

**数字媒体技术实验报告**

题 目 跨媒体的检索与分析

学 院 计算学部

专 业 软件工程

学 号 120L022109

学 生 李世轩

任 课 教 师 刘绍辉

哈尔滨工业大学计算机科学与技术学院

2023.3

**实验二:跨媒体的检索与分析**

1. **实验内容或者文献情况介绍**

**实验环境：**

**操作系统 windows 11**

**编程语言 python3.8（Anaconda虚拟环境）**

**编程IDE PyCharm 2023.1 (Professional Edition)**

**实验内容：**

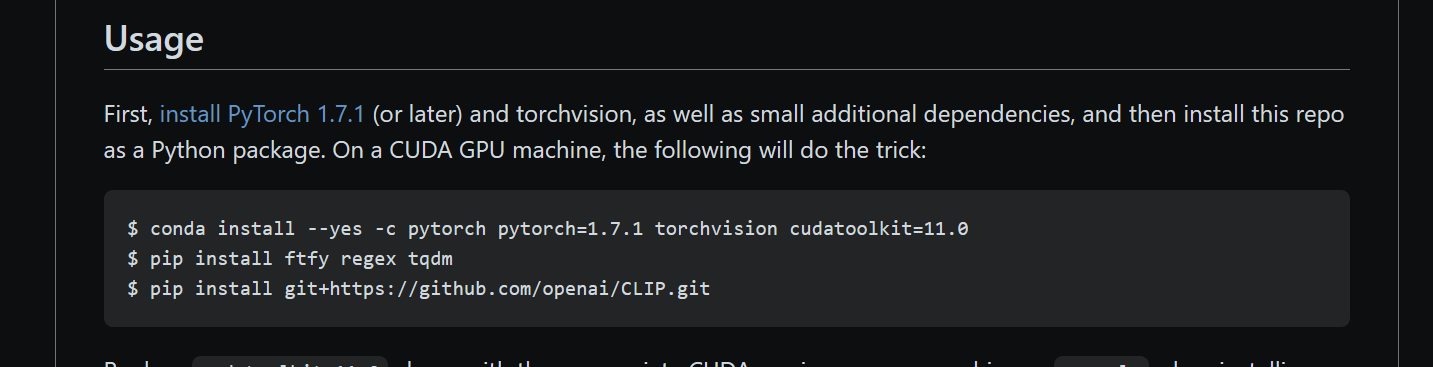
根据图像、视频、音频，分别提取其特征，根据某种媒体的特征来检索别的模态的内容。

因个人能力有限，及对深度学习等等领域和python语言的不熟悉，这里仅实现了使用文字搜索图像的功能。

1. **实现细节**

**首先完成clip及其依赖的安装**

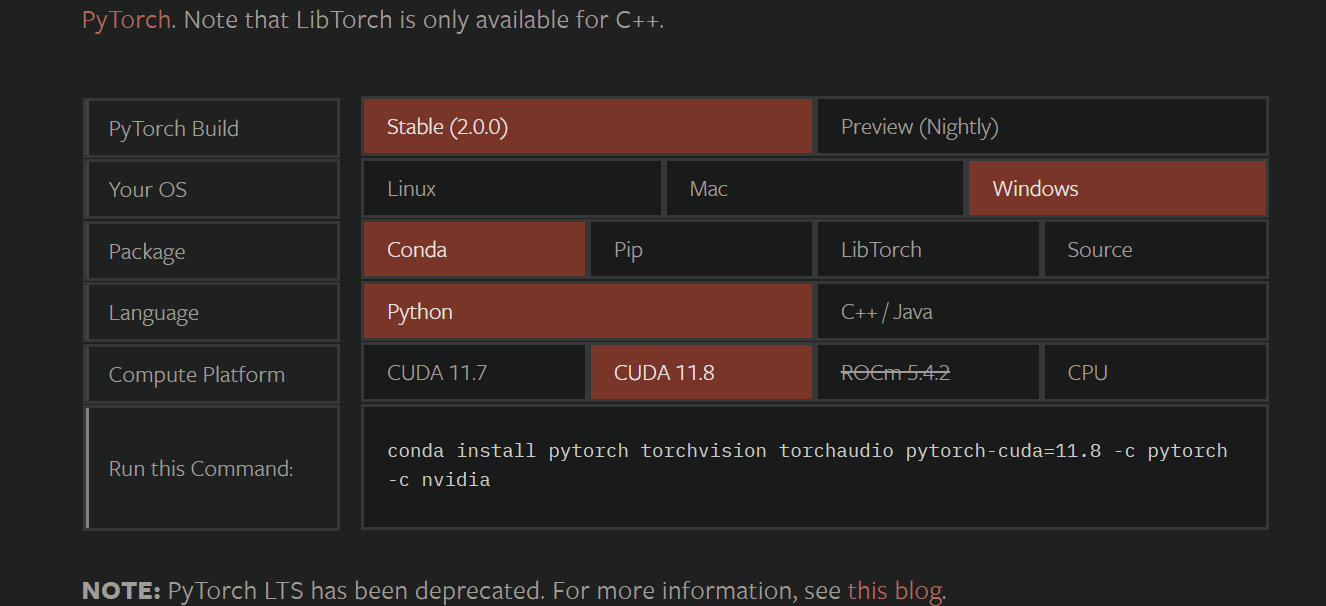
根据其github上的介绍，进行依赖的安装。



首先安装Anaconda，conda是一个用于包管理和环境管理的开源软件。它是Python的一个流行的包管理器，可以帮助用户安装、升级、删除和管理不同版本的软件包。由于conda支持多个操作系统，并且可以在不同的平台上共享环境设置和依赖关系，因此它非常适合Python程序开发和科学计算领域。除了Python之外，conda还支持其他编程语言和工具集的包管理。同时，conda还可以创建和管理虚拟环境，使得用户可以在一个系统中维护多个不同的工作环境，从而方便地进行不同项目或实验的开发。

使用conda安装pytorch。

**conda install pytorch torchvision torchaudio pytorch-cuda=11.8 -c pytorch -c nvidia**



使用conda安装ftfy regex tqdm

**conda install ftfy regex tqdm**

使用conda安装clip。使用git克隆到本地再安装的方式安装，可以避免一些依赖的冲突。

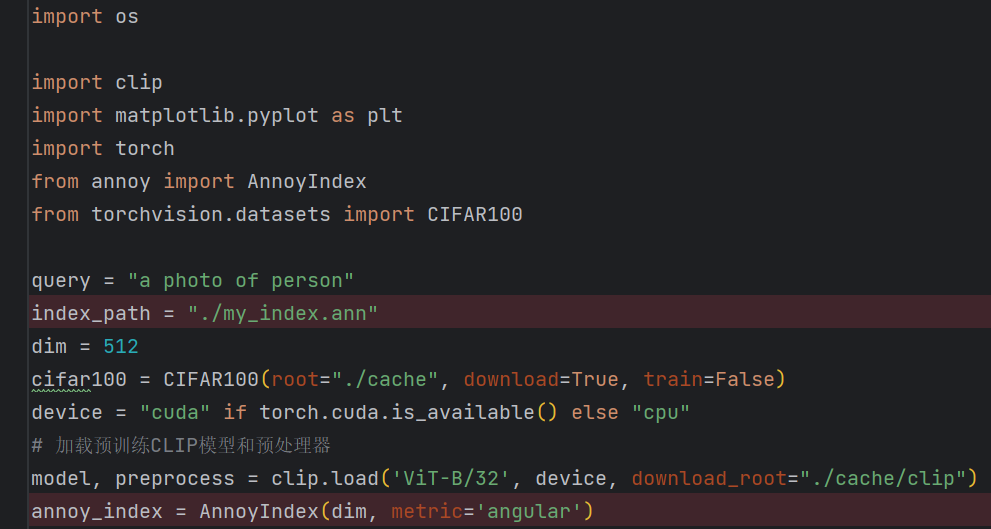
**conda install git**

**git clone https://github.com/openai/CLIP.git**

**cd CLIP/**

**python setup.py install**

**接下来介绍实现文字搜索图像的算法细节**



首先，定义一些变量，字符串query当作用户输入的文字，即用来搜索图像的原文字。定义index\_path，其具体作用在后边介绍。定义整数dim为512，这是因为CLIP模型会将图像特征提取为一个512维的向量，这个参数也将会在后面用到。

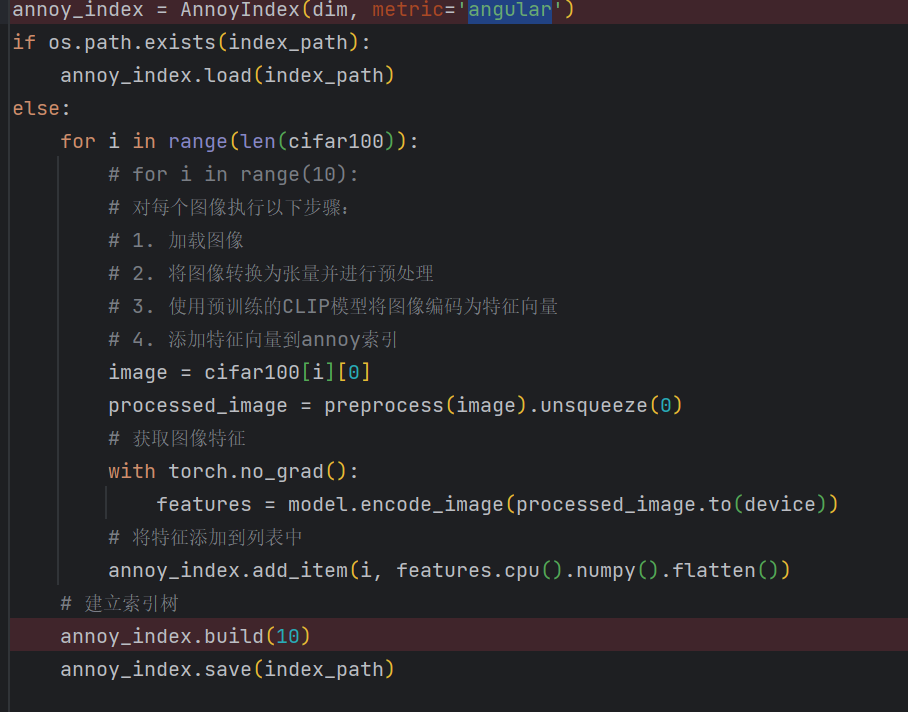
加载CIFAR100数据集，CIFAR-100是一个包含100个不同类别的图像数据集，每个类别都包含600张尺寸为32x32的彩色图像。CIFAR-100是CIFAR-10数据集的扩展版本，旨在测试更高级别的视觉分类能力。该数据集通常用于机器学习和计算机视觉方面的研究和开发工作，如分类、特征提取、图像识别和深度学习模型评估等任务。由于CIFAR-100是一个相对较小的数据集，因此可以在短时间内训练和评估各种算法，并进行快速实验验证。

加载预训练CLIP模型和预处理器。

创建一个annoy索引，其中元素为512维向量，即dim。该索引用于存放所有的图像的特征向量，使用这种方法可以极大的提高检索速度。参数"angular"是指夹角余弦相似性的度量方式。这是一种常见的衡量向量之间相似性的方法，在计算余弦相似性（Cosine Similarity）时经常用到。

夹角余弦相似性度量是通过两个向量的夹角余弦值（Cosine Value of the Angle Between Vectors）来表示它们在空间中的相似程度。具体地说，根据余弦定理，它可以被定义为两个向量点积与它们模长的乘积之比。因此，夹角余弦相似性是一个介于-1和1之间的标量值，其中1表示完全相似，0表示没有相似性，-1表示完全相反/不同方向。

因为CLIP模型在训练过程中使用图像与文字的特征向量的余弦相似度当作图像与文字是否相似的衡量标准，这里将参数 "angular" 传递给Annoy的构造函数。这会告诉Annoy使用夹角余弦相似性度量来计算向量之间的相似性，并相应地构建索引。



接下来为数据集中的每一张图像计算特征向量，并将其保存在索引中，另外这里会将构建完成的索引保存到文件中，可以避免每次搜索都计算所有的图像特征。



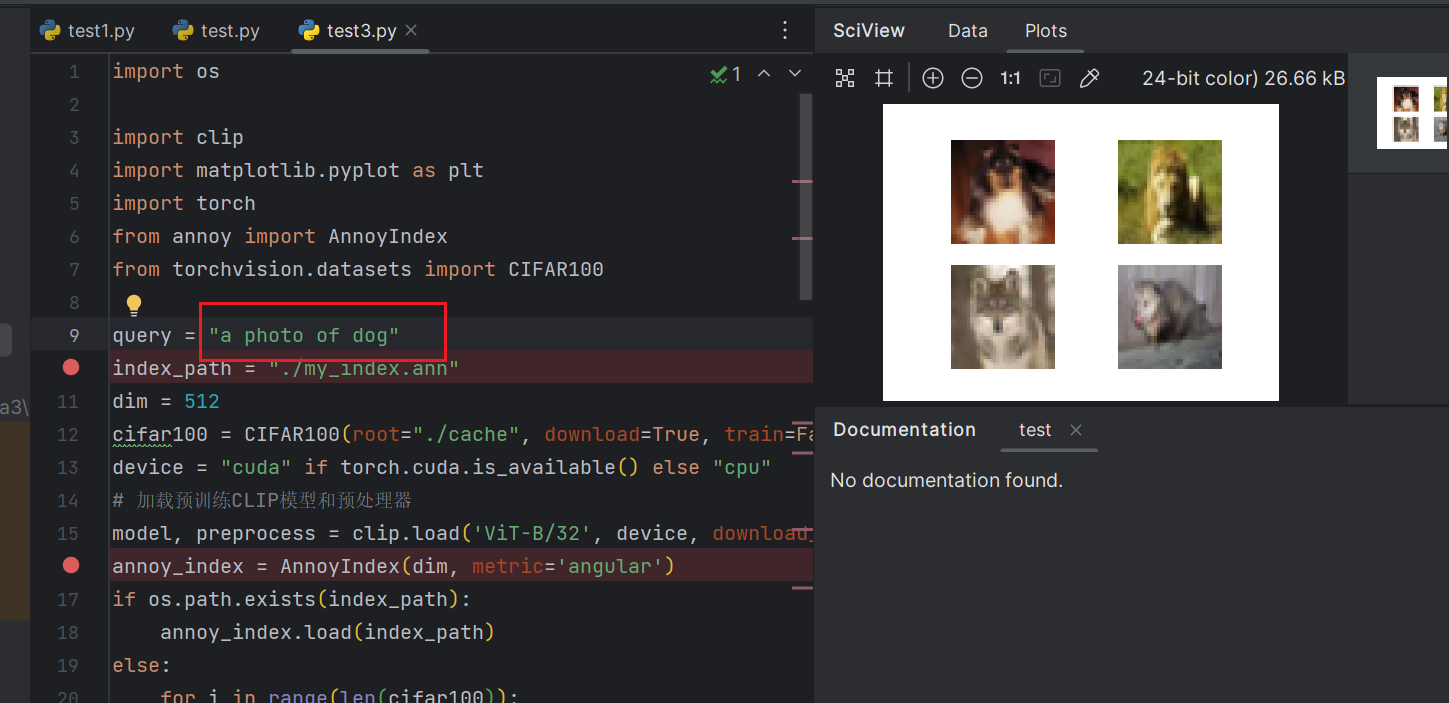
接下来只要提取用户输入字符串的特征向量并在annoy索引中匹配最接近的图像即可。因为CLIP模型在训练过程中使用图像与文字的特征向量的余弦相似度当作图像与文字是否相似的衡量标准，使用参数 "angular" 构造的Annoy索引中匹配到的图像应该会输入文字具有相近的含义。

1. **实验设置及结果分析（包括实验数据集）**

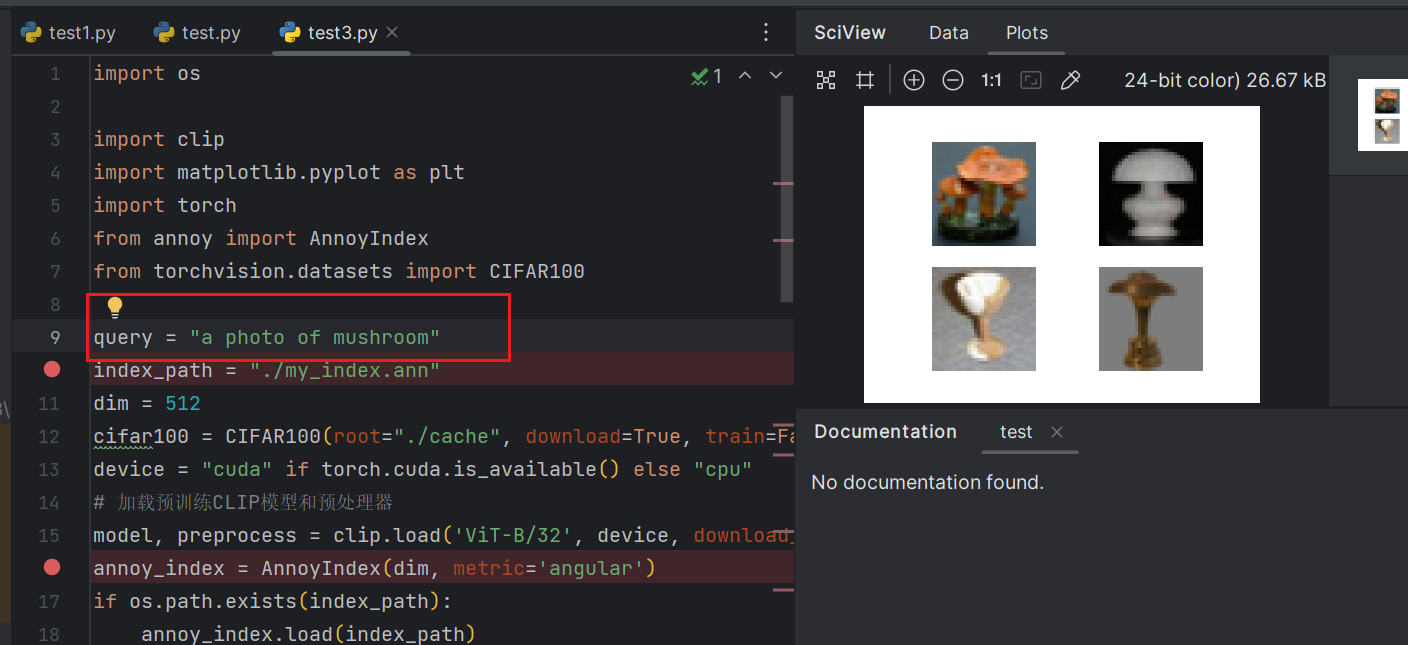
数据集，采用CIFAR-100，CIFAR-100是一个包含100个不同类别的图像数据集，每个类别都包含600张尺寸为32x32的彩色图像。CIFAR-100是CIFAR-10数据集的扩展版本，旨在测试更高级别的视觉分类能力。该数据集通常用于机器学习和计算机视觉方面的研究和开发工作，如分类、特征提取、图像识别和深度学习模型评估等任务。由于CIFAR-100是一个相对较小的数据集，因此可以在短时间内训练和评估各种算法，并进行快速实验验证。

**实验**

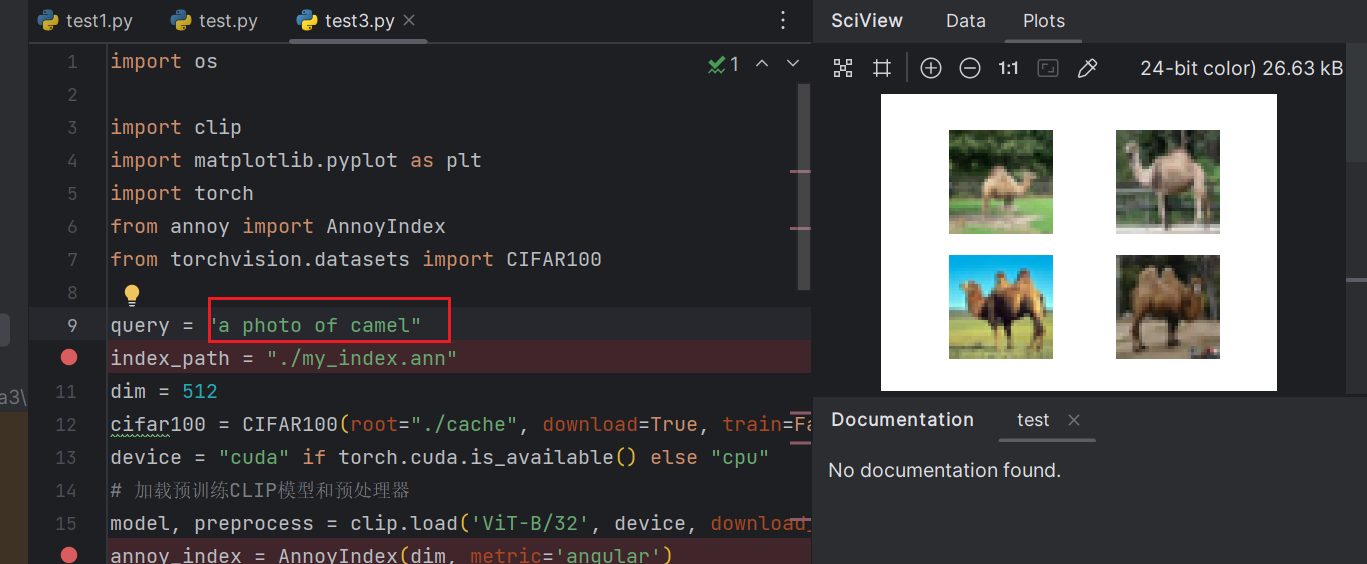
**搜索“a photo of dog”**



**搜索“a photo of mushroom”**



**搜索“a photo of camel”**

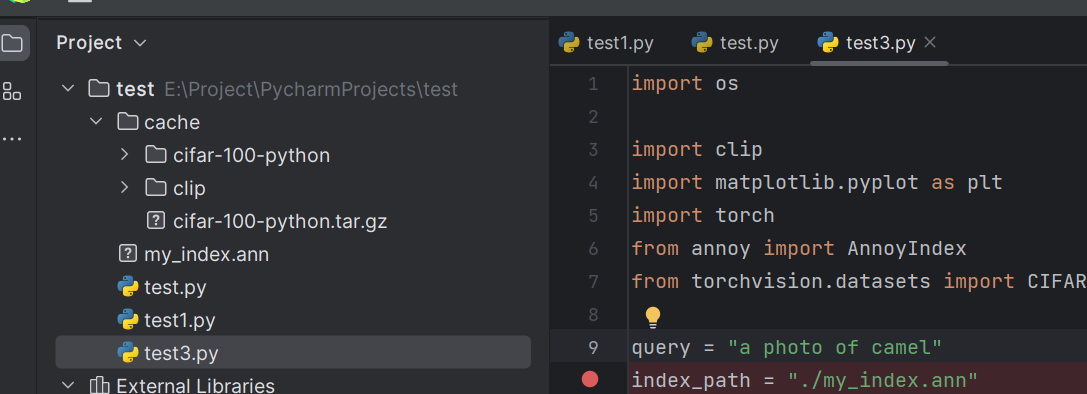


可以看到搜索结果都具有很高的相似性。

1. **结论**

**使用clip模型实现了一个简单的文字搜索图片的跨模态检索应用。**

**最终运行后目录结构如图：**



其中，cache中为clip的与训练模型，cifar-100-python为CIFAR-100数据集。my\_index.ann为将数据集中图像提取特征向量后保存的annoy索引文件。

test.py为CLIP实现Zero-Shot分类的程序，并不属于这次实验的主要程序，用于验证环境的搭建是否完全。test1.py是clip官网的Zero-Shot Prediction，同样用于验证环境。

test3.py中为这次实验的主要程序，即上述展示的程序。

1. **参考文献**
2. clip模型官方github仓库（[openai/CLIP: CLIP (Contrastive Language-Image Pretraining), Predict the most relevant text snippet given an image (github.com)](https://github.com/openai/CLIP)）
3. 神器CLIP：连接文本和图像，打造可迁移的视觉模型（[神器CLIP：连接文本和图像，打造可迁移的视觉模型 - 知乎 (zhihu.com)](https://zhuanlan.zhihu.com/p/493489688)）