PR #17706 完整报告

sgl-project/sglang

[bugfix] avoid attention padding tokens computation in pcg

合并时间：2026-04-14 16:08

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/17706

执行摘要

本 PR 修复了 Piecewise CUDA Graph (PCG) 模式下注意力计算中填充令牌被错误处理的问题，通过切片张量排除填充令牌，避免 FlashInfer 后端产生未定义行为（如 NaN、输出损坏）。修复后，先前禁用的测试 test_qwen3_next_models_pcg 重新启用并通过，提升了 PCG 的鲁棒性和推理可靠性。

功能与动机

为什么做：当 PCG 启用时，注意力元数据使用实际令牌数（real_num_tokens）初始化，但输入张量仍包含填充令牌，导致 FlashInfer 等注意力后端无法正确处理，可能引发 NaN 值、损坏的输出（如重复“!!!!!”）或异常输出长度。PR body 中明确指出“To fix this, *exclude the padded tokens and make PCG more robust*。”

实现拆解

改动模块：

1. 注意力计算层：radix_attention.py 和 radix_linear_attention.py 中，unified_attention_with_output 和 unified_linear_attention_with_output 函数新增切片逻辑：
   - 使用 forward_batch.num_token_non_padded_cpu 获取实际令牌数。
   - 将 query、key、value 等张量切片到该长度，排除填充令牌。
   - 动态修改和恢复 out_cache_loc，确保缓存写入正确位置。python real_num_tokens = forward_batch.num_token_non_padded_cpu query = query[:real_num_tokens] forward_batch.out_cache_loc = original_out_cache_loc[:real_num_tokens]
2. 后端清理：flashinfer_backend.py 移除 PCG 填充相关代码（如 extra_kv 和 pad_tokens 逻辑），简化实现。
3. PCG 运行器：piecewise_cuda_graph_runner.py 添加 num_token_non_padded_cpu 参数传递，确保实际令牌数在重放时可用。
4. 上下文管理：piecewise_context_manager.py 删除 num_tokens 字段，减少冗余。

评论区精华

讨论焦点：

- 字段命名与注释：gemini-code-assist[bot] 建议更新 real_num_tokens 注释，ch-wan 指出其等同于 num_token_non_padded，作者采纳并重命名为 num_token_non_padded_cpu，体现代码清晰度优化。
- 输出缓冲区初始化：Oasis-Git 和 hebiao064 讨论是否用 zeros 代替 empty，结论是移除零初始化以保持设计简洁，避免不必要更改。
- sinks 形状处理：ispobock 提醒 sinks 可能无令牌维度，作者更新代码保持 sinks 不变，确保模型兼容性。

风险与影响

技术风险：

- 回归风险：核心注意力路径变更可能影响所有 PCG 模式下的模型推理，需依赖 CI 测试覆盖。
- 性能开销：添加切片操作可能引入微小延迟，但相比未定义行为修复，利大于弊。
- 测试覆盖不足：边缘情况（如不同填充场景）可能未充分测试，建议补充单元测试。

影响评估：

- 用户：修复输出损坏问题，提升推理可靠性和确定性，尤其对 Qwen3-next 等模型用户有益。
- 系统：PCG 路径更稳健，减少未定义行为，可能间接改善性能。
- 团队：代码简化便于维护，移除冗余逻辑降低未来错误概率。

关联脉络

历史 PR 关联：

- PR #21452：被本 PR 回滚，原 PR 可能引入了填充处理逻辑，但本 PR 提供了更彻底的修复。
- PR #17404：评论中提及修复了 Mamba 缓存问题，与本 PR 共同提升 PCG 稳定性。
- 近期 PR 趋势：仓库近期多个 PR 涉及 PCG、填充和模型特定修复（如 PR #22739），显示团队持续优化推理路径的稳健性和性能。演进方向：本 PR 是 PCG 优化线的一部分，旨在通过简化逻辑和排除填充令牌，提升大规模模型推理的确定性和效率。