PR #1828 完整报告

THUDM/slime

Bugfix: use cpu instead of cuda in convert_torch_dist_to_hf.py when --add-missing-from-origin-hf is set

合并时间：2026-04-15 14:37

原文链接：http://prhub.com.cn/THUDM/slime/pull/1828

执行摘要

- 一句话：修复权重转换工具中补充缺失权重时错误使用 CUDA 设备的问题。
- 推荐动作：该 PR 代码变更简单明了，适合快速了解权重转换工具的设备处理逻辑。值得关注的是工具设计中设备选择的考量：在离线预处理任务中优先保证兼容性而非性能。

功能与动机

根据 PR 标题和 body 描述，修复 convert_torch_dist_to_hf.py 脚本在设置 --add-missing-from-origin-hf 参数时错误使用 CUDA 设备的问题。该问题会导致在没有 GPU 的环境中运行脚本时失败，因为脚本尝试将 safetensors 文件加载到 CUDA 设备。

实现拆解

1. 定位问题代码段：在 tools/convert_torch_dist_to_hf.py 文件的 save_tensors 函数中，当 origin_hf_dir 不为 None 时，脚本会遍历原始 HF 检查点的 safetensors 文件。
2. 修改设备参数：将 safetensors.safe_open 调用的 device 参数从 "cuda" 改为 "cpu"（第 131 行）。这样确保文件始终加载到 CPU 内存，避免对 GPU 设备的依赖。
3. 影响分析：修改后，补充缺失权重的逻辑可以在任何环境中运行，包括没有 GPU 的机器。由于权重转换通常是离线预处理任务，使用 CPU 加载不会影响最终转换结果的质量。

关键文件：

- tools/convert_torch_dist_to_hf.py（模块 工具脚本；类别 source；类型 core-logic；符号 save_tensors）：这是本次 PR 唯一修改的文件，包含权重转换工具的核心逻辑。修复了补充缺失权重时的设备选择问题。

关键符号：save_tensors

关键源码片段

tools/convert_torch_dist_to_hf.py

这是本次 PR 唯一修改的文件，包含权重转换工具的核心逻辑。修复了补充缺失权重时的设备选择问题。

```
def save_tensors(args, model_name, state_dict, output_dir, chunk_size, vocab_size=None, origin_hf_dir=None):
```

```
    # ... 前面的权重转换逻辑 ...

    if origin_hf_dir is not None:
        # 当提供了原始HF检查点目录时，补充当前转换中缺失的权重
        safetensors_files = [f for f in os.listdir(origin_hf_dir) if f.endswith(".safetensors")]
        for filename in safetensors_files:
            # 修复：将device参数从"cuda"改为"cpu"，确保在没有GPU的环境中也能正常运行
            with safetensors.safe_open(os.path.join(origin_hf_dir, filename), framework="pt",
            device="cpu") as f:
                for k in f.keys():
                    if k not in converted_names:
                        # 补充缺失的权重张量
                        converted_name = k
                        print(f"add {k} from origin hf checkpoint")
                        converted_param = f.get_tensor(k)
                        converted_names.add(k)
                        # ... 后续的存储逻辑 ...

    # ... 后续的元数据生成和文件保存逻辑 ...
```

评论区精华

本次 PR 没有 review 评论，直接由 zhuzilin 合并。从提交历史看，这是一个简单的单行修复，没有引发设计讨论或争议。

- 暂无高价值评论线程

风险与影响

- 风险：低风险：
  1. 功能回归风险：极低。仅修改了设备参数，不改变算法逻辑。补充的权重数据内容不变，只是存储位置从 GPU 内存改为 CPU 内存。
  2. 性能影响：在 GPU 环境中，从 CPU 内存访问张量可能比从 GPU 内存稍慢，但权重转换是离线任务，通常不要求实时性。
  3. 兼容性风险：无。修改后工具兼容性更好，既支持 GPU 环境也支持纯 CPU 环境。
  4. 安全风险：无新增安全漏洞。
- 影响：影响范围：
  1. 用户影响：使用 --add-missing-from-origin-hf 参数的用户现在可以在没有 GPU 的机器上运行权重转换工具，提高了工具的可移植性。
  2. 系统影响：工具的内存使用模式从可能使用 GPU 内存变为只使用 CPU 内存，在 GPU 环境中可能增加 CPU-GPU 数据传输开销，但影响轻微。
  3. 团队影响：简化了开发环境配置要求，开发者不再需要 GPU 即可测试权重补充功能。
- 风险标记：低风险变更

关联脉络

- PR #1812 feat: add support for including missing weights from origin HF checkp…: PR #1812 新增了 --add-missing-from-origin-hf 功能，本次 PR 修复了该功能中的一个设备选择 bug。两者都修改了同一个文件的相同代码区域。