PR #22088 完整报告

sgl-project/sglang

[sgl] add support for weight update function in spedec

合并时间：2026-04-21 07:26

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/22088

执行摘要

此 PR 为 sglang 仓库的推测解码模块添加了权重更新支持，通过在 multi_layer_eagle_worker_v2.py 和 eagle_worker_v2.py 中新增 update_weights_from_disk 和 update_weights_from_ipc 方法，使草案工作者能够从磁盘或 IPC 加载新模型权重。同时更新调度器混合类以协调更新流程，增强了系统在线模型热更新能力，但缺乏直接测试覆盖，需关注后续补充。

功能与动机

动机源于支持草案工作者中的 update_weights_from_disk 和 update_weights_from_ipc 功能，如 PR 描述所示。这允许在推测解码过程中动态更新模型权重，例如从检查点加载新参数或通过进程间通信传递权重，提升部署灵活性和运行时适应性。

实现拆解

1. 入口变更：两个工作者文件首先更新导入语句，添加 UpdateWeightFromDiskReqInput 和 UpdateWeightsFromIPCReqInput 类型引用。
2. 核心逻辑扩展：
   - 在 multi_layer_eagle_worker_v2.py 中，新增方法循环遍历 speculative_num_steps 个草案运行器，确保所有步骤同步更新。
   - 在 eagle_worker_v2.py 中，类似方法直接调用草案运行器，简化逻辑。
3. 调度器集成：修改 scheduler_update_weights_mixin.py，在 update_weights_from_disk 和 update_weights_from_ipc 方法中添加对草案工作者的调用，并处理缓存刷新，确保整体一致性。
4. 测试与部署配套：本次改动未包含测试文件，Issue 评论中提及后续 PR 应添加测试；无配置或部署变更。

评论区精华

Reviewer Qiaolin-Yu 提出关键问题：> 'Where do ipc_path and load_format come from? It seems UpdateWeightsFromIPCReqInput doesn't have these attributes'。这暴露了早期实现可能存在的参数传递错误，但最终代码已修正为直接传递整个请求对象，避免了类型不匹配风险，体现了代码审查对正确性的重视。

风险与影响

- 技术风险：权重同步失败可能导致草案模型与主模型不一致，影响解码准确性；循环更新可能引入性能开销；缺少测试增加回归风险。
- 影响评估：用户可通过 API 实现模型热更新，提升操作灵活性；系统增强在线权重管理能力；团队需跟进测试补充以确保功能稳定。

关联脉络

从近期历史 PR 看，PR 22832 同样修改了 multi_layer_eagle_worker_v2.py 文件，修复 CUDA Graph 相关 bug，显示该文件是推测解码的核心组件。PR 21599 引入自适应推测解码步骤，与本 PR 同属推测解码功能演进线，表明该模块正持续扩展以优化性能和功能。