

PR #29520 完整报告

sgl-project/sglang

fix: fix prefill-aware SWA floor tracking

合并时间: 2026-06-28 14:18

原文链接: <http://prhub.com.cn/sgl-project/sglang/pull/29520>

执行摘要

- 一句话: 修复预填充感知 SWA floor 跟踪逻辑
- 推荐动作: 值得精读。该修改展示了在 chunked prefill 场景下如何正确设置 SWA eviction floor, 对于理解 SGLang 的 SWA 缓存管理和调度逻辑有帮助。建议同步查看 PR #29186 了解完整背景。

功能与动机

Unlimited-OCR 模型使用 prefill-aware SWA 策略, 需要在 decode 阶段保留 prompt 和 image 的 KV 缓存。原实现中 `swa_evict_floor` 被设置为 `fill_len`, 但 `fill_len` 是请求的完整输入长度, 在 chunked prefill 场景下会导致 floor 过高, 使得后续 chunk 和在 decode 阶段可能错误地 evict 必要的 KV。通过提交注释和代码审查可知, 修正方法是将 floor 改为 `extend_range.end`, 即当前实际被处理到的位置。

实现拆解

1. 修正 scheduler.py 中的 floor 赋值: 在 `_get_new_batch_prefill_raw` 方法中, 将 `req.swa_evict_floor = req.fill_len` 改为 `req.swa_evict_floor = req.extend_range.end`, 确保 floor 仅保护当前 extend chunk 之前的内容。
2. 更新 schedule_batch.py 中的衍生注释: 修正 `Req.__init__` 中对 `full_untruncated_fill_ids` 的注释, 将 admission 更新字段从 `fill_len` 改为 `extend_range`, 以与事实一致。
3. 新增 Unlimited-OCR 集成测试: 添加 `test/registered/vlm/test_unlimited_ocr_server.py`, 使用 `TestOpenAIMLLMServerBase` 框架启动真实模型服务, 验证单图对话完成请求能够正确识别图像中的文字 ("12345" 和 "sglang")。
4. 简化 cookbook 配置: 将 `docs_new/src/snippets/configs/baidu/unlimited-ocr.jsx` 中的硬件支持从 6 种缩减为仅 h100, 并标记该配置已验证 (`verified: true`), 移除冗余的硬件匹配项。

关键文件:

- `python/sglang/srt/managers/scheduler.py` (模块 调度器; 类别 source; 类型 core-logic ; 符号 `_get_new_batch_prefill_raw`, `swa_evict_floor`) : 核心修复文件, 修改了 `swa_evict_floor` 的赋值表达式, 从 `fill_len` 改为 `extend_range.end`, 直接影响 SWA 缓存的 eviction 行为。

- python/sglang/srt/managers/schedule_batch.py (模块 批处理; 类别 source; 类型 core-logic; 符号 swa_evict_floor) : 修正了 Req.__init__ 中关于 full_untruncated_fill_ids 的注释, 将 admission 更新字段从 fill_len 改为 extend_range, 与其实际行为保持一致。
- test/registered/vlm/test_unlimited_ocr_server.py (模块 集成测试; 类别 test; 类型 test-coverage; 符号 TestUnlimitedOCRServer, _font, _make_ocr_image_url, test_single_image_chat_completion) : 新增的集成测试, 验证 Unlimited-OCR 模型在真实服务器端到端推理的正确性, 覆盖图像 OCR 场景。
- docs_new/src/snippets/configs/baidu/unlimited-ocr.jsx (模块 文档配置; 类别 source; 类型 configuration) : 简化了配置文件, 将硬件支持缩减为仅 H100, 并标记已验证。减少文档维护成本。

关键符号: _get_new_batch_prefill_raw, Req.init, test_single_image_chat_completion

关键源码片段

python/sglang/srt/managers/scheduler.py

核心修复文件, 修改了 swa_evict_floor 的赋值表达式, 从 fill_len 改为 extend_range.end, 直接影响 SWA 缓存的 eviction 行为。

```
# 在 _get_new_batch_prefill_raw 方法中 (调度器核心路径)
if self.tp_worker.model_runner.prefill_aware_swa:
    for req in can_run_list:
        # 修正: 使用 extend_range.end 而非 fill_len
        # 原代码: req.swa_evict_floor = req.fill_len # 错误: fill_len 是整个 prompt 长度
        # 在 chunked prefill 中, 这个值可能远大于当前 chunk 的结束位置,
        # 导致 floor 过高, 后续 decode 时可能错误保护了不应保护的 token。
        req.swa_evict_floor = req.extend_range.end
```

评论区精华

本 PR 的 review 评论全部来自自动机器人 gemini-code-assist, 仅反馈无具体意见, 无人类讨论。

- 暂无高价值评论线程

风险与影响

- 风险: 该修改涉及调度器核心路径和 SWA eviction 逻辑, 虽然改动量小 (仅一行赋值), 但错误设置会导致模型输出不正确。新增的集成测试提供了基本保障。风险等级中等, 建议在更多硬件 (非 H100) 上运行测试以确认兼容性。
- 影响: 影响范围集中于使用 prefill-aware SWA 的模型 (当前仅 Unlimited-OCR), 以及未来可能采用该策略的模型。对性能无负面影响, 但正确性得到保障。对系统其他模块无直接影响。cookbook 配置简化减少了文档维护量。
- 风险标记: 核心路径变更, 微小改动但影响正确性, 测试覆盖新增

关联脉络

- PR #29186 [New Model] Add Baidu Unlimited-OCR: 该 PR 引入了 Unlimited-OCR 模型和 prefill-aware SWA 逻辑，是当前修复的原始上下文和依赖。