实验 3：影响函数与可解释安全

一、实验目的：

掌握数据分类模型的训练方法，理解影响函数的作用，能使用影响函数工具分析数据集对模型的影响。

二、实验原理：

影响函数是用来衡量一个训练样本对一个测试样本的影响力的。近年来，机器学习模型的预测准确度越来越高。但是要去解释系统为什么会做出这样的预测并不容易，使用影响函数对模型进行解释是一种突破性的尝试。模型学习的所有信息和“知识”都是从训练集中得出的，所以可以查询特定预测受到各个训练样本的影响有多大。如果去掉或微调某些训练数据，导致预测的置信水平变化，那么这个差值就可以反映该训练数据对整个模型的影响。找到对模型性能影响最大的训练样本对于优化模型具有重要意义，这有助于理解人工智能可解释安全性的多样性。本实验可以帮助同学理解影响函数的计算流程和作用，引导同学进行更深层次的算法可解释性研究。

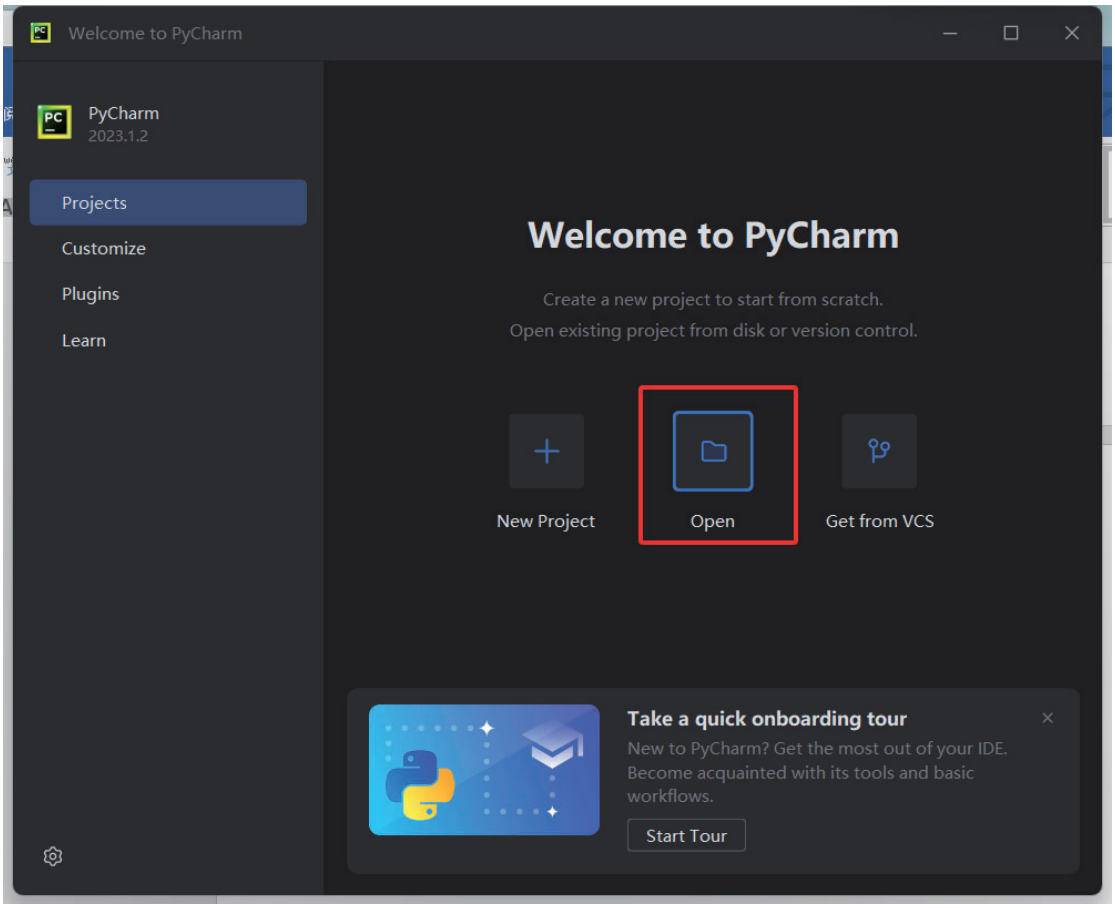
三、实验内容：

使用 anaconda 创建实验环境，并安装软件包；训练一个 CIFAR10 数据分类模型，用于后续实验；范围查询训练集中的图片；查询测试集中的特定图片；计算测试集中某个测试用例对应所有训练样本的影响函数；查询对某个测试用例正面影响最大的训练样本和负面影响最大的训练样本。

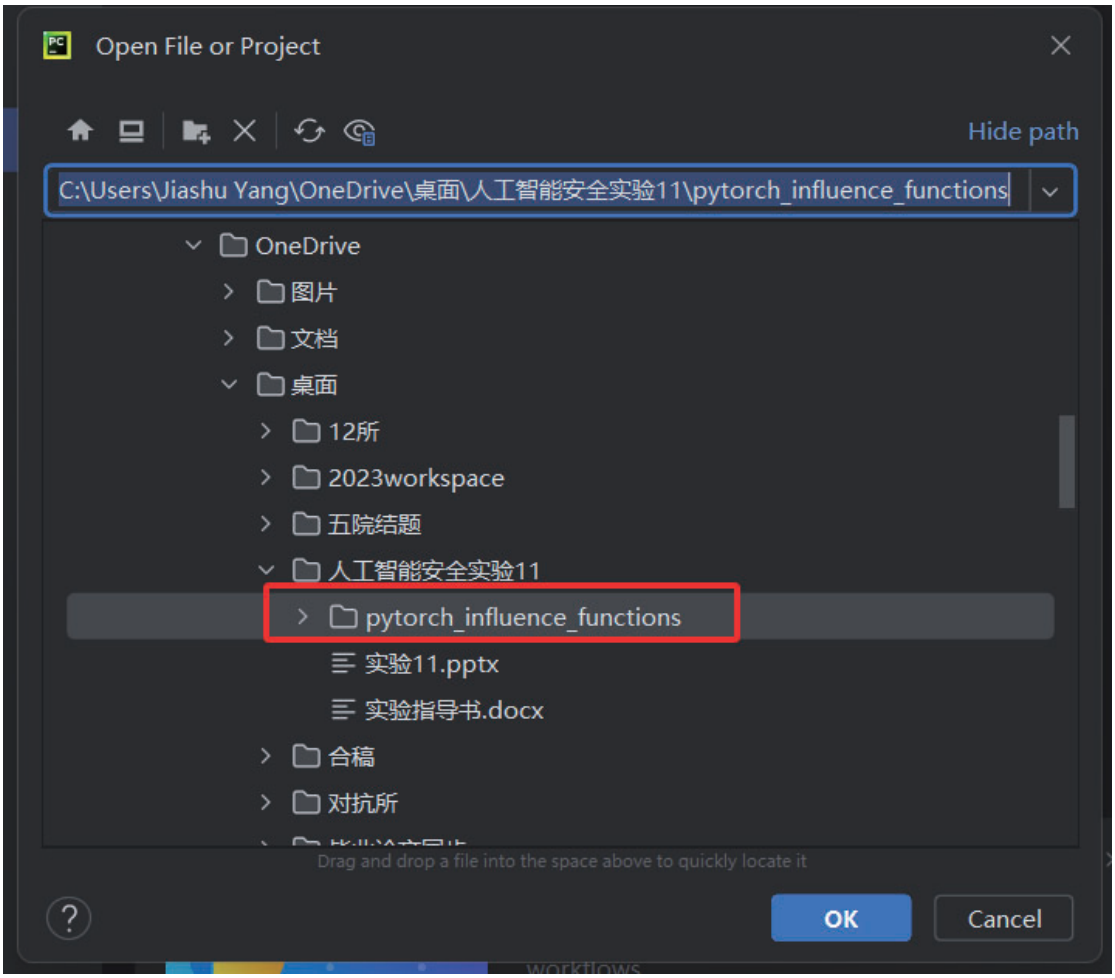
四、实验步骤：

1. 项目准备

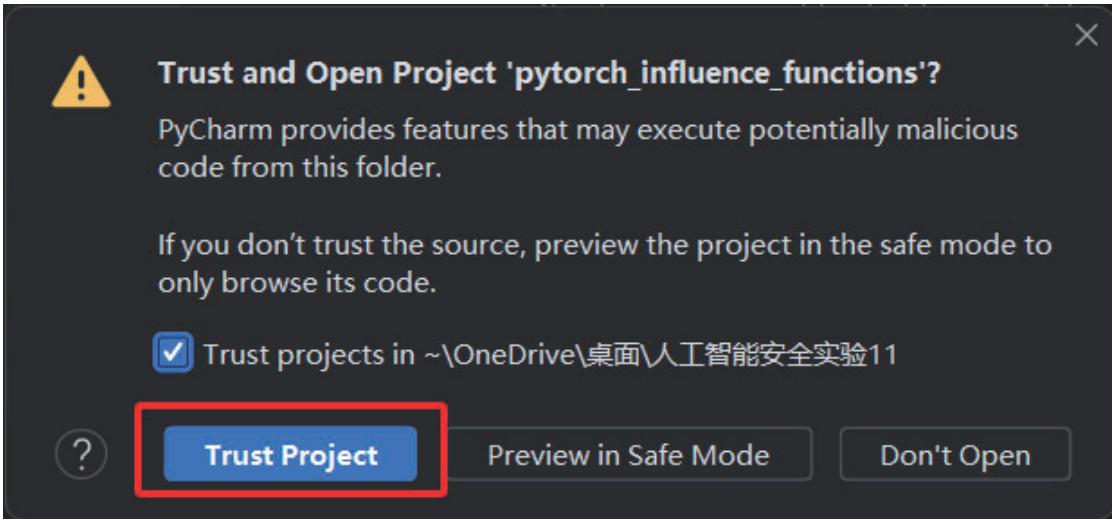
打开 pycharm，点击 open 打开项目：



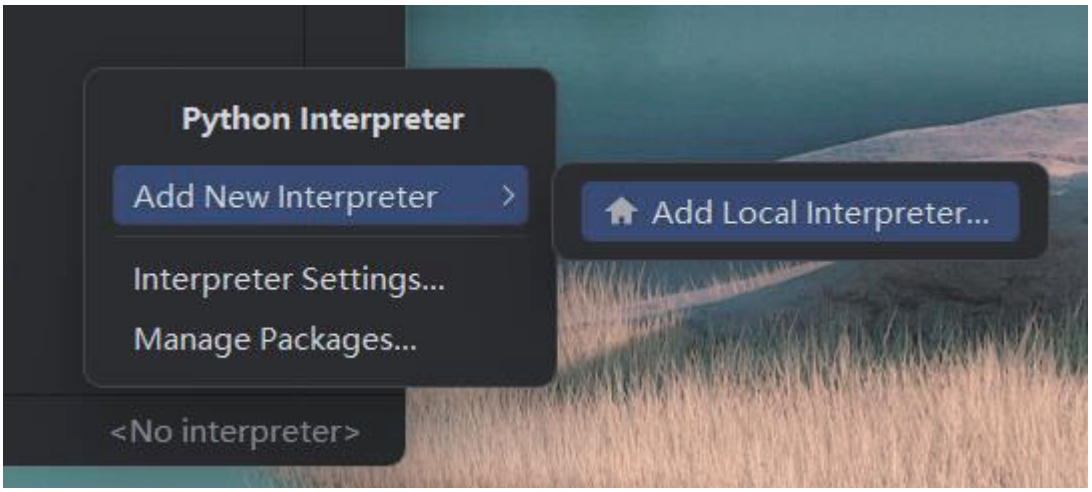
选择实验代码文件 pytorch_influence_functions：



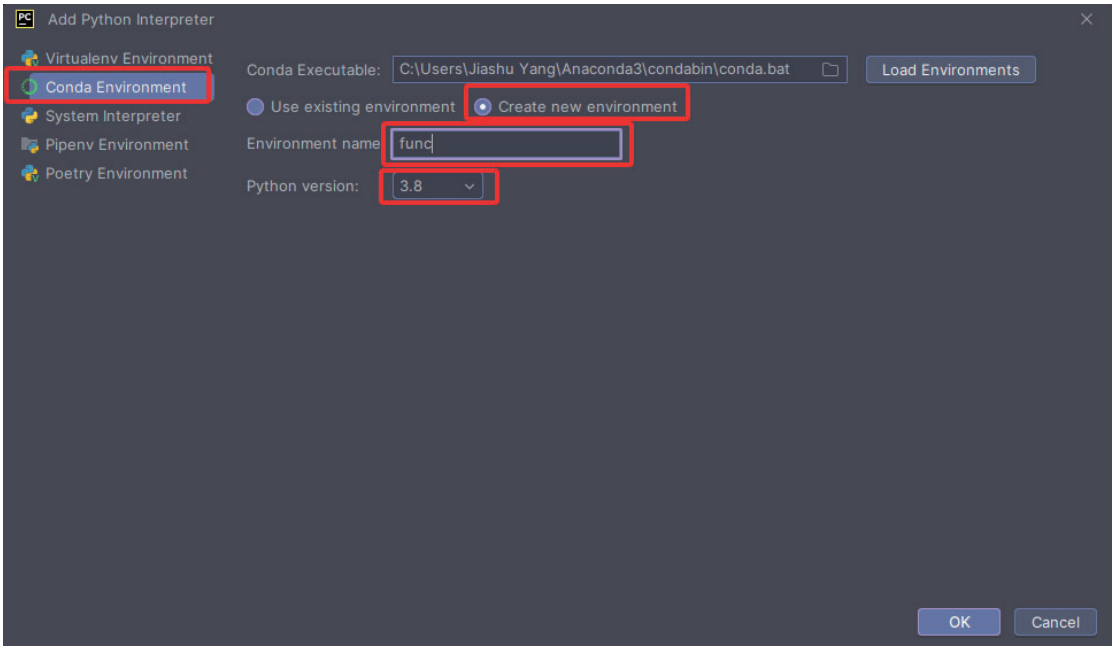
如果出现警告弹框，先勾选，再点击 Trust Project：



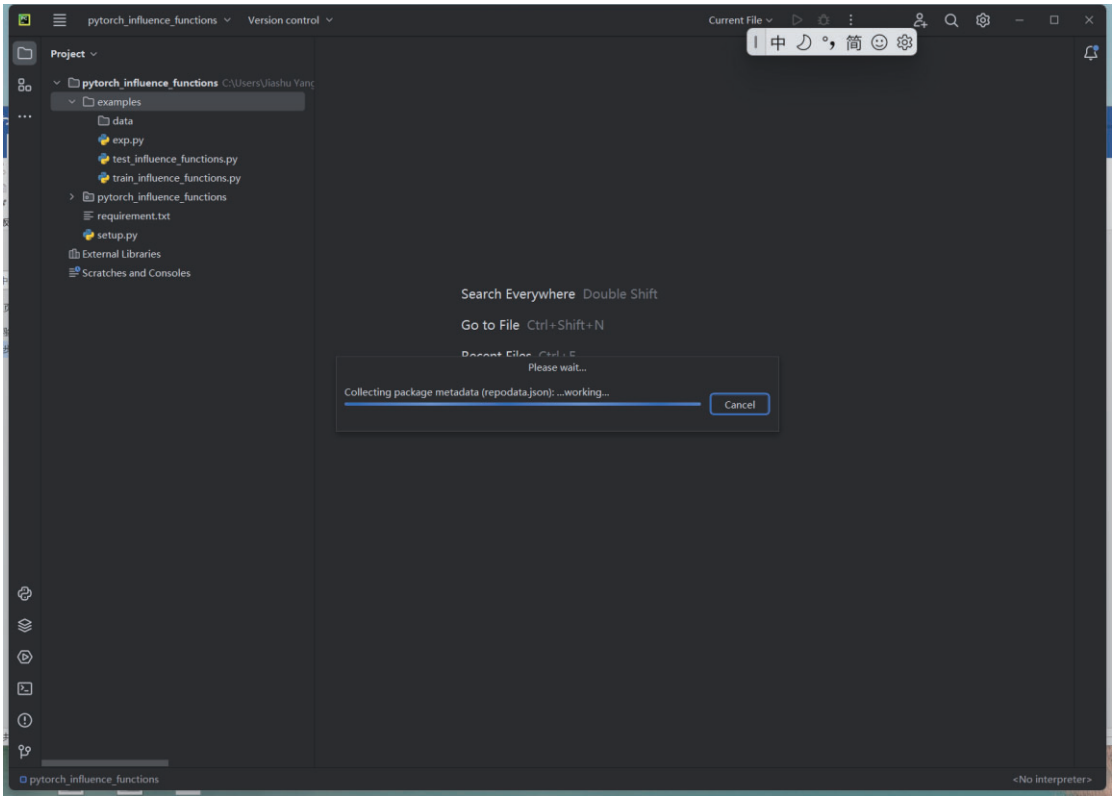
在 pycharm 右下角依次点击<No interpreter>——Add New Interpreter ——Add Local Interpreter 设置 python 环境：



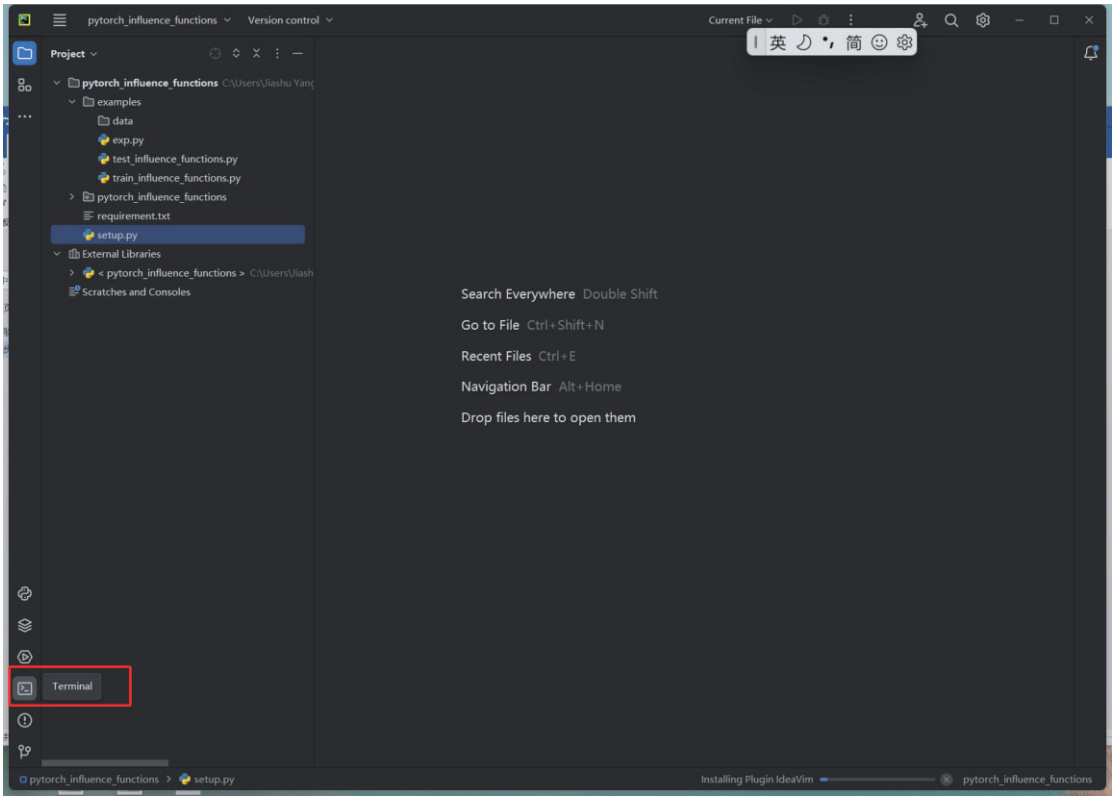
点击 Conda Environment 创建新的 conda 环境，选项如图所示：



创建后等待几分钟，初始化环境：



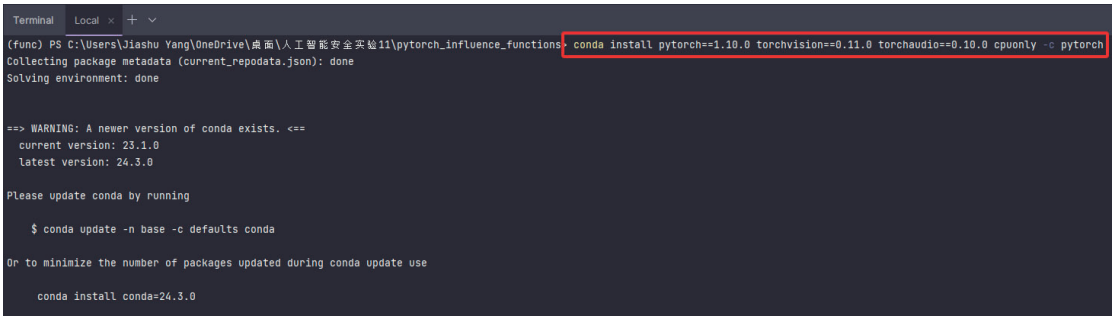
点击左下角的 Terminal，启动命令行程序：



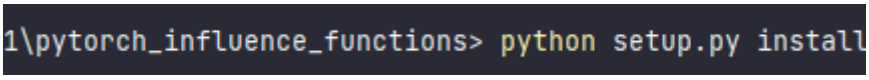
在命令行中输入

conda install matplotlib pytorch==1.10.0 torchvision==0.11.0 torchaudio==0.10.0 cpuonly -c pytorch

等待依赖包安装完成：



在命令行中输入 python setup.py install 安装影响函数工具：



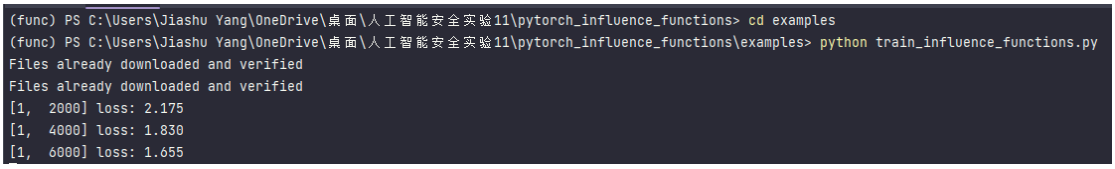
至此，实验环境配置完成。

2. 训练 CIFAR10 数据分类模型

在命令行中依次输入以下命令：

cd examples

python train_influence_functions.py



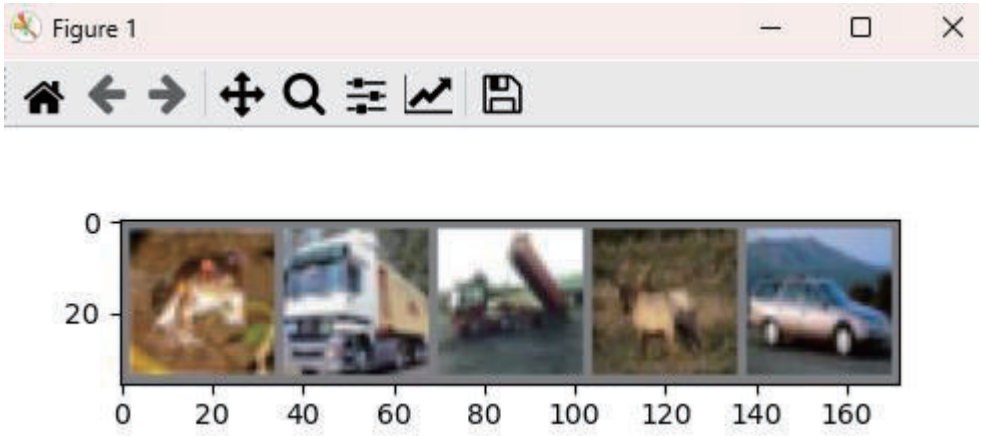
训练结束后会出现类似信息：

```
Finished Training
Accuracy of the network on the 10000 test images: 62 %
Accuracy of plane : 67 %
Accuracy of   car : 67 %
Accuracy of  bird : 43 %
Accuracy of   cat : 47 %
Accuracy of  deer : 59 %
Accuracy of   dog : 50 %
Accuracy of  frog : 76 %
Accuracy of horse : 69 %
Accuracy of  ship : 66 %
Accuracy of truck : 79 %
```

请截图保存用于验收和实验报告。

3. 范围查询训练集中的图片

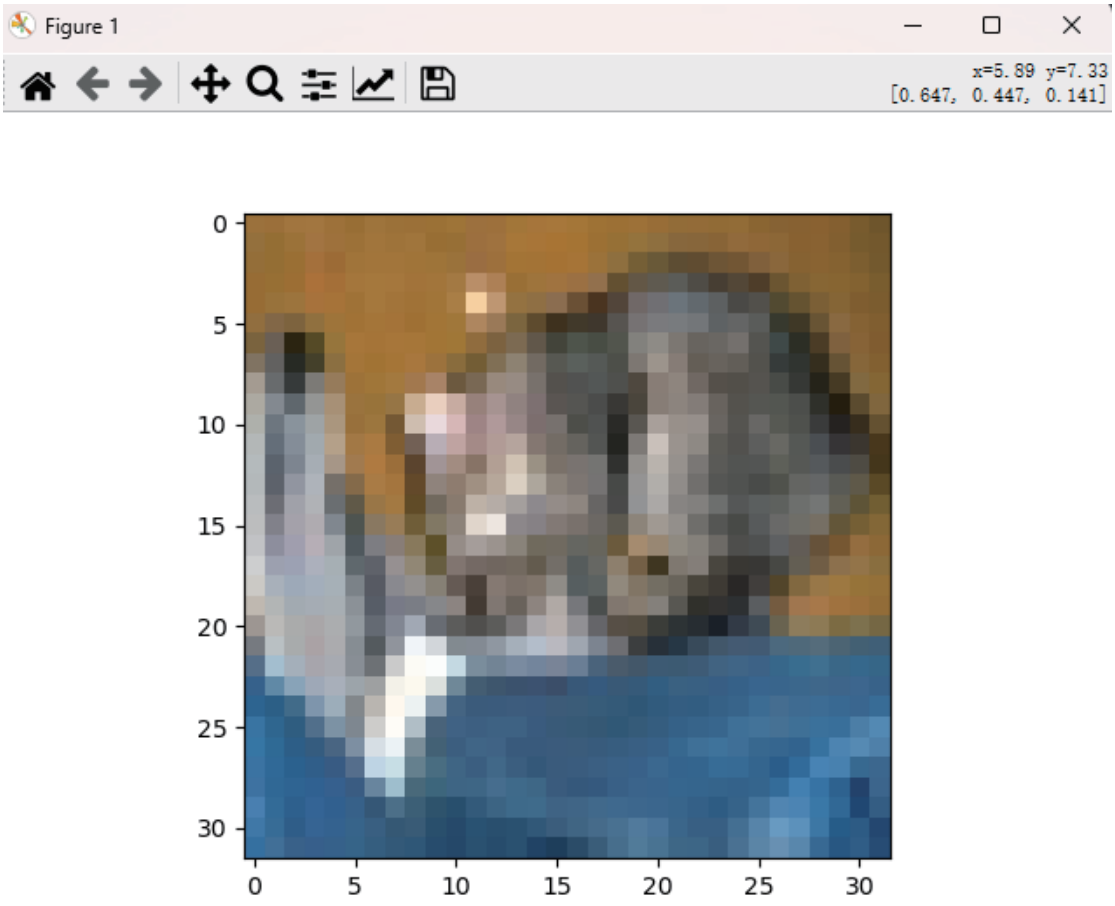
在 examples 文件夹下运行 python exp.py --exam 1，运行结束后会展示训练集中的前 5 张图片，如图所示：



任务：打开代码 exp.py，修改第 19 行代码为 PIC_FOR_SHOW = 16，再次运行命令 python exp.py --exam 1，观察结果并截图保存。

4. 查询测试集中的特定图片。

在 examples 文件夹下运行 python exp.py --exam 2，运行结束后会展示测试集中的第 1 张图片，如图所示：



任务：打开代码 exp.py，修改第 33 行代码为 i = 5，

再次运行命令 python exp.py --exam 2，观察结果并截图保存。

5. 计算训练样本的影响函数

在 examples 文件夹下运行 python exp.py --exam 3，运行结束后会得到测试集中的第 1 张图片对应训练样本的影响函数，如图所示：

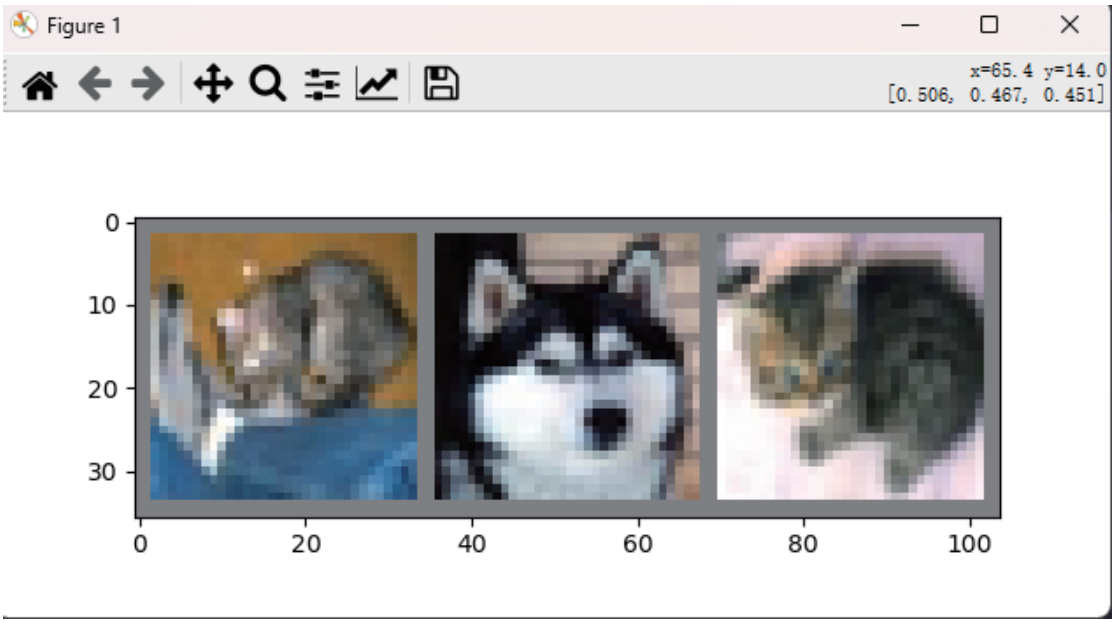
```
(func) PS C:\Users\Jiashu Yang\OneDrive\桌面\人工智能安全实验11\pytorch_influence_functions\examples> python exp.py --exam 3
Files already downloaded and verified
Files already downloaded and verified
C:\Users\Jiashu Yang\anaconda3\envs\func\lib\site-packages\pytorch_influence_functions-0.1.1-py3.8.egg\pytorch_influence_funct
ted. Change the call to include dim=X as an argument.
Calc. s_test recursions: [==================================================] 1 / 1
Averaging r-times: [==================================================] 1 / 1
Calc. influence function: [==========================================] 49999 / 50000
Calc. influence function: [==========================================] 50000 / 50000
负面影响最大的训练图片id: 30159, 影响函数值为-1.58975e-03
正面影响最大的训练图片id: 37699, 影响函数值为2.17170e-03
```

任务：打开代码 exp.py，修改第 40 行代码为 i = 50，

再次运行命令 python exp.py --exam 3，观察结果并截图保存。

6. 查询正面影响最大的训练样本和负面影响最大的训练样本

在 examples 文件夹下运行 python exp.py --exam 4，运行结束后会得到测试集中的第 1 张图片对应训练样本的影响函数，如图所示：



任务：打开代码 exp.py，修改第 54 行代码为 i = 500，

再次运行命令 python exp.py --exam 4，观察结果并截图保存。

五、验收要求：

1. 完成模型训练并保存 CIFAR10 分类模型的训练结果。
2. 完成范围查询训练集中的图片的任务，并保存结果。
3. 完成查询测试集中特定图片的任务，并保存结果。
4. 计算测试集中某个测试用例对应所有训练样本的影响函数，并保存结果。
5. 查询对某个测试用例正面影响最大的训练样本和负面影响最大的训练样本，并保存结果。