PR #7096 完整报告

PaddlePaddle/FastDeploy

[XPU] Add TP broadcast after sampling in XPU model runner.

合并时间：2026-04-08 19:26

原文链接：http://prhub.com.cn/PaddlePaddle/FastDeploy/pull/7096

执行摘要

该 PR 修复了 FastDeploy 中 XPU 设备在 Tensor Parallel（TP）模式下各 rank 采样结果不一致的问题。通过在 fastdeploy/worker/xpu_model_runner.py 的采样逻辑后添加 TP 组内广播操作，确保所有 rank 以 rank0 的采样令牌为准，提升了多卡推理的确定性。变更简单但关键，影响范围限于 XPU 且启用 TP 的场景，已合并至 develop 分支。

功能与动机

问题背景：在 TP 模式下，每个 rank 独立进行随机采样，由于随机种子可能不同，导致各 rank 生成不同的令牌序列，造成多卡推理输出不一致。

解决方案：根据 PR body 中的明确表述，"Added a broadcast operation after sampling in the XPU model runner to synchronize the sampled tokens from rank 0." 即通过广播同步 rank0 的采样结果，强制所有 rank 使用相同令牌。

实现拆解

修改仅涉及一个文件，在采样后添加条件广播逻辑：

```
if self.parallel_config.tensor_parallel_size > 1:
    paddle.distributed.broadcast(
        sampler_output.sampled_token_ids,
        self.parallel_config.data_parallel_rank * self.parallel_config.tensor_parallel_size,
        group=self.parallel_config.tp_group,
    )
```

关键改动点： | 路径 | 广播张量 | 说明 | |------|----------|------| | 非投机解码 | sampled_token_ids | 同步采样出的令牌 ID | | 投机解码 | accept_tokens, accept_num, step_idx, stop_flags | 同步投机解码相关的四个状态张量 |

src rank 计算公式为 data_parallel_rank * tensor_parallel_size，使用 tp_group 通信组进行广播。

评论区精华

Copilot 在 review 中提出了代码优化建议，但未被采纳：

> " 这里 src（root rank）的计算表达式在多处重复使用 ... 建议先用局部变量保存，再传给 broadcast，避免复制粘贴带来的维护风险。"

> "speculative 分支里连续多次调用 broadcast... 建议复用同一个 tp_src_rank 变量，并考虑用一个 key 列表循环广播这些张量，降低后续新增 / 修改字段时遗漏的概率。"

这些建议指向代码重复和维护性问题，但 PR 最终以原始实现合并，由 cmcamdy 批准（"LGTM"）。

风险与影响

技术风险：

1. 通信开销：新增广播操作会增加 TP 模式下的通信延迟，特别是在投机解码路径需要广播四个张量。
2. 维护风险：重复的 data_parallel_rank * tensor_parallel_size 计算逻辑，如 Copilot 指出，增加了未来修改时出错的风险。
3. 设备一致性：仅修改 XPU 模型运行器，未同步修改 GPU 等其他设备的对应逻辑，可能导致不同硬件平台行为不一致。

影响评估：

- 正面影响：彻底解决了 TP 模式下采样结果不一致的问题，提升了多卡推理的可靠性。
- 性能影响：通信开销是必要代价，影响仅限于启用 TP 的 XPU 场景。
- 范围：变更集中，易于理解和回滚。

关联脉络

与历史 PR 的关系：

1. PR#7159：同样修改模型运行器（GPU 版本），涉及采样和令牌处理，表明采样同步是分布式推理的通用需求。
2. PR#7165：关注 GPU 模型运行器的性能优化（TBO），而本 PR 关注 XPU 的正确性修复，体现不同设备模型的并行维护策略。
3. PR#7147：曾修改同一文件但仅修复拼写错误，凸显本 PR 的功能性价值。

演进趋势：FastDeploy 近期多个 PR（如 #7136、#7215）聚焦推测解码优化，本 PR 的投机解码路径广播与之协同，确保分布式环境下推测解码的正确性。XPU 支持作为重要方向，本 PR 是确保其 TP 模式可靠性的关键一步。