 时, `safetensors index` 可能从错误的 `ref` 下载。

实现拆解

1. 修改 `vllm/model_executor/model_loader/bitsandbytes_loader.py` 中 `_prepare_weights` 方法对 `download_safetensors_index_file_from_hf` 的调用, 改为关键字参数 `cache_dir=self.load_config.download_dir` 和 `revision=revision`。
2. 同样修改 `vllm/model_executor/model_loader/runai_streamer_loader.py` 中的相同调用。
3. 在 `tests/models/quantization/test_bitsandbytes.py` 中添加 `test_bitsandbytes_passes_revision_by_name` 测试, 使用 `mock` 模拟 `download_safetensors_index_file_from_hf` 并断言 `revision` 关键字参数正确传递。
4. 在 `tests/model_executor/model_loader/runai_streamer_loader/test_runai_model_streamer_loader.py` 中添加 `test_runai_passes_revision_by_name` 测试, 类似验证。

关键文件:

- `vllm/model_executor/model_loader/bitsandbytes_loader.py` (模块 量化加载; 类别 `source`; 类型 `core-logic`; 符号 `_prepare_weights`): `BitsAndBytes` 量化加载器, 修复了 `download_safetensors_index_file_from_hf` 调用参数传递 bug
- `vllm/model_executor/model_loader/runai_streamer_loader.py` (模块 流式加载; 类别 `source`; 类型 `core-logic`; 符号 `_prepare_weights`): `Run:ai` 流式加载器, 修复了相同调用 bug

- tests/models/quantization/test_bitsandbytes.py (模块 量化测试; 类别 test; 类型 test-coverage; 符号 test_bitsandbytes_passes_revision_by_name) : 新增测试验证 bitsandbytes loader 参数传递正确性
- tests/model_executor/model_loader/runai_streamer_loader/test_runai_model_streamer_loader.py (模块 流式测试; 类别 test; 类型 test-coverage; 符号 test_runai_passes_revision_by_name) : 新增测试验证 Run:ai loader 参数传递正确性

关键符号: BitsAndBytesModelLoader._prepare_weights,
RunaiModelStreamerLoader._prepare_weights,
test_bitsandbytes_passes_revision_by_name, test_runai_passes_revision_by_name

关键源码片段

vllm/model_executor/model_loader/bitsandbytes_loader.py

BitsAndBytes 量化加载器, 修复了 download_safetensors_index_file_from_hf 调用参数传递 bug

```
# bitsandbytes_loader.py 中的 _prepare_weights 方法片段
# 关键修改: 将 download_safetensors_index_file_from_hf 的调用改为关键字参数
# 避免 revision 被错误赋给 subfolder 参数
```

```
if use_safetensors:
    if not is_local:
        download_safetensors_index_file_from_hf(
            model_name_or_path,
            index_file,
            cache_dir=self.load_config.download_dir, # 改为关键字参数
            revision=revision, # 明确指定 revision 关键字
        )
    hf_weights_files = filter_duplicate_safetensors_files(
        hf_weights_files, hf_folder, index_file
    )
else:
    hf_weights_files = filter_files_not_needed_for_inference(hf_weights_files)
```

vllm/model_executor/model_loader/runai_streamer_loader.py

Run:ai 流式加载器, 修复了相同调用 bug

```
# runai_streamer_loader.py 中的 _prepare_weights 方法片段
# 修复前: download_safetensors_index_file_from_hf(
# model_name_or_path, index_file, self.load_config.download_dir, revision)
# 导致 revision 被绑定到 subfolder, 生效的 revision 为 None
# 修复后:
```

```
if not is_local and not is_object_storage_path:
    download_safetensors_index_file_from_hf(
        model_name_or_path,
        index_file,
```

```
        cache_dir=self.load_config.download_dir, # 关键字参数
        revision=revision, # 关键字参数
    )
```

评论区精华

无公开 review 讨论。PR 由 simon-mo 直接批准，无评论。

- 暂无高价值评论线程

风险与影响

- 风险：风险极低。修复集中且明确，新增测试覆盖了参数传递路径。潜在风险是其他调用点可能仍有类似问题，但 DefaultModelLoader 已正确使用关键字参数。此外，mock 测试仅验证参数传递，不覆盖真实下载流程，但针对 bug 本身是充分的。
- 影响：影响范围限于使用 Run:ai 流式加载器或 bitsandbytes 量化加载器且指定非默认 revision（如 --revision）的用户。修复后这类用户将能正确下载对应 revision 的 safetensors index，避免加载错误的权重或报错。影响程度中等偏低。
- 风险标记：参数传递陷阱，非默认 revision 路径，模型加载核心路径

关联脉络

- 暂无明显关联 PR