LLaMA | 部署教程

PhiltreX
2 年前

LLaMA在Windows环境下的安装部署教程

LLaMA的安装过程其实非常简单，只需要几条CMD命令行即可完成。其实个人感觉效果不如ChatGPT，而且对硬件要求较高，本站并不推荐个人部署。

介绍

关于什么是LLaMa，详细情况可以看下面这篇文章。

LLaMA | 开源语言模型

Meta最新开源语言模型LLaMA惨遭泄漏，程序员狂喜。本站提供部署教程、泄露版模型免磁力国内网盘下载，该语言模型据说比openAI的ChatGPT能力更强的，属于是本地版免费使用的ChatGPT了。

openAI维基百科 1

Git安装

该软件的作用是拉取远程Github仓库代码，可以让你的Stable Diffusion远程实时更新，及时使用全新功能。如果您的电脑内还没有安装Git，请参考下面的文章来安装Git。

Git | Windows安装教程

本文将详细介绍如何在Windows系统上安装Git，包括下载、安装步骤以及常见问题解决方案。我们将确保您在最短的时间内完成Git安装并开始使用Git进行版本控制。

openAI维基百科 0

Conda安装

Conda有Anaconda和Miniconda两个版本可以选择，Anaconda属于完整版，默认包含很多库，但我们用不到，本教程使用的是Miniconda，所以推荐大家也选择Miniconda。

备注：如果您想安装Anaconda也是可以的，教程所使用的命令行完全一样，不用担心不适配的问题。

Anaconda | Miniconda | Windows安装教程

在本文中，我们将向您展示如何在Windows系统中安装Conda，Conda是一个强大的Python包管理工具，可以方便地管理和创建虚拟环境，并安装所需的Python包。

openAI维基百科 8

环境搭建

首先，我们在电脑本地磁盘中创建一个文件夹，用来存放该项目。本站在电脑的D盘创建了一个openai.wiki文件夹用来存放数据，完整的路径为 `D:/openai.wiki` 目录。

指定路径

我们打开CMD，首先执行如下命令，使CMD的盘符为我们的工作路径。

```
1.  cd /d D:\openai.wiki
```

执行如下命令，获取LLaMA官方远程Github仓库项目文件到本地。

```
1.   git clone https://github.com/facebookresearch/
```

此时，会在 `D:/openai.wiki` 目录下看到已经自动创建了一个名字llama的文件夹，我们在CMD内执行下面的命令，进行到刚刚获取的llama项目。

创建Conda环境

```
1.  conda create -n name llama
```

激活Conda环境

在CMD窗口内输入如下命令，激活我们刚刚创建的环境。

```
1.  conda activate llama
```

进入工作目录，CMD中执行如下合集。

```
1.  cd llama
```

执行如下命令，安装项目依赖，此时将会自动安装使用LLaMA的所有依赖库，时间因网络环境而定，可能会比较久。

⚠注意：如果遇到安装报错等问题，可以尝试按键盘的上方向键，然后按回车即可重试。

依赖库安装

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

```
pip install -e .
```

模型下载

关于模型下载，可以前往下面的文章去下载，本站已提供国内网盘下载地址，不限速、不用客户端、免广告、免登录。

> **LLaMA | 模型下载**
>
> Meta（Facebook）开源语言模型LLaMA泄漏版国内网盘下载，该语言模型据说比openAI的ChatGPT能力更强。虽然是开源语言模型，但仅可做为科学研究使用，本站已为大家整理好国内网盘下载。
>
> openAI维基百科 5

使用方法

```python
from llama import LLaMA, ModelArgs, Tokenizer,

os.environ['RANK'] = '0'
os.environ['WORLD_SIZE'] = '1'
os.environ['MP'] = '2'
os.environ['MASTER_ADDR'] = '127.0.0.1'
os.environ['MASTER_PORT'] = '2223'

def setup_model_parallel() -> Tuple[int, int]:
    local_rank = int(os.environ.get("LOCAL_RANK
    world_size = 2

    torch.distributed.init_process_group("gloo"
    initialize_model_parallel(world_size)
    torch.cuda.set_device(local_rank)

    # seed must be the same in all processes
    torch.manual_seed(1)
    return local_rank, world_size
```

如果您希望不使用GPU，而是CPU来运行LLaMA，可以尝试根据下面这篇教程部署。

markasoftware/llama-cpu：在CPU上运行的Facebook的LLaMa模型的分支 (github.com)

LLaMA # 语言模型

3

4,783 18

LLaMA | 开源语言模型

上一篇

LLaMA | 模型下载

下一篇

评论 (18)

昵称（不填则显示匿名）

邮箱

提交

kiritoyu

torchrun –nproc_per_node 1 example.py –ckpt_dir /home/ubuntu/llama-model/llama-7b-hf/7B –tokenizer_path /home/ubuntu/llama-model/llama-7b-hf/tokenizer.model
官方命令中的7B文件去了哪?

2 年前 浙江省 回复

leo

@kiritoyu 同问，楼主有解决办法没?

2 年前 湖南省 回复

南极海豹

在“创建Conda环境”这一步报错，尝试按网上搜索的解决方案，换了清华源（也尝试过替换https为http）、恢复默认、关闭VPN隧道，都还是没能解决，求助。
具体指令反馈信息如下：

E:\openai.wiki>conda create -n name llama
Collecting package metadata (current_repodata.json): done
Solving environment: failed with repodata from current_repodata.json, will retry with next repodata source.
Collecting package metadata (repodata.json): done
Solving environment: failed
PackagesNotFoundError: The following packages are not available from current channels:
– llama
Current channels:
– http://mirrors.tuna.tsinghua.edu.cn/anaconda/cloud/conda-forge/win-64
– http://mirrors.tuna.tsinghua.edu.cn/anaconda/cloud/conda-forge/noarch
– http://mirrors.tuna.tsinghua.edu.cn/anaconda/cloud/pytorch/win-64
– http://mirrors.tuna.tsinghua.edu.cn/anaconda/cloud/pytorch/noarch
– http://mirrors.tuna.tsinghua.edu.cn/anaconda/pkgs/main/win-64
– http://mirrors.tuna.tsinghua.edu.cn/anaconda/pkgs/main/noarch
– http://mirrors.tuna.tsinghua.edu.cn/anaconda/pkgs/free/win-64
– http://mirrors.tuna.tsinghua.edu.cn/anaconda/pkgs/free/noarch
– https://repo.anaconda.com/pkgs/main/win-64
– https://repo.anaconda.com/pkgs/main/noarch
– https://repo.anaconda.com/pkgs/r/win-64
– https://repo.anaconda.com/pkgs/r/noarch
– https://repo.anaconda.com/pkgs/msys2/win-64
– https://repo.anaconda.com/pkgs/msys2/noarch

looking for, navigate to

https://anaconda.org

and use the search bar at the top of the page.

2 年前 浙江省 回复

PhiltreX

@南极海豹 好像镜像源还不支持LLaMA，换成默认的吧。

2 年前 浙江省 回复

南极海豹

@PhiltreX 感谢指导，之前在网上也搜到有这个说法，已尝试恢复默认源，但是涛声依旧，问题仍未能排除。

2 年前 浙江省 回复

gh

@南极海豹 你好，请问你找到解决的方法了吗

2 年前 未知地区 回复

PhiltreX

@gh 建议魔法上网

2 年前 陕西省 回复

南极海豹

@gh 还没解决，暂时搁置了。

2 年前 浙江省 回复

gh

@南极海豹 你好，请问你现在解决这个问题了吗

2 年前 未知地区 回复

jczaza

@gh 都试过了，我的还不行

2 年前 浙江省 回复

GPU

这个泄漏版本还是需要通过Edu邮箱申请资格吗?

2 年前 湖南省 回复

PhiltreX

@GPU 不需要，直接可以用。

2 年前 陕西省 回复

xjq284

怎么把bin文件合一起?

2 年前 浙江省 回复

123

@xjq284 怎么合?

1 年前 浙江省 回复

会飞的蛇

请问部署完成后如何使用?

2 年前 浙江省 回复

PhiltreX

@会飞的蛇 有示例代码呀

2 年前 浙江省 回复

alhua

@PhiltreX 请教大佬，最后的示例代码怎么运行呢？代码

1 年前 湖南省 回复

PhiltreX

@alhua 讲真的，我没跑起来，这个需要的配置不是家用电脑能够使用的。这个示例代码是官方的。

1 年前 浙江省 回复

猜你喜欢

Baichuan2 | 开源语言模型百川2代

llama2.c | Baby LLaMA

LLaMA-2 | FreeWilly2微调模型