

PR #47697 完整报告

vllm-project/vllm

[fix][run_batch]: respect proxy env vars when downloading media URLs

合并时间: 2026-07-06 20:45

原文链接: <http://prhub.com.cn/vllm-project/vllm/pull/47697>

执行摘要

- 一句话: 修复 run_batch 下载媒体 URL 时不遵守代理环境变量
- 推荐动作: 建议阅读 run_batch.py 中的 download_bytes_from_url 函数变更, 了解 vLLM 如何通过全局 HTTP 连接统一管理代理配置、超时与重试。此 PR 展示了从零散 aiohttp 客户端到集中连接管理的小型重构思路, 对理解 vLLM 网络层设计有参考价值。

功能与动机

aiohttp.ClientSession 默认不读取 HTTP_PROXY/HTTPS_PROXY 环境变量, 导致在需要代理的网络环境(如企业网络)中 download_bytes_from_url 连接超时失败, 即使 curl/wget 等其他工具正常工作。PR 旨在通过添加 trust_env=True 或改用已配置好代理支持的 global_http_connection 来修复此问题。

实现拆解

1. 在文件头部添加 from vllm.connections import global_http_connection 导入。
2. 将 download_bytes_from_url 函数中原本直接创建 aiohttp.ClientSession() 并调用 session.get(url) 的代码替换为一行 return await global_http_connection.async_get_bytes(url, allow_redirects=...)。
3. 移除了原来手动检查 resp.status != 200 的异常抛出逻辑, 因为 global_http_connection 内部已处理错误。
4. 未添加新测试, 但全局连接已在 vLLM 多处使用, 回归风险较低。

关键文件:

- vllm/entrypoints/openai/run_batch.py (模块 入口服务; 类别 source; 类型 dependency-wiring; 符号 download_bytes_from_url): 核心变更文件, 将依赖从直接 aiohttp.ClientSession 替换为 global_http_connection, 修复代理环境兼容性。

关键符号: download_bytes_from_url

关键源码片段

[vllm/entrypoints/openai/run_batch.py](#)

核心变更文件, 将依赖从直接 aiohttp.ClientSession 替换为 global_http_connection, 修复代理环境兼容性。

```

async def download_bytes_from_url(
    url: str,
    allowed_media_domains: Optional[set[str]] = None,
) -> bytes:
    """
    Downloads the content from a given URL and returns it as bytes.
    Supports http, https, and data (base64) schemes.
    """

    parsed = urlparse(url)

    # Handle data URLs (base64 encoded) - not subject to domain restrictions
    if parsed.scheme == "data":
        if "," in url:
            header, data = url.split(",", 1)
            if "base64" in header:
                return base64.b64decode(data)
            raise ValueError(f"Unsupported data URL encoding: {header}")
        raise ValueError(f"Invalid data URL format: {url}")

    # Handle HTTP/HTTPS URLs
    elif parsed.scheme in ("http", "https"):
        if allowed_media_domains is not None:
            url_spec = parse_url(url)
            if url_spec.hostname not in allowed_media_domains:
                raise ValueError(
                    f"The URL must be from one of the allowed domains: "
                    f"{allowed_media_domains}. Input URL domain: "
                    f"{url_spec.hostname}"
                )
            # Use the normalized URL to prevent parsing discrepancies
            # between urllib3 and aiohttp (e.g. backslash-@ attacks).
            url = url_spec.url

        # 使用全局 HTTP 连接，该连接已配置 trust_env = True
        # 以支持 HTTP_PROXY / HTTPS_PROXY 环境变量
        return await global_http_connection.async_get_bytes(
            url, allow_redirects=envs.VLLM_MEDIA_URL_ALLOW_REDIRECTS
        )

    else:
        raise ValueError(
            f"Unsupported URL scheme: {parsed.scheme}. "
            "Supported schemes: http, https, data"
        )

```

评论区精华

jikunshang 在 review 中询问“添加 trust_env=True 是否安全？应该留给用户设置还是通过环境变量？”。DarkLight1337 回应“这与 HTTPConnection.get_async_client 一致，建议改用

global_http_connection”。作者接受建议并更新 PR，从简单添加 trust_env 改为复用全局 HTTP 连接。最终 DarkLight1337 批准合并。

- 使用 trust_env 还是改用 global_http_connection (design): 改用 global_http_connection，因为其已整合代理支持，且复用现有基础设施。

风险与影响

- 风险：低风险。变更仅影响 run_batch 模块中的 URL 下载路径，且使用了 vLLM 中已有大量使用的 global_http_connection，其配置（包括代理支持）已稳定。SSRF 域名白名单检查在调用全局连接之前执行，安全机制未受影响。但缺少专门针对代理场景的测试覆盖，且最终实现依赖全局连接的默认行为（如超时、重试策略），与原实现略有差异，需确认兼容。
- 影响：影响范围：限于使用 vllm run_batch 命令且需要从 HTTP/HTTPS 媒体 URL 下载数据的用户。企业代理网络用户从此能正常工作；无代理用户行为完全一致。影响程度小但关键——修复了一个功能断裂问题，提升了企业环境兼容性。
- 风险标记：缺少测试覆盖，依赖全局连接默认行为

关联脉络

- 暂无明显关联 PR