THUDM-chatglm-6b 部署命令

Stephen CUI

2023-09-12

目录

1 安装 Python 与环境创建

这个命令可以安装 Python 其他版本，系统带有 Python3.8，可以不安装
`sudo apt-get install python3.10`，其中 3.10 可以换成任意其他的 Python 版本。

```
% 删除默认的软连接，因为默认 python 链接到 python2
sudo rm -rf /usr/bin/python3
% 将python 默认设置为 Python3.8，Python3.8 应该是 Ubuntu20 默认安装的版本
sudo ln -s /usr/bin/python3.8 /usr/bin/python
```

安装 Python3 的 pip
`sudo apt-get install python3-pip`
可以不运行，除非版本不够（升级至最新的 pip）：
`pip install --upgrade pip`
从 Python 3.6 开始，创建虚拟环境的推荐方法是使用 venv 模块。因此我们先要安装提供 venv 模块的 python3-venv 软件包。运行命令：

```
sudo apt install python3-venv
```

切换项目目录。在目录中，运行

```
python -m venv llm
```

命令来创建新的虚拟环境，llm 可以更改为任意环境名。

要开始使用此虚拟环境，您需要通过运行 activate 脚本将其激活。source 命令将会加载 python 的虚拟环境

```
source llm/bin/activate
% conda activate llm
```

一旦激活，虚拟环境的 bin 目录将添加到 PATH 变量的开头。此外，您的 Shell 提示符也会更改，并且会显示您当前正在使用的虚拟环境的名称。

使用 Python 包管理器 pip 安装 python 包，在虚拟环境中，可以使用命令 pip 代替 pip3，并使用 python 代替 python3

```
%  测试 numpy 的安装
pip install numpy
```

完成工作后停用虚拟环境，只需键入 deactivate，您将返回到常规 shell。

```
deactivate
```

2 下载模型与代码

2.1 下载权重

2.1.1 方式一

直接从 Hugging Face 的网站上下载，地址为：THUDM/chatglm2-6b

2.1.2 手动下载权重的话不需要运行以下步骤

```
sudo apt-get update
```

安装 git-lfs 因为较大文件 git 不管理，需要使用 git-lfs 下载，但是下载较慢，建议在清华云盘中下载。

```
sudo apt-get install git-lfs
```

2.1.3 方式二(推荐)

从清华云盘中下载模型权重：地址为：chatglm-6b

2.2 下载运行代码

注意：**必须在 Git Bash 中运行**
需要安装 git，下载地址为：git
如果需要，可以设置用户名和邮箱：

```
git config --global user.name "FIRST_NAME LAST_NAME"
git config --global user.email "MY_NAME@example.com"
```

如果有代理的话，设置代理的端口（否则连接错误）

```
git config --global http.proxy http://127.0.0.1:port
```

port 设置为机子的端口。

执行如下命令可以下载代码，代码会保存在当前目录下的 chatglm-6b 文件夹中，可以移动到权重相同文件夹下：

```
GIT_LFS_SKIP_SMUDGE=1 git clone https://huggingface.co/THUDM/chatglm-6b
```

3 本地文件上传至服务器

3.1 基于 vscode 的拓展 SFTP

安装完 SFTP 后键盘按 `ctrl+shift+p` 输入 `> SFTP: config` 回车进入 sftp.json 文件。

使用 SFTP 传输本地文件的配置：remotePath，host，password，name 按实际修改，ignore为不传输文件类型或文件

```
{
    "name": "ECS-aeqk",
    "host": "IP地址",
    "protocol": "sftp",
    "port": 22,
    "username": "用户名",
    "password": "你的密码",
    "remotePath": "llm/glm",
    "uploadOnSave": true,
    "useTempFile": false,
    "openSsh": false,
    "ignore": [
        "**/.vscode/**",
        "**/.git/**",
        "**/.DS_Store/**",
        "**/__pycache__/**"
    ]
}
```

4 安装运行依赖的包

运行如下代码：

```
pip install -r requirements.txt -i https://pypi.tuna.tsinghua.edu.cn/simple
```

点击获取 requirements.txt 地址

4.1 运行测试代码

可以在 Python 交互式界面中运行以下命令（逐行）（进行交互式页面 python -i）

```
from transformers import AutoTokenizer, AutoModel
tokenizer = AutoTokenizer.from_pretrained("./", trust_remote_code=True, revision='v1.1.0')
model = AutoModel.from_pretrained("./", trust_remote_code=True, revision='v1.1.0').half().cuda()
model = model.eval()
response, history = model.chat(tokenizer, "你好", history=[])
print(response)
response, history = model.chat(tokenizer, "如何学习语言大模型", history=history)
print(response)
```

也可以直接写入文件，比如 running_test.py，然后运行

```
python running_test.py
```

5 网页版

5.1 基于 streamlit 推荐

安装网页版本所需的包

```
pip install streamlit -i https://pypi.tuna.tsinghua.edu.cn/simple
pip install streamlit-chat -i https://pypi.tuna.tsinghua.edu.cn/simple
```

安装完成，直接运行下面命令就可以网页版：

```
streamlit run web_demo2.py
```

5.2 基于 gradio 国内不推荐

Gradio 是通过友好的 Web 界面演示机器学习模型的最快方式，以便任何人都可以在任何地方使用它!

```
pip install gradio mdtex2html
```

Gradio 直接使用 Python 运行，国内访问速度慢。

```
python web_deom.py
```

6 微调模型

6.1 微调所需依赖包

```
pip install rouge_chinese nltk jieba datasets -i https://pypi.tuna.tsinghua.edu.cn/simple
```

6.2 软件依赖

cuda 更多版本

更新 cuda 使用以下命令

```
wget https://developer.download.nvidia.com/compute/cuda/11.7.1/local_installers/cuda_11.7.1_515.65
```

```
sudo sh cuda_11.7.0_515.43.04_linux.run
```

如果你安装了驱动，在安装的时候需要取消驱动的勾选：

```
apt install ubuntu-drivers-common
ubuntu-drivers devices
```

执行微调的文件，用以下命令：

```
bash train.sh
```

6.3 微调后的 API

使用微调后的权重，已经有写好的文件可以运行（web_demo.py），只需要将 web_demo.sh 文件的参数 `--model_name_or_path` 改成我们本地对应的文件夹或者路径即可。

如果需要打开公网接口，需要将 `demo.queue().launch(share=False, inbrowser=True)` 中的 `share` 参数设置为 `True`，该行位置在 web_demo.py Line 162.

打开公网接口，会出现一些问题，解决见 常见问题。

7 其他

7.1 常见问题

- RuntimeError: Internal: src/sentencepiece_processor.cc(1101) [model_proto->ParseFromArray(serialized.data(), serialized.size())] 可以重新在 Hugging Face 上下载文件 下载运行代码。
- RuntimeError: Default process group has not been initialized, please make sure to call init_process_group. 将 transformers 降级至 4.27.1，使用 `pip install transformers==4.27.1`
- Could not create share link. Missing file: /xxxx/frpc_linux_amd64_v0.2. 解决步骤如下

  1. 下载必要的文件 frpc_linux_amd64，会被识别为病毒，需要在防火墙开启权限并执行
  2. 将下载的文件名修改为 frpc_linux_amd64_w0.2
  3. 将文件移动到此路径 /root/username/lib/pthon3.8/site-packages/gradio 下，username 是你自己的用户，移动命令为

     ```
     mv frpc_linux_amd64_w0.2 /root/username/lib/pthon3.8/site-packages/gradio
     ```
  4. 给 frpc_linux_amd64_w0.2 添加一个可执行权限

     ```
     chmod +x /root/username/lib/pthon3.8/site-packages/gradio/frpc_linux_amd64_w0.2
     ```

7.2 参考链接

如何在 Ubuntu 18.04 创建 Python 虚拟环境
yizhongw/self-instruct
清华大学GLM-6B
模型权重所在云盘位置