PR #7159 完整报告

PaddlePaddle/FastDeploy

[Feature] Support set PREEMPTED_TOKEN_ID in GET_SAVE_OUTPUT_V1

合并时间：2026-04-08 19:30

原文链接：http://prhub.com.cn/PaddlePaddle/FastDeploy/pull/7159

执行摘要

本次 PR 为 FastDeploy 的 GPU 模型运行器增加了请求抢占信号传递支持。当启用 FD_USE_GET_SAVE_OUTPUT_V1 环境变量时，被抢占的请求会将其采样令牌 ID 设置为特殊值 PREEMPTED_TOKEN_ID(-9)，以此通知服务器端中断操作已完成。这是一个小规模但重要的功能增强，确保了抢占流程的正确结束。

功能与动机

根据 PR 描述，该变更的目的是 " 支持在开启 FD_USE_GET_SAVE_OUTPUT_V1 时，同步中断完成信号 "。具体来说，在使用 GET_SAVE_OUTPUT_V1 模式时，当请求被抢占（preempted）后，没有对应的采样令牌。通过设置 PREEMPTED_TOKEN_ID(-9)，可以通知服务器端 abort 操作已完成，使抢占流程能够正确结束。

实现拆解

实现集中在 fastdeploy/worker/gpu_model_runner.py 文件的 _postprocess 方法中：

1. 导入变更：从 fastdeploy.config 导入 PREEMPTED_TOKEN_ID 常量
2. 核心逻辑：在 FD_USE_GET_SAVE_OUTPUT_V1 启用时，检查 last_preempted_idx 数组
3. 令牌替换：使用 paddle.where 条件操作，将被抢占请求的 sampled_token_ids 设置为 PREEMPTED_TOKEN_ID

关键代码段：

```
if envs.FD_USE_GET_SAVE_OUTPUT_V1:
    paddle.assign(
        paddle.where(
            self.share_inputs["last_preempted_idx"][: sampler_output.sampled_token_ids.shape[0]] ==
            = 1,
            PREEMPTED_TOKEN_ID,
            sampler_output.sampled_token_ids,
        ),
        sampler_output.sampled_token_ids,
    )
```

评论区精华

AI 代码审查机器人 fastdeploy-bot 指出了关键问题：

> 🤔 Bug变量名错误：envs.GET_SAVE_OUTPUT_V1 不存在，应为 envs.FD_USE_GET_SAVE_OUTPUT_V1

这个错误会导致运行时 AttributeError，使条件分支永远无法进入。开发者随后修复了这个问题。审查还提到：

> 代码逻辑清晰，与已有的 token_processor.py 中处理 PREEMPTED_TOKEN_ID 的逻辑保持一致。

风险与影响

风险分析：

- 环境变量名错误已修复，但原始错误可能导致运行时异常
- 需要确保服务器端能正确处理 PREEMPTED_TOKEN_ID(-9)，否则可能导致协议不一致
- 代码覆盖率报告显示有 1 行代码缺少测试覆盖

影响评估：

- 对使用 FD_USE_GET_SAVE_OUTPUT_V1 模式的用户：增强了请求抢占处理能力
- 对系统：提高了 GPU 模型运行器在抢占场景下的健壮性
- 对团队：变更规模小，易于维护，与现有处理逻辑保持一致

关联脉络

从近期历史 PR 分析可以看出：

1. 相同文件修改：PR #7165、#7147、#7215 都修改了 fastdeploy/worker/gpu_model_runner.py 文件，表明这是 FastDeploy 中频繁修改的核心组件
2. GPU 优化趋势：多个 PR 涉及 GPU 性能优化（如 #7165 的 TBO 优化、#7215 的 CUDA 图自动缩放），本次 PR 延续了这一方向
3. 引擎改进：与 #7102（修复流式解码令牌丢失）类似，都是对请求处理生命周期的完善

本次 PR 虽然规模小，但体现了 FastDeploy 在完善请求生命周期管理和异步处理信号传递方面的持续改进。