

基于流水线并行和卸载策略的大模型精调 系统 V1.0 软件说明书

目录

1 简介 错误! 未定义书签。

 1.1 编写目的 3

 1.2 使用对象 3

2 系统需求 3

3 软件流程概述 3

 3.1 总体流程 3

 3.2 训练流程 4

4 安装和运行方法 6

 4.1 获取软件源代码 6

 4.2 依赖环境准备 7

 4.3 配置和使用 7

 4.4 命令行参数说明 8

 4.5 程序执行结果 8

5 软件运行截图 9

 5.1 显示帮助内容 9

 5.2 运行核心功能 9

1 产品概述

本软件提供了一个大模型精调框架，优化大模型精调的显存占用及设备间通信效率，提升性能与资源利用率。旨在通过高效的流水线并行和模型卸载策略，帮助用户在有限的硬件条件下实现大模型的精调任务。软件可以帮助用户实现的功能表现为从开源网站引用大模型，使用开源数据集，在单节点上高效利用多设备资源并行精调大模型。在保证模型精度和训练效率的前提下，本技术将可精调的模型规模提升至单设备方法的 5 倍。例如，在传统单设备精调方法仅支持 1B 参数规模的情况下，本技术可扩展至 5B 参数规模的大模型精调，突破单设备显存瓶颈，实现更大规模模型的高效训练。

1.1 编写目的

本文档为使用说明文档，为产品的使用与维护提供信息基础。

1.2 使用对象

本文档的使用对象主要为产品的使用人员。产品的目标用户是有精调大模型需求但硬件资源有限的用户群体。

2 环境需求

本软件适用系统：Linux 系统。推荐 Ubuntu 22.04.5 LTS 操作系统。

3 软件流程概述

3.1 总体流程

基于流水线并行和卸载策略的大模型精调系统总体流程如图 3-1 总体流程图所示。

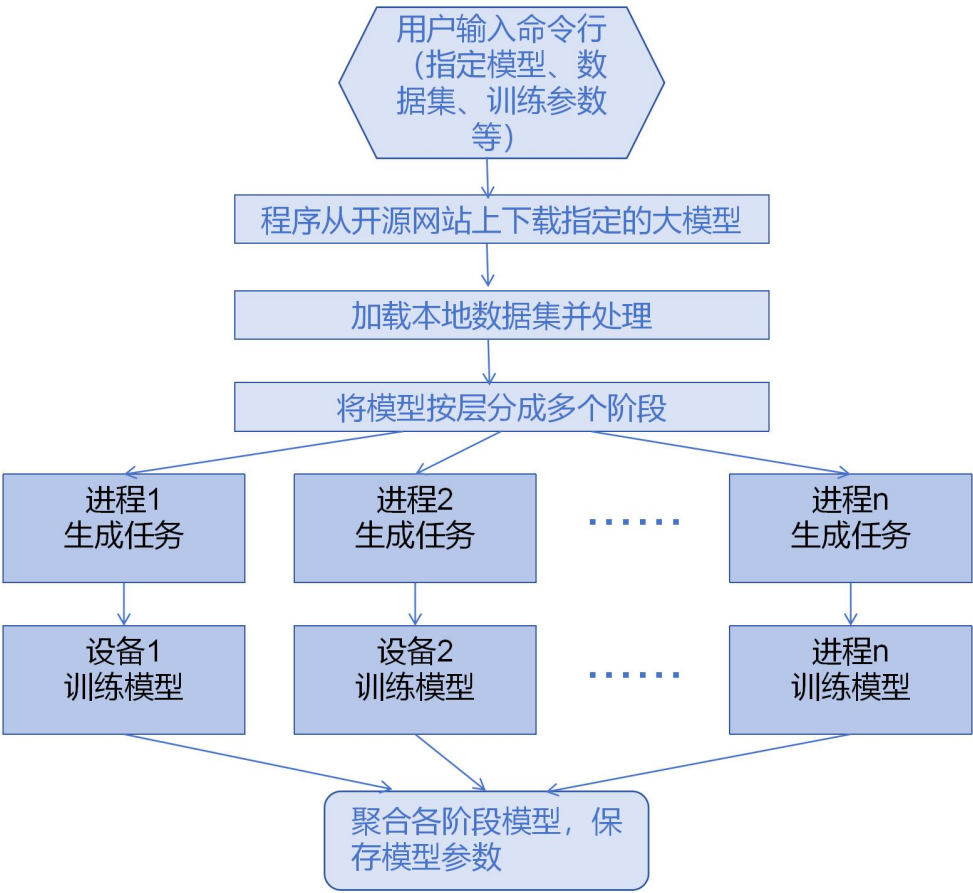


图 3-1 总体流程图

3.2 训练流程

3.2.1 背景和目的

本流程图展示了在特定模型和硬件设备条件下，采用流水线并行技术和模型卸载/加载技术进行模型精调 (Fine-tuning) 的训练过程。通过将模型划分为多个阶段，并利用多个加速设备并行处理不同阶段的任务，本系统显著提高了训练效率。通过在训练过程中卸载/加载模型参数，本系统显著提高了显存利用率。

3.2.2 模型精调的一般过程

模型精调 (Fine-tuning) 是指在预训练模型的基础上，通过进一步训练使其适应特定任务的过程。通常，模型精调包括以下步骤：

- ✧ 加载预训练模型：从预训练模型中加载权重和参数。
- ✧ 划分数据集：将训练数据划分为多个微批次（Mini-batches）。
- ✧ 前向传播：计算每个微批次的输出。
- ✧ 计算损失：根据输出结果和真实标签计算损失值。
- ✧ 反向传播：根据损失值更新模型参数。
- ✧ 迭代训练：重复上述步骤，直到模型收敛。

3.2.3 训练流程图以及内容详细说明

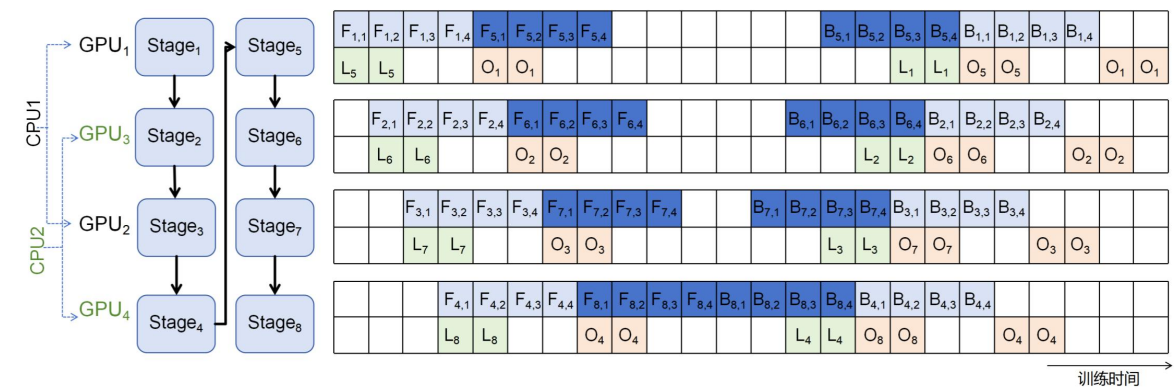


图 3-2 训练流水线示意图

- ✧ 模型划分：将模型划分为 8 个阶段，每个阶段由不同的加速设备负责。例如：
 - 设备 1 负责阶段 1 和阶段 5。
 - 设备 2 负责阶段 2 和阶段 6。
 - 以此类推。
- ✧ 操作符号
 - $F_{i,j}$ ：表示对阶段 i 的第 j 个微批次进行前向传播操作。
 - $B_{i,j}$ ：表示对阶段 i 的第 j 个微批次进行反向传播操作。
 - L_i ：表示对阶段 i 的模型参数进行加载操作。
 - O_i ：表示对阶段 i 的模型参数进行卸载操作。
- ✧ 流水线并行
 - 每个加速设备同时处理不同微批次的不同阶段任务。例如：
 - 设备 1 在处理阶段 1 的第 j 个微批次的同时，设备 2 已经开始处理阶段 2 的第 j 个微批次。
 - 通过这种流水线并行的方式，系统能够充分利用硬件资源，减少空闲时间，提高训练效率。

✧ 动态模型移动

在前向传播/反向传播计算前, 将模型参数加载至加速设备, 传播计算完成后, 将模型参数卸载至 CPU。这种方法能够提高显存利用率, 为中间结果提供更多可用的显存。

3.2.4 总结技术优势

本系统的创新点在于:

- ✧ 单设备多阶段流水线并行: 传统流水线并行将模型层分割为 N 个阶段, 其中 N 表示 GPU 数量。我们采用虚拟流水线策略, 将模型层分割为 $v \times N$ 个阶段, 每个 GPU 轮流执行 v 个流水线阶段。以 $N=4$, $v=2$ 为例。8 个虚拟阶段 ($v \times N=2 \times 4$) 在 4 个 GPU 间循环分布: GPU1 处理阶段 1 和 5, GPU2 处理阶段 2 和 6, GPU3 处理阶段 3 和 7, GPU4 处理阶段 4 和 8。动态加载和卸载: 通过动态加载 (Li) 和卸载 (Oi) 模型参数, 减少了设备间的通信开销。
- ✧ 模型卸载: 在训练过程中动态将模型参数卸载到 CPU, 为激活值和梯度节省高带宽内存 (HBM)。
- ✧ 通信-计算重叠: 优化流水线执行, 使得模型移动带来的通信和计算操作尽量重叠, 减少设备计算资源空闲时间。
- ✧ 交叉映射: 战略性地放置流水线阶段以最小化通信开销。
- ✧ 多流多线程执行: 并发数据移动和计算以提高吞吐量。

4 安装和运行方法

4.1 获取软件源代码

用户需从 GitHub 获取本软件的源代码, 可使用以下命令下载:

```
git clone https://github.com/MLSysU/Mobius/commits/bishe/
cd Gpipe
```

4.2 依赖环境准备

本软件支持在不同硬件环境上运行, 用户需根据自身的硬件选择合适的依赖环境。推荐使用 conda 创建虚拟环境, 安装必要的依赖项:

```
conda create -n finetune-env python=3.9  
conda activate finetune-env
```

然后安装依赖:

```
pip install -r requirements.txt
```

如果使用 GPU, 请确保已正确安装 CUDA 和 cuDNN, 并安装适配的 PyTorch 版本。

或者, 可以直接使用 docker 镜像:

```
docker pull coir1hat1man/mobius:latest
```

4.3 配置和使用

在运行软件前, 用户需要准备数据集和开源模型地址, 并指定相关参数。可以通过命令行运行 (举例) :

下方命令解释: 如果使用 GPU, CUDA_VISIBLE_DEVICES 用于指定 GPU 的编号; 使用 torchrun 运行程序; nproc_per_node 用于指定单节点上使用的设备数; master_port 用于指定运行程序的端口; main.py 是程序的主文件; 其他程序命令行参数说明见表 4-1。

```
CUDA_VISIBLE_DEVICES=0,1,2,3 torchrun --nproc_per_node 4 --master_port  
29502 ./main.py --model_path='meta-llama/Llama-2-7b-hf'  
--dataset='xsum'  
--num_iterations=20 --batch_size=64 --num_stage=8 --use_prefetch  
--use_offload
```

4.4 命令行参数说明

表 4-1 命令行参数表

序号	参数名	参数类型	默认值	参数描述
1	model_path	字符串	'meta-llama/Llama-2-7b-hf'	开源模型名称
2	dataset	字符串	'xsum'	本地数据集名称
3	save_params	字符串	'finetune_params.pt'	精调完成之后，保存模型新参数的文件名
4	use_prefetch	布尔		使用 prefetch 技巧
5	no_prefetch	布尔		不使用 prefetch 技巧
6	use_offload	布尔		使用模型参数卸载技巧
7	no_offload	布尔		不使用模型参数卸载技巧
8	batch_size	整数	128	批次大小
9	num_chunks	整数	4	微批次数
10	seq_length	整数	512	句长
11	embedding_dim	整数	4096	嵌入层维度
12	ff_dim	整数	4096	前馈层维度
13	num_iterations	整数	2	迭代次数
14	num_stages	整数	8	将模型切分成的阶段数
15	num_layers	整数	8	训练的模型层数
16	num_heads	整数	32	多注意力头数

4.5 程序执行结果

在精调完成后，将优化后的模型参数保存至指定的 .pt 文件中，以便用户在后续推理或其他下游任务中高效加载并使用该模型。

4.6 常见故障排查

问题 1: 安装 PyTorch 时出现版本冲突

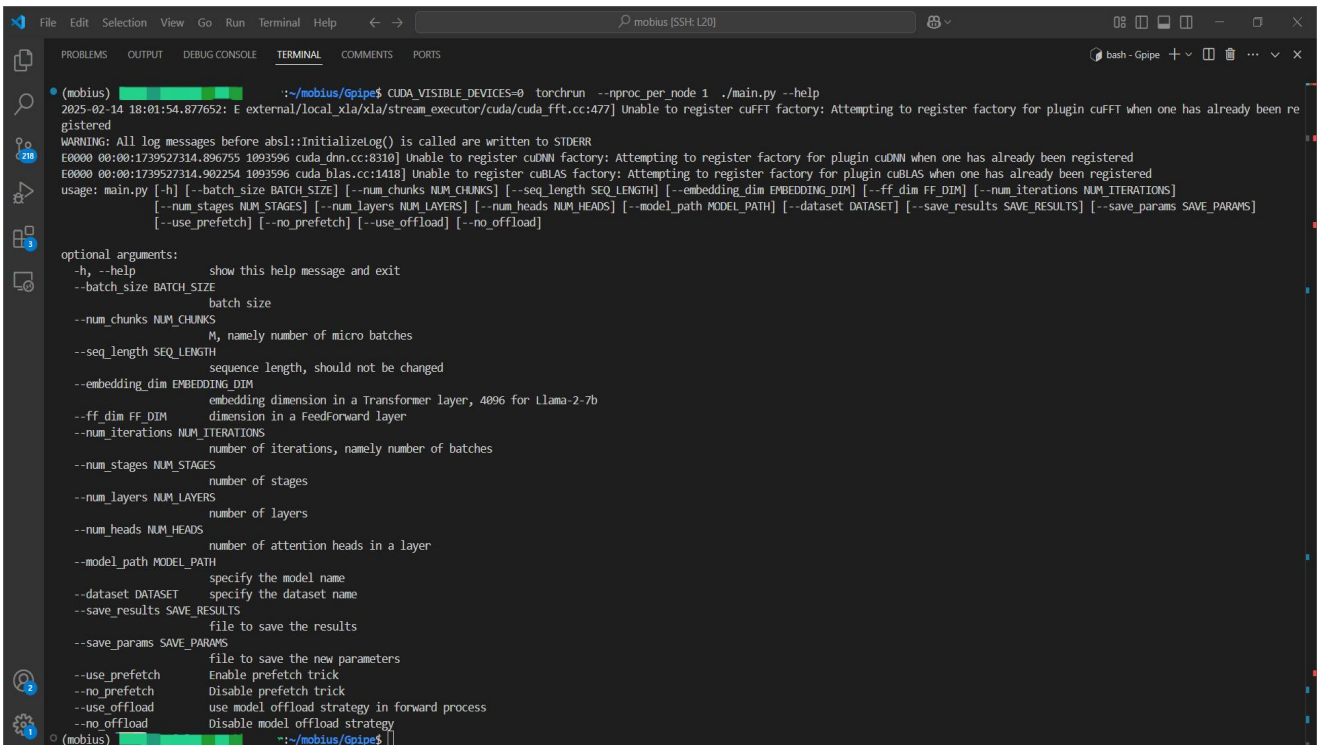
解决方法: 确保 CUDA 和 PyTorch 版本匹配。参考 [PyTorch 官方安装指南](#)。

问题 2: 运行时报错 "NCCL 未初始化"

解决方法: 检查环境变量 LOCAL_RANK 和 WORLD_SIZE 是否已正确设置。

5 软件运行截图

5.1 显示帮助内容



```
(mobius) ~/mobius/gpipe$ CUDA_VISIBLE_DEVICES=0 torchrun --nproc_per_node 1 ./main.py --help
2025-02-14 18:01:54.877652: E external/local_xla/xla/stream_executor/cuda/cuda_fft.cc:477] Unable to register cuFFT factory: Attempting to register factory for plugin cuFFT when one has already been re
gistered
WARNING: All log messages before absl::InitializeLog() is called are written to STDERR
E0000 00:00:1739527314.896755 1093596 cuda.dnn.cc:8310] Unable to register cuDNN factory: Attempting to register factory for plugin cuDNN when one has already been registered
E0000 00:00:1739527314.902254 1093596 cuda.blas.cc:1418] Unable to register cuBLAS factory: Attempting to register factory for plugin cuBLAS when one has already been registered
usage: main.py [-h] [--batch_size BATCH_SIZE] [--num_chunks NUM_CHUNKS] [--seq_length SEQ_LENGTH] [--embedding_dim EMBEDDING_DIM] [--ff_dim FF_DIM] [--num_iterations NUM_ITERATIONS]
               [--num_stages NUM_STAGES] [--num_layers NUM_LAYERS] [--num_heads NUM_HEADS] [--model_path MODEL_PATH] [--dataset DATASET] [--save_results SAVE_RESULTS] [--save_params SAVE_PARAMS]
               [--use_prefetch] [--no_prefetch] [--use_offload] [--no_offload]

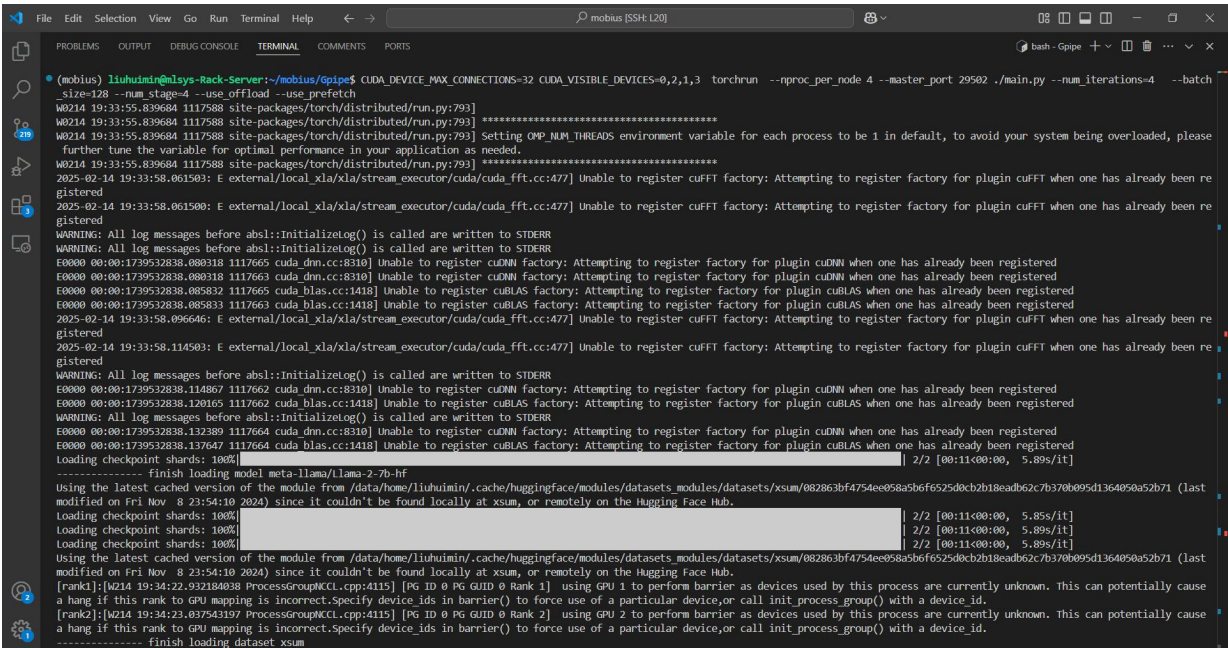
optional arguments:
  -h, --help            show this help message and exit
  --batch_size BATCH_SIZE
                        batch size
  --num_chunks NUM_CHUNKS
                        M, namely number of micro batches
  --seq_length SEQ_LENGTH
                        sequence length, should not be changed
  --embedding_dim EMBEDDING_DIM
                        embedding dimension in a Transformer layer, 4096 for Llama-2-7b
  --ff_dim FF_DIM       dimension in a FeedForward layer
  --num_iterations NUM_ITERATIONS
                        number of iterations, namely number of batches
  --num_stages NUM_STAGES
                        number of stages
  --num_layers NUM_LAYERS
                        number of layers
  --num_heads NUM_HEADS
                        number of attention heads in a layer
  --model_path MODEL_PATH
                        specify the model name
  --dataset DATASET     specify the dataset name
  --save_results SAVE_RESULTS
                        file to save the results
  --save_params SAVE_PARAMS
                        file to save the new parameters
  --use_prefetch        Enable prefetch trick
  --no_prefetch         Disable prefetch trick
  --use_offload         use model offload strategy in forward process
  --no_offload          Disable model offload strategy
```

图 5-1 显示帮助运行图

5.2 运行核心功能

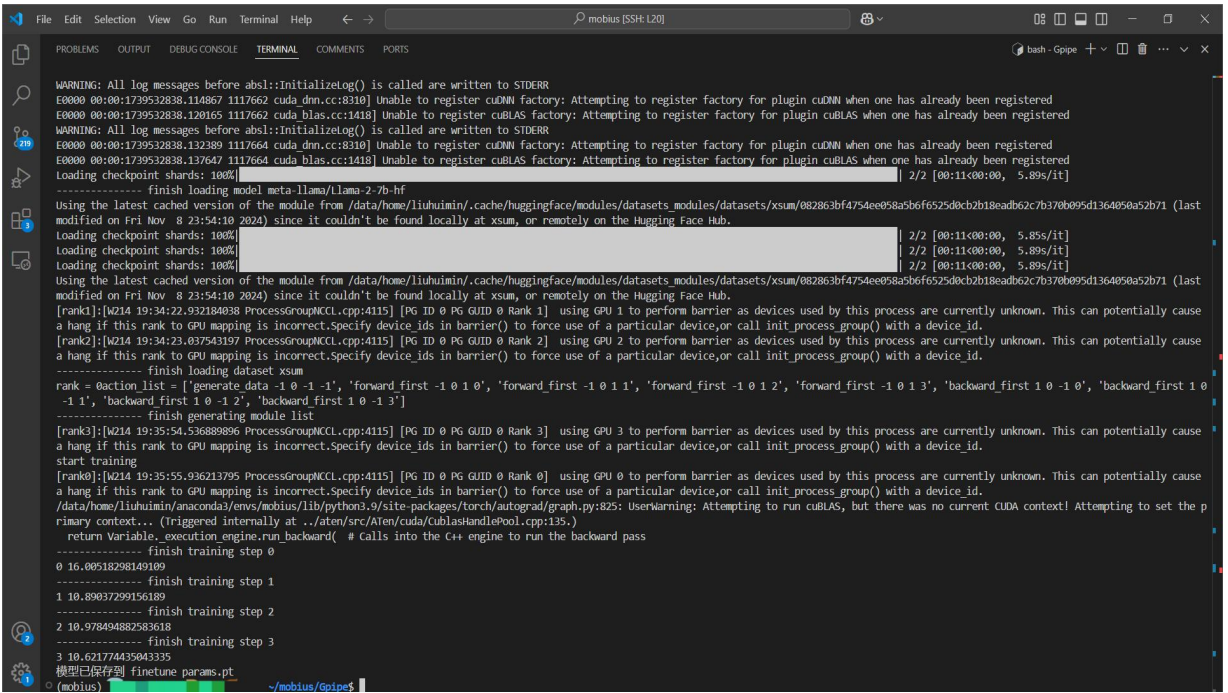
执行运行指令后, 根据终端提示语, 可以看到程序依次完成模型下载、加载数据集、切分模型、迭代训练模型。最终精调完成的模型参数以.pt 文件的形式存储在本地。

基于流水线并行和卸载策略的大模型精调系统 V1.0 软件说明书



```
(mobius) liuhuimin@mlsys-Rack-Server:~/mobius/gpipe$ CUDA_DEVICE_MAX_CONNECTIONS=32 CUDA_VISIBLE_DEVICES=0,2,1,3 torchrun --nproc_per_node 4 --master_port 29502 ./main.py --num_iterations=4 --batch_size=128 --num_stage=4 --use_offload --use_prefetch
W0214 19:33:55.839684 1117588 site-packages/torch/distributed/run.py:793]
W0214 19:33:55.839684 1117588 site-packages/torch/distributed/run.py:793] *****
W0214 19:33:55.839684 1117588 site-packages/torch/distributed/run.py:793] Setting OMP_NUM_THREADS environment variable for each process to be 1 in default, to avoid your system being overloaded, please further tune the variable for optimal performance in your application as needed.
W0214 19:33:55.839684 1117588 site-packages/torch/distributed/run.py:793] *****
2025-02-14 19:33:58.061503: E external/local_xla/xla/stream_executor/cuda/cuda_fft.cc:477] Unable to register cuFFT factory: Attempting to register factory for plugin cuFFT when one has already been registered
2025-02-14 19:33:58.061503: E external/local_xla/xla/stream_executor/cuda/cuda_fft.cc:477] Unable to register cuFFT factory: Attempting to register factory for plugin cuFFT when one has already been registered
WARNING: All log messages before absl::InitializeLog() is called are written to STDERR
WARNING: All log messages before absl::InitializeLog() is called are written to STDERR
E0000 00:00:1739532838.080318 1117665 cuda.dnn.cc:8310] Unable to register cuDNN factory: Attempting to register factory for plugin cuDNN when one has already been registered
E0000 00:00:1739532838.080318 1117663 cuda.dnn.cc:8310] Unable to register cuDNN factory: Attempting to register factory for plugin cuDNN when one has already been registered
E0000 00:00:1739532838.085832 1117665 cuda.blas.cc:1418] Unable to register cuBLAS factory: Attempting to register factory for plugin cuBLAS when one has already been registered
E0000 00:00:1739532838.085833 1117663 cuda.blas.cc:1418] Unable to register cuBLAS factory: Attempting to register factory for plugin cuBLAS when one has already been registered
2025-02-14 19:33:58.096466: E external/local_xla/xla/stream_executor/cuda/cuda_fft.cc:477] Unable to register cuFFT factory: Attempting to register factory for plugin cuFFT when one has already been registered
2025-02-14 19:33:58.114503: E external/local_xla/xla/stream_executor/cuda/cuda_fft.cc:477] Unable to register cuFFT factory: Attempting to register factory for plugin cuFFT when one has already been registered
WARNING: All log messages before absl::InitializeLog() is called are written to STDERR
E0000 00:00:1739532838.114867 1117662 cuda.dnn.cc:8310] Unable to register cuDNN factory: Attempting to register factory for plugin cuDNN when one has already been registered
E0000 00:00:1739532838.120165 1117662 cuda.blas.cc:1418] Unable to register cuBLAS factory: Attempting to register factory for plugin cuBLAS when one has already been registered
WARNING: All log messages before absl::InitializeLog() is called are written to STDERR
E0000 00:00:1739532838.132389 1117664 cuda.dnn.cc:8310] Unable to register cuDNN factory: Attempting to register factory for plugin cuDNN when one has already been registered
E0000 00:00:1739532838.137647 1117664 cuda.blas.cc:1418] Unable to register cuBLAS factory: Attempting to register factory for plugin cuBLAS when one has already been registered
Loading checkpoint shards: 100% | 2/2 [00:11:00:00, 5.89s/it]
----- finish loading model meta-llama/llama-2-7b-hf
Using the latest cached version of the module from /data/home/liuhuimin/.cache/huggingface/modules/datasets_modules/datasets/xsum/082863bf4754ee058a5b6f6525d0cb2b18eadb62c7b370b095d1364050a52b71 (last modified on Fri Nov 8 23:54:10 2024) since it couldn't be found locally at xsum, or remotely on the Hugging Face Hub.
Loading checkpoint shards: 100% | 2/2 [00:11:00:00, 5.85s/it]
Loading checkpoint shards: 100% | 2/2 [00:11:00:00, 5.89s/it]
Loading checkpoint shards: 100% | 2/2 [00:11:00:00, 5.89s/it]
Using the latest cached version of the module from /data/home/liuhuimin/.cache/huggingface/modules/datasets_modules/datasets/xsum/082863bf4754ee058a5b6f6525d0cb2b18eadb62c7b370b095d1364050a52b71 (last modified on Fri Nov 8 23:54:10 2024) since it couldn't be found locally at xsum, or remotely on the Hugging Face Hub.
[rank1]:[W0214 19:34:22.92184038 ProcessGroupMCCl.cpp:4115] [PG ID 0 PG GUID 0 Rank 1] using GPU 1 to perform barrier as devices used by this process are currently unknown. This can potentially cause a hang if this rank to GPU mapping is incorrect.Specify device_ids in barrier() to force use of a particular device,or call init_process_group() with a device id.
[rank2]:[W0214 19:34:23.037543197 ProcessGroupMCCl.cpp:4115] [PG ID 0 PG GUID 0 Rank 2] using GPU 2 to perform barrier as devices used by this process are currently unknown. This can potentially cause a hang if this rank to GPU mapping is incorrect.Specify device_ids in barrier() to force use of a particular device,or call init_process_group() with a device id.
----- finish loading dataset xsum
rank = 0 action_list = ['generate_data -1 0 -1 -1', 'forward_first -1 0 1 0', 'forward_first -1 0 1 1', 'forward_first -1 0 1 2', 'forward_first -1 0 1 3', 'backward_first 1 0 -1 0', 'backward_first 1 0 -1 1', 'backward_first 1 0 -1 2', 'backward_first 1 0 -1 3']
----- finish generating module list
[rank3]:[W0214 19:35:54.536889896 ProcessGroupMCCl.cpp:4115] [PG ID 0 PG GUID 0 Rank 3] using GPU 3 to perform barrier as devices used by this process are currently unknown. This can potentially cause a hang if this rank to GPU mapping is incorrect.Specify device_ids in barrier() to force use of a particular device,or call init_process_group() with a device id.
start training
[rank0]:[W0214 19:35:55.936213795 ProcessGroupMCCl.cpp:4115] [PG ID 0 PG GUID 0 Rank 0] using GPU 0 to perform barrier as devices used by this process are currently unknown. This can potentially cause a hang if this rank to GPU mapping is incorrect.Specify device_ids in barrier() to force use of a particular device,or call init_process_group() with a device id.
/data/home/liuhuimin/anaconda3/envs/mobius/lib/python3.9/site-packages/torch/autograd/graph.py:825: UserWarning: Attempting to run cuBLAS, but there was no current CUDA context! Attempting to set the primary context... (Triggered internally at ../aten/src/ATen/cuda/CudaLashHandlePool.cpp:135.)
return Variable._execution_engine.run_backward( # calls into the C++ engine to run the backward pass
----- finish training step 0
0 16.00518298149109
----- finish training step 1
1 10.89037729156189
----- finish training step 2
2 10.978494882583618
----- finish training step 3
3 10.621774435043335
模型已保存到 finetune_params.pt
(mobius)
```

图 5-2-a 核心功能运行图



```
WARNING: All log messages before absl::InitializeLog() is called are written to STDERR
E0000 00:00:1739532838.114867 1117662 cuda.dnn.cc:8310] Unable to register cuDNN factory: Attempting to register factory for plugin cuDNN when one has already been registered
E0000 00:00:1739532838.120165 1117662 cuda.blas.cc:1418] Unable to register cuBLAS factory: Attempting to register factory for plugin cuBLAS when one has already been registered
WARNING: All log messages before absl::InitializeLog() is called are written to STDERR
E0000 00:00:1739532838.132389 1117664 cuda.dnn.cc:8310] Unable to register cuDNN factory: Attempting to register factory for plugin cuDNN when one has already been registered
E0000 00:00:1739532838.137647 1117664 cuda.blas.cc:1418] Unable to register cuBLAS factory: Attempting to register factory for plugin cuBLAS when one has already been registered
Loading checkpoint shards: 100% | 2/2 [00:11:00:00, 5.89s/it]
----- finish loading model meta-llama/llama-2-7b-hf
Using the latest cached version of the module from /data/home/liuhuimin/.cache/huggingface/modules/datasets_modules/datasets/xsum/082863bf4754ee058a5b6f6525d0cb2b18eadb62c7b370b095d1364050a52b71 (last modified on Fri Nov 8 23:54:10 2024) since it couldn't be found locally at xsum, or remotely on the Hugging Face Hub.
Loading checkpoint shards: 100% | 2/2 [00:11:00:00, 5.85s/it]
Loading checkpoint shards: 100% | 2/2 [00:11:00:00, 5.89s/it]
Loading checkpoint shards: 100% | 2/2 [00:11:00:00, 5.89s/it]
Using the latest cached version of the module from /data/home/liuhuimin/.cache/huggingface/modules/datasets_modules/datasets/xsum/082863bf4754ee058a5b6f6525d0cb2b18eadb62c7b370b095d1364050a52b71 (last modified on Fri Nov 8 23:54:10 2024) since it couldn't be found locally at xsum, or remotely on the Hugging Face Hub.
[rank1]:[W0214 19:34:22.92184038 ProcessGroupMCCl.cpp:4115] [PG ID 0 PG GUID 0 Rank 1] using GPU 1 to perform barrier as devices used by this process are currently unknown. This can potentially cause a hang if this rank to GPU mapping is incorrect.Specify device_ids in barrier() to force use of a particular device,or call init_process_group() with a device id.
[rank2]:[W0214 19:34:23.037543197 ProcessGroupMCCl.cpp:4115] [PG ID 0 PG GUID 0 Rank 2] using GPU 2 to perform barrier as devices used by this process are currently unknown. This can potentially cause a hang if this rank to GPU mapping is incorrect.Specify device_ids in barrier() to force use of a particular device,or call init_process_group() with a device id.
----- finish loading dataset xsum
rank = 0 action_list = ['generate_data -1 0 -1 -1', 'forward_first -1 0 1 0', 'forward_first -1 0 1 1', 'forward_first -1 0 1 2', 'forward_first -1 0 1 3', 'backward_first 1 0 -1 0', 'backward_first 1 0 -1 1', 'backward_first 1 0 -1 2', 'backward_first 1 0 -1 3']
----- finish generating module list
[rank3]:[W0214 19:35:54.536889896 ProcessGroupMCCl.cpp:4115] [PG ID 0 PG GUID 0 Rank 3] using GPU 3 to perform barrier as devices used by this process are currently unknown. This can potentially cause a hang if this rank to GPU mapping is incorrect.Specify device_ids in barrier() to force use of a particular device,or call init_process_group() with a device id.
start training
[rank0]:[W0214 19:35:55.936213795 ProcessGroupMCCl.cpp:4115] [PG ID 0 PG GUID 0 Rank 0] using GPU 0 to perform barrier as devices used by this process are currently unknown. This can potentially cause a hang if this rank to GPU mapping is incorrect.Specify device_ids in barrier() to force use of a particular device,or call init_process_group() with a device id.
/data/home/liuhuimin/anaconda3/envs/mobius/lib/python3.9/site-packages/torch/autograd/graph.py:825: UserWarning: Attempting to run cuBLAS, but there was no current CUDA context! Attempting to set the primary context... (Triggered internally at ../aten/src/ATen/cuda/CudaLashHandlePool.cpp:135.)
return Variable._execution_engine.run_backward( # calls into the C++ engine to run the backward pass
----- finish training step 0
0 16.00518298149109
----- finish training step 1
1 10.89037729156189
----- finish training step 2
2 10.978494882583618
----- finish training step 3
3 10.621774435043335
模型已保存到 finetune_params.pt
(mobius)
```

图 5-2-b(接 5-2-a) 核心功能运行图

最终生成的模型参数文件:

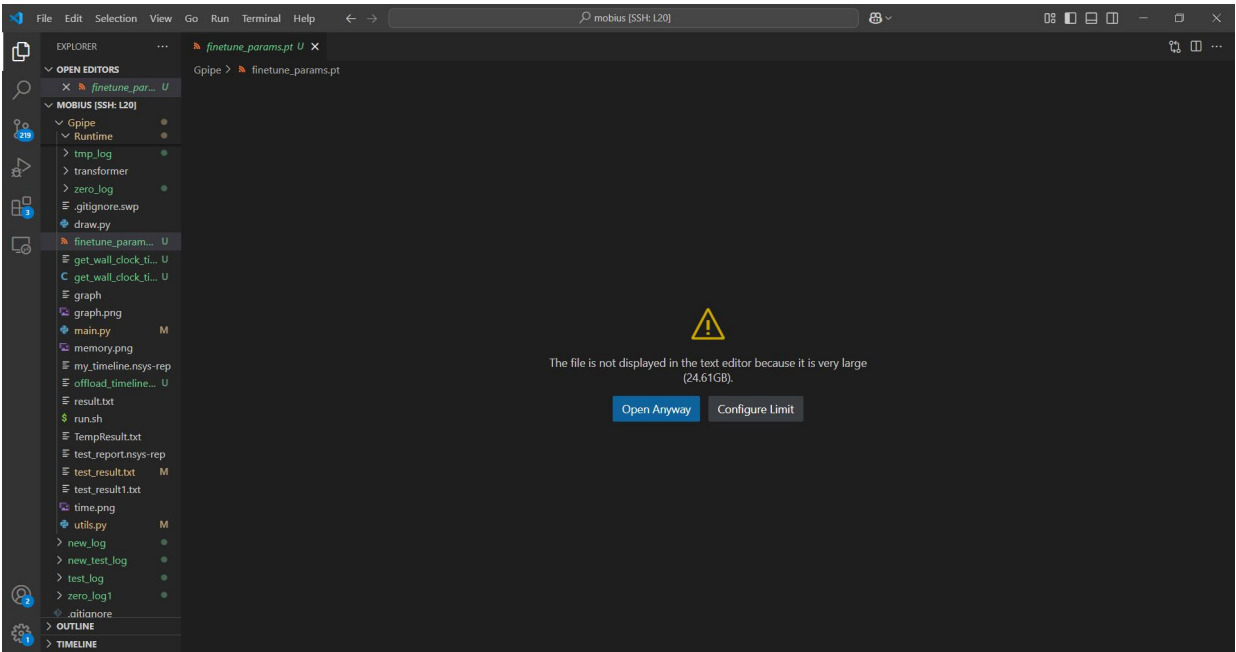


图 5-3 程序运行结果图