

PR #41649 完整报告

vllm-project/vllm

Limit gpu utils and lower max BS on test_transcription_api_correctness.py

合并时间: 2026-05-05 01:30

原文链接: <http://prhub.com.cn/vllm-project/vllm/pull/41649>

执行摘要

- 一句话: 降低转录测试内存占用防 OOM
- 推荐动作: 值得快速合并, 以稳定 CI 流水线。显存限制策略可复用为其他类似测试的参考模式。

功能与动机

PR#41478 已尝试降低 batch size 修复 OOM, 但在 18GB MIG H200 的 CI 机器上仍发生 OOM (参见 Buildkite 构建 #64258)。需要更激进的 GPU 显存限制以适配小显存环境。

实现拆解

1. 降低并发序列数: 将 MAX_SEQS_FOR_TRANSCRIPTION_TEST 从 32 改为 8, 以减少同时处理的请求数量, 降低显存峰值。
2. 新增 GPU 显存利用率: 引入 GPU_UTIL_FOR_TRANSCRIPTION_TEST = 0.5 常量和对应的 --gpu_memory_utilization 参数, 限制 vLLM 服务端使用的显存比例, 避免超出 18GB 上限。
3. 传递新增参数: 在 test_wer_correctness 的 server_args 列表中添加 f"--gpu_memory_utilization={GPU_UTIL_FOR_TRANSCRIPTION_TEST}", 确保启动时生效。

关键文件:

- tests/entrypoints/openai/correctness/test_transcription_api_correctness.py (模块 转录测试; 类别 test; 类型 test-coverage): 唯一变更文件, 调整测试资源限制以适配 18GB MIG H200 CI 环境

关键符号: 未识别

关键源码片段

[tests/entrypoints/openai/correctness/test_transcription_api_correctness.py](#)

唯一变更文件, 调整测试资源限制以适配 18GB MIG H200 CI 环境

```
# SPDX-License-Identifier: Apache-2.0
```

```
# ... imports ...
```

```
# Tuned to prevent OOM on 18GB GPUs in transcription correctness tests.
```

```
MAX_SEQS_FOR_TRANSCRIPTION_TEST = 8 # 从 32 降低到 8, 减少并发请求数
GPU_UTIL_FOR_TRANSCRIPTION_TEST = 0.5 # 新增: 限制 GPU 显存利用率为 50%

# ...

def test_wer_correctness(model_config, dataset_repo, n_examples=-1, max_concurrent_request=
None):
    # ...
    server_args = [
        "--enforce-eager",
        f"--tokenizer_mode={model_info.tokenizer_mode}",
        f"--max_num_seqs={MAX_SEQS_FOR_TRANSCRIPTION_TEST}",
        f"--gpu_memory_utilization={GPU_UTIL_FOR_TRANSCRIPTION_TEST}", # 新增参数
    ]
    # ...
```

评论区精华

无人工讨论。gemini-code-assist[bot] 的自动审查确认了变更内容并认为无反馈。
DarkLight1337 直接批准。

- 暂无高价值评论线程

风险与影响

- 风险: 风险极低。仅涉及测试配置调整, 不影响任何生产代码。max_num_seqs=8 可能轻微降低测试压力, 但原定 32 已属宽松; gpu_memory_utilization=0.5 配合 18GB 显存提供了充足余量, 同时仍在合理范围内。
- 影响: 仅影响 tests/entrypoints/openai/correctness/test_transcription_api_correctness.py 这一个测试文件。CI 中学生转录测试的显存压力显著降低, 预期可稳定通过。用户无感知。
- 风险标记: 测试仅微调配置无逻辑变更

关联脉络

- PR #41478 Lower max BS on test_transcription_api_correctness.py: 前驱 PR, 同样尝试降低 batch size 修复 OOM 但不足, 本 PR 进一步收紧限制