PR #7215 完整报告

PaddlePaddle/FastDeploy

[Speculative Decoding] Auto-scale CUDA graph capture sizes for speculative decoding

合并时间：2026-04-07 20:22

原文链接：http://prhub.com.cn/PaddlePaddle/FastDeploy/pull/7215

执行摘要

本 PR 为推测解码（Speculative Decoding）功能自动缩放 CUDA 图捕获尺寸，解决了用户需手动计算的痛点。通过修改配置初始化和 GPU 运行器逻辑，简化了设置流程，提升易用性。变更影响范围集中，风险较低，但需注意变量名拼写错误未修正。

功能与动机

动机：根据 PR body 描述，当前使用推测解码时，用户需要手动计算 CUDA 图捕获尺寸（batch size 乘以 num_speculative_tokens + 1），这种方式容易出错且不便。PR 旨在自动化这一过程，减少配置错误。

实现拆解

实现分为两个关键文件改动：

1. fastdeploy/config.py：修改 GraphOptimizationConfig.init_with_cudagrpah_size 方法，新增 num_speculative_tokens 参数，并在推测解码启用时自动缩放捕获尺寸。核心逻辑如下： python if not self.flag_cudagraph_capture_sizes_initlized and num_speculative_tokens != 0: self.cudagraph_capture_sizes = [size * (num_speculative_tokens + 1) for size in self.cudagraph_capture_sizes if (size * (num_speculative_tokens + 1)) <= max_capture_size] 同时构建 real_bsz_to_captured_size 映射，用于批量大小到捕获尺寸的转换。
2. fastdeploy/worker/gpu_model_runner.py：在 capture_model 方法中，将推测解码的条件从仅 SpecMethod.MTP 扩展为包含 SpecMethod.SUFFIX，确保两者使用相同的捕获流程。

评论区精华

review 中仅有一个来自 fastdeploy-bot 的评论：

> 🤔 建议变量名拼写错误 flag_cudagraph_capture_sizes_initlized 应为 flag_cudagraph_capture_sizes_initialized（缺少字母 i）。

该建议未被采纳，变量名在 PR 中保持原样，可能影响代码可读性。

风险与影响

- 风险：变量名拼写错误虽不影响功能，但降低代码整洁度；自动缩放逻辑依赖 num_speculative_tokens 参数，若传入错误值可能导致捕获尺寸计算偏差；修改涉及 CUDA 图核心配置，需测试确保在推测解码和非推测解码场景下均正常。
- 影响：对用户简化配置，减少手动错误；对系统可能优化内存和性能，但需验证；对团队影响范围可控，集中在配置和运行器模块。

关联脉络

从近期历史 PR 看，本 PR 与多个推测解码相关优化（如 PR#7136、#7172、#7201）形成脉络，共同提升推测解码性能和易用性。这些 PR 涉及 GPU kernel 优化、bug 修复和 OP 简化，表明团队正持续改进推测解码模块。本 PR 的自动缩放功能是这一趋势的延续，旨在降低用户使用门槛。