PR #1812 完整报告

THUDM/slime

feat: add support for including missing weights from origin HF checkp…

合并时间：2026-04-07 14:56

原文链接：http://prhub.com.cn/THUDM/slime/pull/1812

执行摘要

- 一句话：在权重转换工具中添加从原始 HF 检查点补充缺失权重的功能，提升 Qwen3.5 模型转换完整性。
- 推荐动作：该 PR 值得关注其设计思路：通过维护已转换权重名称集合和从原始检查点补充缺失项的方式，优雅地解决了部分权重转换问题。建议精读 save_tensors 函数中新增的权重补充逻辑，理解其如何保持转换完整性同时避免重复。

功能与动机

根据 PR 标题和 body 描述，该变更旨在解决将部分 Qwen3.5 Megatron 权重转换回完整 HuggingFace 格式时，某些权重可能缺失的问题。PR body 明确指出“The code is to convert partial Qwen3.5 Megatron weights back to the complete Huggingface format”，表明这是针对特定模型转换场景的功能增强。

实现拆解

主要修改集中在 tools/convert_torch_dist_to_hf.py 文件：

1. 在 save_tensors 函数签名中添加 origin_hf_dir 参数，默认为 None
2. 在转换过程中维护 converted_names 集合记录已转换的权重名称
3. 当 origin_hf_dir 不为 None 时，遍历原始 HF 检查点的 safetensors 文件，将未在 converted_names 中的权重补充到输出中
4. 在命令行参数解析器中添加 --add-missing-from-origin-hf 标志（简写 -a）
5. 修改主函数调用逻辑，根据标志决定是否传递 origin_hf_dir 参数

关键文件：

- tools/convert_torch_dist_to_hf.py（模块 tools）：这是唯一被修改的文件，包含了权重转换工具的核心逻辑变更。新增的权重补充功能直接影响模型转换的完整性和正确性。

关键符号：save_tensors

评论区精华

由于 review_comments_count 为 0，没有实际的 review 讨论发生。PR 由 zhuzilin 直接合并，表明变更被认为是直接且低风险的。

- 暂无高价值评论线程

风险与影响

- 风险：1. 功能风险：新增的权重补充逻辑可能错误地包含不应转换的权重（如优化器状态），但通过 converted_names 集合的过滤机制降低了此风险。 2. 兼容性风险：修改了 save_tensors 函数签名，但通过默认参数保持了向后兼容性，不影响现有调用。 3. 性能风险：需要额外读取原始 HF 检查点文件，可能增加转换时间和内存使用，但仅当显式启用 --add-missing-from-origin-hf 标志时才触发。 4. 逻辑风险：如果原始 HF 检查点与目标模型结构不匹配，补充的权重可能无效，但这是用户责任而非工具问题。
- 影响：1. 对用户：为使用 Qwen3.5 等模型进行部分权重转换的用户提供了更完整的转换能力，避免因缺失权重导致的模型加载失败。 2. 对系统：工具功能增强，不影响核心训练/推理流程。 3. 对团队：简化了模型权重转换的工作流，减少了手动处理缺失权重的需求。
- 风险标记：函数签名变更，外部依赖读取

关联脉络

- PR #1769 Support FP8 conversion for Qwen3.5: 同样涉及 Qwen3.5 模型的权重转换工具（tools/convert_hf_to_fp8.py），属于同一技术领域的功能增强。
- PR #1799 fix qwen3.5 397B converting error when enable expert parallel: 都针对 Qwen3.5 模型的转换问题，1799 修复权重转换错误，本 PR 增强转换完整性，存在功能关联。