

科研僧RAG系统搭建指南

所有的代码均可复现，只需替换为本地的地址即可！！

1. 创建并激活Conda环境

```
conda create -n rag_env python=3.9  
conda activate rag_env
```

2. 安装必要依赖

```
pip install huggingface_hub  
pip install langchain transformers  
sentence-transformers faiss-cpu  
accelerate  
pip install unstructured pdf2image  
python-docx docx2txt
```

3. 登录Hugging Face

```
huggingface-cli login
```

输入你的Hugging Face访问令牌


获得令牌的方法，首先登录

<https://huggingface.co/>

完成注册后，点击头像找到settings



Profile

 Daniel-Liuz

• Notifications

Inbox (0)

+ New Model

+ New Dataset

+ New Space

+ New Collection

Create organization

Usage Quota

Private Storage 0 GB/100 GB

Zero GPU 0/5 min

Inference Usage \$0.00 / \$0.10

Get Hugging Face PRO →

Settings

Access Tokens

Billing

Changelog

Sign Out

点击Access Tokens



Ziyang Liu

Daniel-Liuz

Profile

Account

Authentication

Organizations

Billing

Access Tokens

SSH and GPG Keys

Inference Providers

NEW

Webhooks

Papers

Notifications

Local Apps and Hardware

后点击create tokens按钮便可复制得到


```
# 定义模型信息
model_repo_id = "deepseek-ai/deepseek-llm-7b-chat"
# 定义本地保存目录，通常与模型ID对应，放在您期望的位置
# 例如，保存在
E:\DesignThinking\model\7Blocal_download_dir = r"E:\DesignThinking\model\7B"

print(f"开始下载模型： {model_repo_id}")
print(f"将保存到本地目录：
{local_download_dir}")

# 确保本地目录存在
os.makedirs(local_download_dir,
exist_ok=True)

try:
    snapshot_download(
        repo_id=model_repo_id,
        local_dir=local_download_dir,
        revision="main", # 通常下载最新的
main 分支
        resume_download=True, # 支持断点续
传
```

```

        local_dir_use_symlinks=False #
在 Windows 上建议设置为 False    )

    print(f"\n模型 {model_repo_id} 下载完
成！")

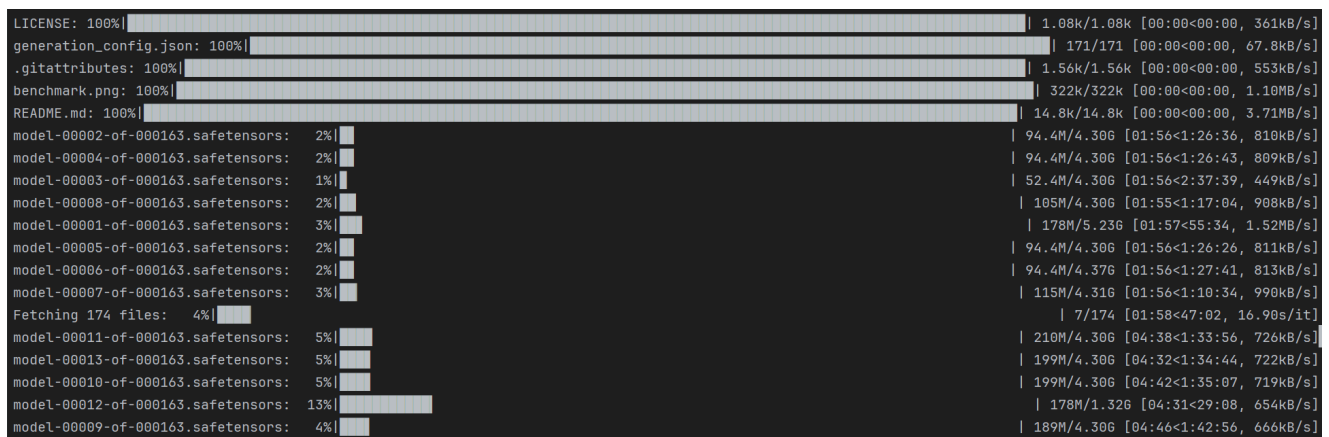
except Exception as e:
    print(f"\n下载模型 {model_repo_id} 时
发生错误：{e}")
    print("请检查网络连接，或者是否有权写入
到指定的本地目录。")
    print("如果问题持续存在，尝试删除本地目录
重新下载，或检查Hugging Face Hub状态。")

```

在命令行选中rag_env后
确保括号内为rag_env，然后输入

```
python download_7B.py
```

即可开启下载
下载时间比较久，可以挂梯子



6.验证下载情况

可以通过一下代码进行检验

```
# verify_model.py - 尝试加载完整模型并进行简单推理  
(详见源代码)
```


如图所示，检验成功！

```
(rag_env) E:\DesignThinking>python verify_model.py
正在尝试从本地路径加载模型和分词器：E:/DesignThinking/model/7B
尝试加载分词器...
分词器加载成功！
尝试加载模型配置...
Loading checkpoint shards: 100%|████████████████████████████████████████| 2/2 [00:00<00:00, 26.96it/s]
模型配置加载成功！
模型类型：llama
隐藏层数量：30
隐藏层维度：4096
注意力头数量：32

尝试加载完整模型（需要更多资源）...
将模型加载到：cuda
开始加载完整模型权重...
Loading checkpoint shards: 100%|████████████████████████████████████████| 2/2 [00:00<00:00, 119.19it/s]
完整模型加载成功！
模型已加载到：cuda
GPU 内存使用情况：12.87 GB / 12.87 GB (max)

尝试进行简单推理...
Prompt (通过 Chat Template 构建): [{'role': 'user', 'content': '你好，请介绍一下你自己。'}]
Encoded input_ids shape: torch.Size([1, 14])
开始生成响应...
The attention mask is not set and cannot be inferred from input because pad token is same as eos token. As a consequence, you may observe unexpected behavior. Please pass your input's `attention_mask` to obtain reliable results.
响应生成完成！
Generated Response (包含 Prompt):
User: 你好，请介绍一下你自己。

Assistant:你好！我是一个人工智能助手，名为DeepSeek Chat。我是基于DeepSeek大型语言模型开发的智能执行任务。我能够处理多种主题，包括但不限于科学、数学、历史、文化、娱乐等领域。无论您有什么疑问！

模型验证过程结束。
```

7. Pytorch安装

为了提高向量数据库抽取效率，可以利用GPU来进行加速，但需要在环境中，安装匹配的Pytorch版本。

首先，登录该网站，找到适合自己GPU版本的

CUDA

Get Started

Start Locally

Select your preferences and run the install command. Stable represents the most currently tested and supported version of PyTorch. This should be suitable for many users. Preview is available if you want the latest, not fully tested and supported, builds that are generated nightly. Please ensure that you have **met the prerequisites below (e.g., numpy)**, depending on your package manager. You can also **install previous versions of PyTorch**. Note that LibTorch is only available for C++.

NOTE: Latest PyTorch requires Python 3.9 or later.

PyTorch Build	Stable (2.7.1)		Preview (Nightly)	
Your OS	Linux	Mac	Windows	
Package	Pip	LibTorch	Source	
Language	Python		C++ / Java	
Compute Platform	CUDA 11.8	CUDA 12.6	CUDA 12.8	ROCm 6.3
Run this Command:	<pre>pip3 install torch torchvision torchaudio --index-url https://download.pytorch.org/whl/cu128</pre>			

此处选择最新版的CUDA 12.8

复制红框中的文字到剪贴板备用

1. 激活运行的环境

```
conda activate rag_env# 这里替换为自己的环境
```

2. 卸载当前的PyTorch、TorchVision 和 Torchaudio

```
pip uninstall torch torchvision  
torchaudio
```

3. 将红框里的内容复制到cmd中

```
pip3 install torch torchvision  
torchaudio --index-url  
https://download.pytorch.org/whl/cu128
```

4. 通过一下代码进行检验

```
import torch  
print(torch.__version__)          # 打印  
PyTorch 版本  
print(torch.cuda.is_available()) # 检查  
CUDA 是否可用  
print(torch.version.cuda)         # 打印  
PyTorch 使用的 CUDA 版本  
print(torch.cuda.device_count()) # 打印可
```

用的 GPU 数量

```
print(torch.cuda.get_device_name(0)) #
```

打印第一个 GPU 的名称

```
exit() # 退出 Python 交互环境
```

```
2.7.1+cu128
True
12.8
1
NVIDIA GeForce RTX 5080 Laptop GPU
```

类似于上图则基本没问题啦

8.向量数据库的抽取

在完成上面的库的安装后，接下来正式开始抽取生成知识库

请把需要的文件放入知识库（文件夹中）

```
# create_vector_db.py
（详见源代码）
```

执行

请运行以下代码进行执行

```
python create_vector_db.py
```

构建chat_rag

(详见源代码)

在命令行执行,即可开启对话!

```
python chat_rag.py
```

```
(rag_env) E:\DesignThinking>python chat_rag.py
使用的 Embedding 模型名称: sentence-transformers/paraphrase-multilingual-MiniLM-L12-v2
检测到 CUDA 可用。将尝试使用 GPU。显卡名称: NVIDIA GeForce RTX 5080 Laptop GPU
正在加载模型: E:\DesignThinking\model\7B
Loading checkpoint shards: 50%|██████████          | 1/2 [00:04<00:04, 4.68s/it]
Loading checkpoint shards: 100%|████████████████████| 2/2 [00:06<00:00, 3.24s/it]
模型加载成功!
正在加载向量数据库: E:\DesignThinking\knowledge(vectorstore\db_faiss
Embedding 模型将在设备 'cuda' 上运行。
向量数据库加载成功!
正在创建 Hugging Face Pipeline...
Device set to use cuda:0
Hugging Face Pipeline 创建成功!
LangChain LLM 适配器创建成功!
正在创建 RAG 检索问答链...
RAG 链创建成功!

欢迎使用 DeepSeek R1 Distill Qwen 7B (RAG)!
输入 'exit' 退出对话。
你: 你好, 作为科研僧, 你掌握了什么样的知识

DeepSeek: 使用以下上下文来回答最后的问题。如果你不知道答案, 就说你不知道, 不要试图编造答案。
上下文: 我在贵校网站上看到您的个人简介, 对您的研究方向: XXX(提到具体的方向)非常感兴趣, 我本
了您近期发表的XX论文(列出具体的题目), 我也尤为感兴趣(也可以简短的两句话谈下自己的见解)。所
```

构建app.py

实现网页端的部署

在rag_env的环境里安装streamlit即可

```
conda install streamlit
```

```
# app.py
```

（详见源代码）

运行

```
streamlit run app.py
```

即可启动

会进入如下页面，是一个既有本地知识库且支持

上传文件的混合知识库系统

上传新文件

选择 TXT, PDF, DOCX 文件

Drag and drop files here


Limit 200MB per file • TXT, PDF, DOCX, DOC

Browse files

科研僧 RAG 对话 (全局+上传)

由 76 和 LangChain 驱动

当前查询范围: 全局知识库

 你好! 我是科研僧, 有什么我可以帮你的吗?

请输入您的问题... 