PR #7231 完整报告

PaddlePaddle/FastDeploy

[Speculative Decoding] Remove arctic_inference deps

合并时间：2026-04-08 15:25

原文链接：http://prhub.com.cn/PaddlePaddle/FastDeploy/pull/7231

执行摘要

- 一句话：将 arctic_inference 从硬依赖改为可选依赖，仅在 Suffix Decoding 功能使用时按需安装。
- 推荐动作：该 PR 变更简单，值得快速浏览以了解依赖管理策略。关注点：版本号不一致问题是否需后续修复，以及文档是否需同步更新。

功能与动机

PR 的 body 中未明确说明动机，但从变更内容推断，目的是将 arctic_inference 从硬依赖改为可选依赖，以减少不必要的依赖安装。只有使用 Suffix Decoding 功能时才需要安装此包。

实现拆解

实现方案包括两个关键改动：1. 在 fastdeploy/spec_decode/suffix.py 中，更新 SuffixProposer 的 __init__方法中的 ImportError 错误提示，从泛化的提示改为具体安装命令和版本号（0.1.2）。2. 在 requirements.txt 中移除 arctic_inference 依赖行。

关键文件：

- fastdeploy/spec_decode/suffix.py（模块 Speculative Decoding）：更新 SuffixProposer 的错误提示，明确 arctic_inference 的安装命令和版本号，是功能可用性的关键文件。
- requirements.txt（模块 infra）：移除 arctic_inference 硬依赖，减少不必要的包安装，影响整个项目的依赖管理。

关键符号：SuffixProposer.init

评论区精华

review 中仅有一个建议：fastdeploy-bot 指出版本号不一致问题，错误信息中指定的是 arctic-inference==0.1.2，但原 requirements.txt 使用的是 arctic_inference-0.1.3 版本。建议确认正确的版本号并保持一致。该建议未在 PR 中解决，但 PR 已被合并。

- 版本号不一致 (correctness): 未在 PR 中解决，PR 已被合并，版本不一致问题可能遗留。

风险与影响

- 风险：风险较低，主要涉及依赖管理：1. 版本不一致风险：错误提示中的版本号（0.1.2）可能与实际可用版本（原为 0.1.3）不匹配，可能导致用户安装错误版本或安装失败。2. 功

能可用性风险：如果用户未安装 arctic_inference 包，Suffix Decoding 功能将无法使用，但错误提示已更新，应能引导用户正确安装。

- 影响：影响范围有限：1. 对用户：减少默认安装的依赖包，简化部署；但使用 Suffix Decoding 功能时需手动安装 arctic_inference。2. 对系统：无性能或安全影响，仅改变依赖配置。3. 对团队：需在文档中更新依赖说明，确保用户知晓可选依赖的安装方式。
- 风险标记：版本不一致，依赖管理变更

关联脉络

- PR #7136 [Optimization]【Hackathon 10th Spring No.49】GPU ngram_match: BlockScan Phase 2 -optimized: 同属 Speculative Decoding 模块，涉及推测解码的优化和 kernel 实现。
- PR #7215 [Speculative Decoding] Auto-scale CUDA graph capture sizes for speculative decoding: 同属 Speculative Decoding 模块，涉及推测解码的功能增强和配置优化。