PR #37873 完整报告

vllm-project/vllm

[Bugfix] RoBERTa position_id accumulation in CUDA graph padding region

合并时间：2026-03-23 22:59

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/37873

执行摘要

该 PR 修复了 RoBERTa-based embedding 模型在 CUDA 图模式下因位置 ID 在 padding 区域累积导致的服务器崩溃问题。通过修改 gpu_model_runner.py 中的 _preprocess 函数，显式零 padding 区域，避免了越界访问，影响多个模型且提升稳定性，变更虽小但关键。

功能与动机

修复一个崩溃问题：当 CUDA 图启用时，所有 RoBERTa-based 模型（如 BAAI/bge-m3、XLM-RoBERTa）在约 max_position_embeddings / 2 次请求后因位置 ID 累积导致越界断言失败而崩溃。根因是 gpu_model_runner 的 persistent GPU buffer 中 padding 区域未重置，RoBERTa 模型的 replace_roberta_positions 函数进行 in-place 累积操作。引用 PR body："Fix a crash that affected all RoBERTa-based embedding models... when CUDA graphs are enabled."

实现拆解

修改仅涉及一个文件 vllm/v1/worker/gpu_model_runner.py，在 _preprocess 函数中添加以下代码：

```
if num_input_tokens > num_scheduled_tokens:
    self.positions.gpu[num_scheduled_tokens:num_input_tokens].zero_()
```

这确保每次请求中，padding 区域的位置 ID 被重置为零，防止累积。关键点：

- 模块：worker 子系统
- 变更：简单条件检查和零操作
- 目标：隔离 CUDA 图重用与模型特定逻辑

评论区精华

review 中，gemini-code-assist[bot] 和 Isotr0py 均认可修复。但在 Issue 评论中，有设计权衡讨论：

- Isotr0py："But seems #37884 would be a better fix which doesn't touch model runner?"
- yanghui1-arch："#37873 protects against this class of bug for models which use positions += offset inner... It's possible that the same issue will appear again in the future." 最终方案被采纳，强调在 model runner 中修复以预防未来问题。

风险与影响

- 风险：低风险，变更简单且测试覆盖充分，但零 padding 可能影响其他依赖非零 padding 的模型逻辑（尽管未观察到）。
- 影响：影响所有 RoBERTa-based 模型在 CUDA 图模式下的服务，修复后避免崩溃，提升生产环境稳定性。测试显示 10000 次请求全通过，无性能退化。

关联脉络

此 PR 与历史 PR #37884 直接相关，后者也修复相同 bug 但在模型层进行。讨论中对比了两种策略：model runner 修复 vs 模型层修复。关联 Issue #37868 报告了原始崩溃问题。整体揭示 vLLM 在 CUDA 图优化中需谨慎处理 tensor 重用和模型特定逻辑，未来可能需加强类似防护。