

PR #2223 完整报告

THUDM/slime

Fix --save-hf

合并时间: 2026-07-20 17:38

原文链接: <http://prhub.com.cn/THUDM/slime/pull/2223>

执行摘要

- 一句话: 修复 --save-hf 的 HuggingFace 权重转换逻辑
- 推荐动作: 值得合并, 修复了明显 bug。建议同步检视相关测试覆盖。

功能与动机

修复 --save-hf 功能, 确保 HuggingFace 权重转换正确性, 特别是对 SGLang 兼容性至关重要的 q_a_proj 与 kv_a_proj 配对。

实现拆解

1. 修改 slime/backends/megatron_utils/hf_checkpoint_saver.py 中 save_hf_model_direct_to_path 函数, 调整 should_convert_chunk 参数从 is_writer_rank and idx % num_save_nodes == save_node_rank 改为仅 is_writer_rank, 使得每个 writer 节点都处理所有 chunk。
2. 添加注释解释原因: Megatron 到 HF 的转换对于某些参数是有状态的, 例如 q_a_proj 和 kv_a_proj 可能位于相邻的 chunk 中, 但必须一起发射以保持 SGLang 兼容性。
3. 实际写入持久化时仍保持取模分片, 非 writer 节点跳过转换。

关键文件:

- slime/backends/megatron_utils/hf_checkpoint_saver.py (模块 后端; 类别 source; 类型 core-logic): 核心变更文件, 修复 HuggingFace 检查点保存时权重转换 chunk 过滤逻辑

关键符号: save_hf_model_direct_to_path

关键源码片段

[slime/backends/megatron_utils/hf_checkpoint_saver.py](#)

核心变更文件, 修复 HuggingFace 检查点保存时权重转换 chunk 过滤逻辑

```
# slime/backends/megatron_utils/hf_checkpoint_saver.py
```

```
# 在 save_hf_model_direct_to_path 函数中, 原先是:
```

```
# should_convert_chunk=lambda idx: is_writer_rank and idx % num_save_nodes == save_node_rank
```

```
# 现在改为:
```

```
should_convert_chunk=lambda _idx: is_writer_rank, # 所有 writer 处理所有 chunk
```

原因说明（代码内注释）：

```
# Megatron-to-HF conversion is stateful for some parameters. For
# example, q_a_proj and kv_a_proj can land in adjacent chunks but
# must be emitted together for SGLang compatibility. Every node
# writer therefore has to observe every chunk so that pairs can
# cross chunk boundaries. Writers still only persist their
# modulo-assigned shards below; non-writer ranks skip conversion.
```

评论区精华

该 PR 无 review 评论。

- 暂无高价值评论线程

风险与影响

- 风险：风险较低，但需注意：所有 writer 节点现在都执行转换，可能小幅增加 CPU 开销；但持久化写操作仍按取模分片，对磁盘写入无影响。
- 影响：影响范围较小，仅影响 `--save-hf` 路径，修复 HuggingFace 检查点保存的正确性。对 SGLang 部署兼容性有正面影响。
- 风险标记：核心路径变更，缺少测试覆盖

关联脉络

- PR #2220 Optimize update weight: 来自同一分支（zilin/optimize_update_weight），其中包含其它优化，此 PR 是修复该分支引入的 bug