PR #24664 完整报告

sgl-project/sglang

Feat: Support SWA (Sliding Window Attention) for EAGLE-3 drafter

合并时间：2026-05-12 09:40

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/24664

执行摘要

- 一句话：为 EAGLE-3 draft 模型添加 SWA 支持
- 推荐动作：建议阅读，特别是参数统一和向后兼容的设计策略。对于需要支持长上下文的 EAGLE-3 部署，此特性有实用价值。注意文档更新未包含在此 PR 中，需后续补充。

功能与动机

EAGLE-3 模型训练长度通常在 2-4K 范围，超过训练长度的上下文会导致准确度显著下降。SWA 能限制注意力窗口，提升当上下文超过训练长度时的接受长度。PR body 引用即将发表的论文指出 SWA 有助于在未经过长上下文训练的模型上提升接受长度。同时需要统一 EAGLE 和 DFLASH 的 SWA 参数名，简化配置。

实现拆解

1. 参数定义与验证 (server_args.py)：添加新的 speculative_draft_window_size 字段（默认 None），移除原有的 speculative_dflash_draft_window_size。在 _handle_speculative_decoding 中更新验证逻辑，使用新字段并修复错误信息字符串格式化。在 CLI 参数解析中添加 --speculative-dflash-draft-window-size 作为别名，指向新字段。
2. EAGLE-3 模型层适配 (llama_eagle3.py)：在 LlamaDecoderLayer 构造函数中新增 draft_window_size 参数，若不为 None 则设置 self.self_attn.attn.sliding_window_size = draft_window_size。LlamaModel 构造函数也新增该参数，并在创建各层时传递。LlamaForCausalLMEagle3 类新增 get_attention_sliding_window_size 方法从全局 ServerArgs 读取参数，并在 __init__ 中调用，将结果传入 LlamaModel。
3. DFLASH worker 统一 (dflash_worker.py)：将 DFlashWorker.__init__ 中读取 draft window size 的来源从 speculative_dflash_draft_window_size 改为新的统一字段 speculative_draft_window_size。

关键文件：

- python/sglang/srt/models/llama_eagle3.py（模块 EAGLE 模型；类别 source；类型 core-logic；符号 get_attention_sliding_window_size）：核心模型层变更，为 EAGLE-3 draft 模型引入滑动窗口注意力支持，包括新的构造函数参数和控制 attention 层的 sliding_window_size。
- python/sglang/srt/server_args.py（模块 服务配置；类别 source；类型 configuration）：参数定义和验证的统一，添加通用参数 speculative_draft_window_size，移除 DFLASH 专

用参数，并通过别名保持向后兼容。

- python/sglang/srt/speculative/dflash_worker.py（模块 投机解码；类别 source；类型 core-logic）：将 draft window size 的读取源从旧参数切换到新统一参数。

关键符号：LlamaDecoderLayer.init, LlamaModel.init, get_attention_sliding_window_size, DFlashWorker.init

关键源码片段

python/sglang/srt/models/llama_eagle3.py

核心模型层变更，为 EAGLE-3 draft 模型引入滑动窗口注意力支持，包括新的构造函数参数和控制 attention 层的 sliding_window_size。

```
# 在 LlamaForCausalLMEagle3 中新增的方法，从全局 ServerArgs 获取 window_size
def get_attention_sliding_window_size(self) -> Optional[int]:
    server_args = get_global_server_args()
    draft_window_size: Optional[int] = (
        int(server_args.speculative_draft_window_size)
        if server_args.speculative_draft_window_size is not None
        else None
    )
    return draft_window_size

# 在 LlamaForCausalLMEagle3.__init__ 中，将 window_size 传入 LlamaModel
LlamaModel(
    config,
    quant_config=quant_config,
    draft_window_size=self.get_attention_sliding_window_size(),
    prefix=add_prefix("model", prefix),
)

# 在 LlamaDecoderLayer.__init__ 中，如果设置了 window_size，则覆盖 attention 的滑动窗口
if draft_window_size is not None:
    self.self_attn.attn.sliding_window_size = draft_window_size
```

评论区精华

Review 中有三个主要讨论：

- Qiaolin-Yu 询问 DFLASH 是否也做了 window_size 减 1，以保证概念一致性。Dogacel 回复说已移除减 1 逻辑，统一使用原始值。
- Qiaolin-Yu 指出错误信息中缺少 f-string 格式化，Dogacel 修正。
- Qiaolin-Yu 建议添加 --speculative-dflash-draft-window-size 作为别名以保持向后兼容，Dogacel 采纳该建议并实现。
- window_size 减 1 的一致性 (correctness): Dogacel 解释称 DFLASH 并未减 1，为保持一致，已移除减 1 逻辑，直接使用原始参数值。
- 错误信息字符串格式化 (style): 已修改为 f-string。

- 向后兼容性别名 (design): Dogacel 采纳建议，添加了别名并设置 dest 为新字段。

风险与影响

- 风险：
  1. 向后兼容风险：移除了旧的 CLI 参数，但通过别名支持，现有脚本若使用旧参数仍可工作，但 dest 已变更，可能影响直接读取 args 的代码。需确保所有内部引用已更新。
  2. 回归风险：SWA 设置可能影响 draft 模型的生成质量。PR 提供了 benchmark 数据，在长上下文场景下吞吐量提升、接受长度增加，但未测试所有模型。需注意默认关闭（None），不会影响现有行为。
  3. 无新增测试：此改动未添加独立的测试文件，仅依赖已有集成测试。建议补充针对 --speculative-draft-window-size 的单元测试。
  4. 性能影响：SWA 会减小注意力范围，理论上降低计算量，但可能降低准确度。对于已训练长上下文的模型，启用 SWA 可能适得其反。文档应说明使用场景。 - 影响：用户影响：用户可通过 --speculative-draft-window-size 控制 draft 模型的滑动窗口。原 DFLASH 用户需迁移配置，但别名保证了无缝过渡。系统影响：改动集中在投机解码模块，影响 EAGLE-3 和 DFLASH worker 的初始化路径。多模型配置（如同时使用 EAGLE 和 DFLASH）的参数解析统一。团队影响：减少一个冗余参数，简化配置逻辑，降低维护成本。
- 风险标记：向后兼容，无测试覆盖，核心路径变更

关联脉络

- PR #24663 Feat: Support newer EAGLE-3 drafters: 本 PR 在 #24663 支持新 EAGLE-3 drafters 的基础上添加 SWA 支持，扩展了其功能，二者构成完整的 EAGLE-3 改进链路。