2024ERQI-MachineLearning

> **关于机器学习**
>
> Ciallo~
> 这或许是你接触人工智能——Artificial Intelligence的最佳机会。
>
> 微积分？线性代数？概率论？没学怎么办？很简单，入门机器学习根本不需要高深的数学功底，看到不会的数学公式，去查去学习就可以了，重要的是拥有对学习的热情。
>
> 很快你就会发现Machine Learning的魅力，用精妙的方法处理并学习数据，并逐步发掘“Computer Vision”和“Natural Language Processing”等细分领域的真相。

> **⚠ 注意**
>
> ***诚然，你可以使用GPT辅助完成以下任何题目，但我们会在面试时着重询问你自己对提交的代码的理解，请确保自己理解代码和基本概念。***

机器学习（AI）

0. 如何提交你的成果

- 提交截止时间：**2024.10.17 23:59**（可能延后）
- 以最新的一次提交作为最终提交
- 文件夹`[姓名]-[学号]`如：`科比布莱恩特-1145141919810`
  - `前置知识`
    - 相关学习笔记
    - 软件安装成功的截图
  - `二分类`
    - 详情请见题目的**提交要求**
  - ……
- 文件夹最终打包为zip压缩包，命名不变，发送至`weldaspica@gmail.com`

> **⚠ 注意**
>
> 每道题的提交内容并非一致，请认真阅读题目的提交要求。
> 同时，只有把压缩包发送到邮箱并收到回复，才算正式提交！

1. 前置知识

> *故事纯属虚构，如有雷同纯属巧合。*

在一个阳光明媚的夏末，成小电终于来到了心心念念的University of Everyday Study and Test of China，俗称UESTC。

成小电是一名普通的UESTC信息与软件工程学院2024级学生，自从入学，他就希望自己加入信软学院最有活力、最年轻的工作室——**尔绮工作室。**
小电在AI的热潮中进入了大学，面对着一眼望过去全是男生的学校，小电忽然觉得这就像自己一眼望到头的人生。

听着路过的学长衷心祝愿着某课程老师全家活光光，班级交流群的头像是个抽象的二次元动漫小人，学校操场上某外部企业大声外放激情的运动会战歌，微博上UESTC官方账号发布了和原神的联动消息，小电觉得自己这辈子有了。

他本想就这样眼前一黑晕倒过去，但他一瞄眼前的路面，发现年久失修的人行道坑坑洼洼，他觉得不看路的话，可能一闭眼一辈子就过去了。破碎的路面，宛如他破碎的心。

——忽然，成小电灵光一闪，AI可以从绝境中缝补他的遗憾！
谁说对着二进制和硅基的智能就不能产生爱呢？从此，小电踏上了一条不归路。

1.1 学习软件的安装

> 小电明白，要打造一个得力的助手，必须要有创造助手的工具。

工欲善其事，必先利其器。
然而，请注意，我们所用的一切工具，最终是为了实现代码，请勿本末倒置！

1.1.1 Anaconda

Anaconda是AI应用和研究中常见且好用的工具。

> *小电，你的各种第三方包怎么有各种冲突啊？你虚拟环境呢，救一下啊。*

- 详细了解什么是Anaconda/miniconda。
- 为什么我们要使用Anaconda?
- 怎么使用Anaconda辅助机器学习的学习?
- 下载Anaconda，若想下载miniconda，请自行寻找资源。

1.1.2 开发平台

众所周知，现代AI常用Python实现，同时，Python程序本身也需要一个合适的平台来编写并运行，这里是学长推荐的平台，以下内容不是你的唯一选择哦。

> *小电，你也不想你的代码跑不起来吧。*

- 下载 PyCharm：JetBrains 出品的用于数据科学和 Web 开发的 Python IDE
- Jupyter Notebook（常与Anaconda一同被下载好）
- VS Code（请自行搜索配置Python环境的方法）

1.2 常用网站推荐

在问问题之前，先看有没有其他人问过了你的问题；学过知识以后，不妨也看看其他人对知识的理解；不了解一份工具的使用，记得先找一找有没有大佬写过精妙的学习手册……
如今初学者95%的问题，都有无数前人踩过坑，给过解决方案。

> *没有学习过提问技术的小电，被前辈送了两个词：RTFM（Read The F??king Manual）和STFW（Search The F??king Web），虽然小电不知道什么是F??king，但他总觉得这不是什么好词，这是怎么绘世呢?*

下面是一些你可能会用到的网站/手册：

- CSDN - 专业开发者社区：中文IT技术交流平台，如今出于各种原因不推荐使用，但你偶尔能找到一些还算有用的文章。
- GitHub：GitHub是一个面向开源及私有软件项目的托管平台，其基于Git这一分布式版本控制系统，提供了代码托管、版本控制、协作开发、项目管理等一系列功能。
- 主页 - PyTorch中文文档 (pytorch-cn.readthedocs.io)
- sklearn (scikitlearn.com.cn)
- OpenCV中文官方文档 (woshicver.com)

1.3 学习视频

通过课程学习是学生的传统美德，或许你可以通过课程得到比较系统的知识框架。
有些人会说**横向无用论**，这常常出现于年轻的技术极客：*所学一切只为实现目标服务，和目标无关的一概不学习，拓宽无用技术的视野是没有意义的。*
也有一些人会说**纵向无用论**，这常常出现于大学老师的说教中：*知识厚度的积累至关重要，有了完整的体系才有未来的长足发展。*

> *小电被学长吐槽不会提问后，终于学会了提问，于是得到了上述的两个回答，他疑惑了——该相信谁啊？都有道理啊?*

笔者认为，走极端不可取，二者都有可取之处，亦有不可取处。
受限于知识面的狭窄，某些客观上可能存在关联的知识会被错判为毫无关联，从而导致纵向的剑走偏锋；困扰于知识面的边边角角，某些真正有意义的事情被忽略去做，从而导致横向的茫然无措。
过于忠于目标，过于重视基础，私以为都不可取。每个人都有适合自己的学习方法，笔者认为有系统的学习还是对未来有帮助的。

若觉得看课程更适合自己，大可以去看；若觉得课程过于冗长，亦可以直接做题。
选择在你们，未来是你们的。

- 吴恩达机器学习系列课程_哔哩哔哩_bilibili
- 吴恩达深度学习deeplearning.ai_哔哩哔哩_bilibili
- 跟李沐学AI的个人空间-跟李沐学AI个人主页-哔哩哔哩视频 (bilibili.com)

通过前两个视频合集可以学到一些理论层面的东西，关于一些东西不懂可以借助AI、学习笔记（csdn，github上面都有很多）、找学长等方法解惑，通过李沐的视频可以学习实践内容。理论和实践结合的方式对学习ML有奇效！

1.4 Python

> *不是，哥们——小电发出疑问，理论会了，怎么实现啊？*

时间充裕的同学可以自行寻找合适课程学习python语言，也可以通过后面的题目学习要用的内容（机器学习并不要求掌握python所有知识，但你学习了python后有助于从语言方向理解代码的含义）。

1.5 基础问题

> *学习也是一个不断提问的过程，于是小电不禁询问AI：你会玩原神吗？*

以下仅为部分基础知识，在面试时我们可能会询问你已学过的，但并不在这些问题中的知识。

1、什么是机器学习?

请用自己的语言描述你对机器学习的认识。

2、谈谈对聚类的理解。

随便谈，建议从原理、应用等方向思考

3、谈谈cost function与gradient descent。

可基于不同的模型简述对其的认识

4、在训练深度学习模型时，正向传播与反向传播分别起什么作用?

5、什么是过拟合，欠拟合？遇到后你会考虑用什么方法去解决问题?

6、分类问题中激活函数是什么，有什么作用，你能举出几个激活函数吗？

7、正则化是怎么样的一个操作？你能简述其过程吗？

8、为什么要设置学习率衰减？方法有哪些？

9、为什么会出现局部最优，有什么方法避免吗？

10、池化层和全连接层的作用？

2. 二分类

> **开始做题啦**
>
> 这道题属于送分题，重点在于学会如何去简单使用机器学习和理解基本概念。
> 建议首先学会科学上网，此题需要使用国外网站**kaggle**参加竞赛。别被所谓参加竞赛吓到了，这道题很简单的，kaggle上也有免费的新手入门课程教你们怎么打这个竞赛。
> **往下做题目之前，先注册kaggle。**

2.1 题目背景

有一天，邓焯同学在边练习剑道边看《泰坦尼克号》，他被感动得泪流满面的同时又在想，船上哪些人在一场沉船的灾难里更可能活下来呢?

于是他想写一段代码，这段代码可以把船上乘客的各种特点转换成1和0的输出。
这些特点包括座舱的级别、性别、年龄、家庭关系等等，看一看如何去处理它们哦！你的代码需要接收这些特点，然后得到输出。
1代表乘客可以活下来，0代表乘客最后未能存活。

题目地址：Titanic - Machine Learning from Disaster | Kaggle

2.2 数据描述

> *训练集，验证集，测试集的概念要明白哦。*

详见Titanic-Data
你将使用train.csv训练你的代码，并读取test.csv的内容，得到对应的1和0的输出。

2.3 如何完成

- 确保你已经通过某些学习渠道了解过机器学习中二分类的基本原理。

- 认真阅读这项Kaggle入门竞赛的Overview部分，它已经非常全面地向你介绍了如何进行一场机器学习竞赛。
- 写一段针对这个竞赛的**二分类机器学习代码**，可以调用你学习的任意框架（如tensorflow、pytorch或mxnet等），使得你的代码更加简洁，也可以从零开始实现一个二分类问题，建议使用框架哦。
- 在Kaggle上提交你的预测结果，**分数至少要在0.75以上。**
- 如果你实在不知道如何下手，请看Titanic Tutorial (kaggle.com)，跟着这个Tutorial就能完全完成这道题，但如果代码是你自己写的而非照搬手册，会有加分哦

2.4 提交要求

- 你在kaggle的leaderboard上的分数截图
- 你的源代码，格式可以为 `.py` 或 `.ipynb`
- 关于这道题你遇见的问题，和你的解决方案
- 关于这道题你的学习笔记

2. (附加题) Linear Regression

> *难度和第二题差不多哦，如果写第三题卡住了，可以再试试这一道题。*
> *这是一个典型的回归问题，和分类问题有一定程度上的差别*

要求

附加题不是必做题，但建议尝试哦！

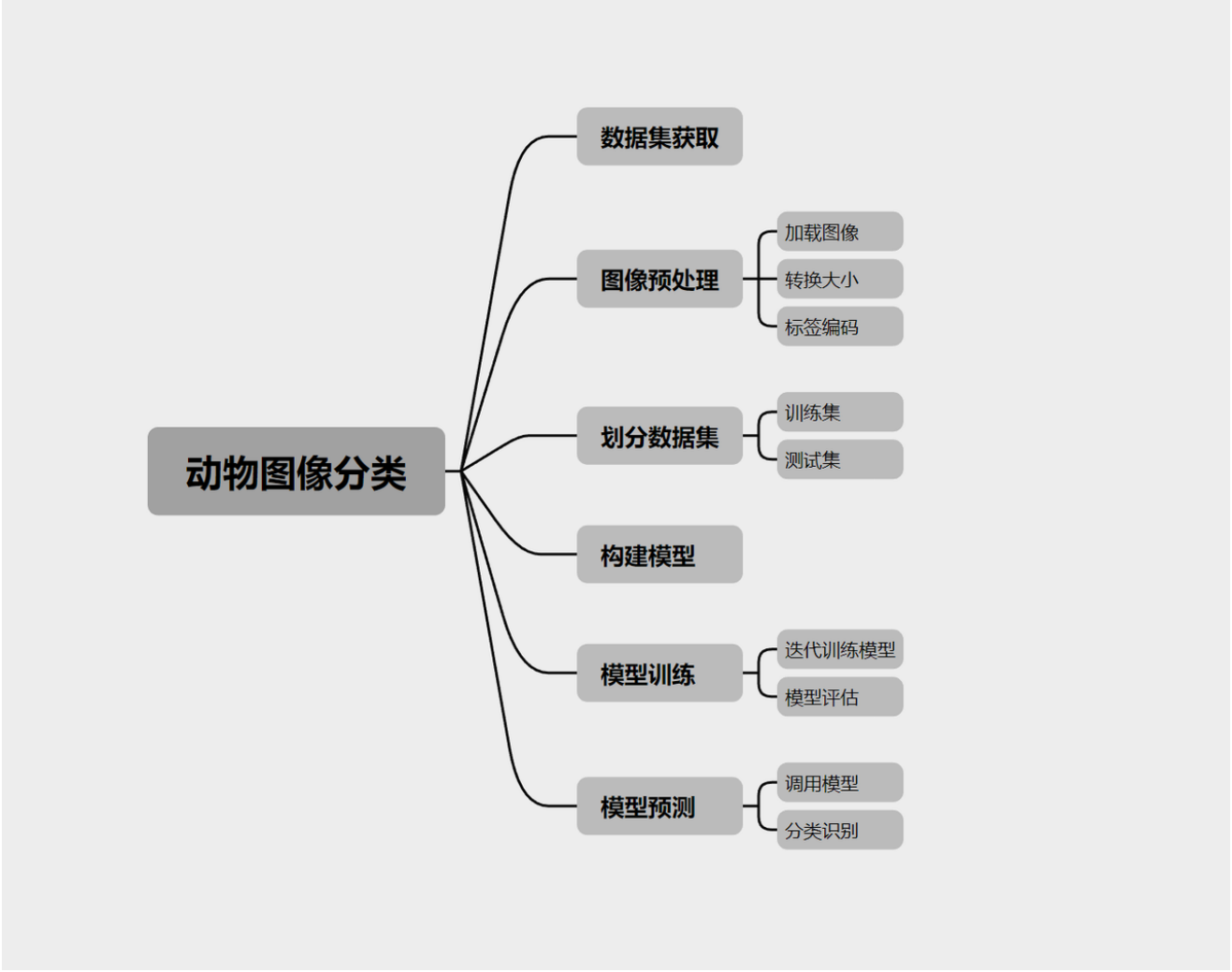
题目地址

House Prices - Advanced Regression Techniques | Kaggle

提交要求

- Leaderboard上的score需要低于0.15，达不到这个要求也没关系，尽量往这个方向靠拢。

3. Classification



看到上面的思维导图大家应该明白这道题考察的是动物图像分类的知识（上图思路仅作参考）。

3.1 背景介绍

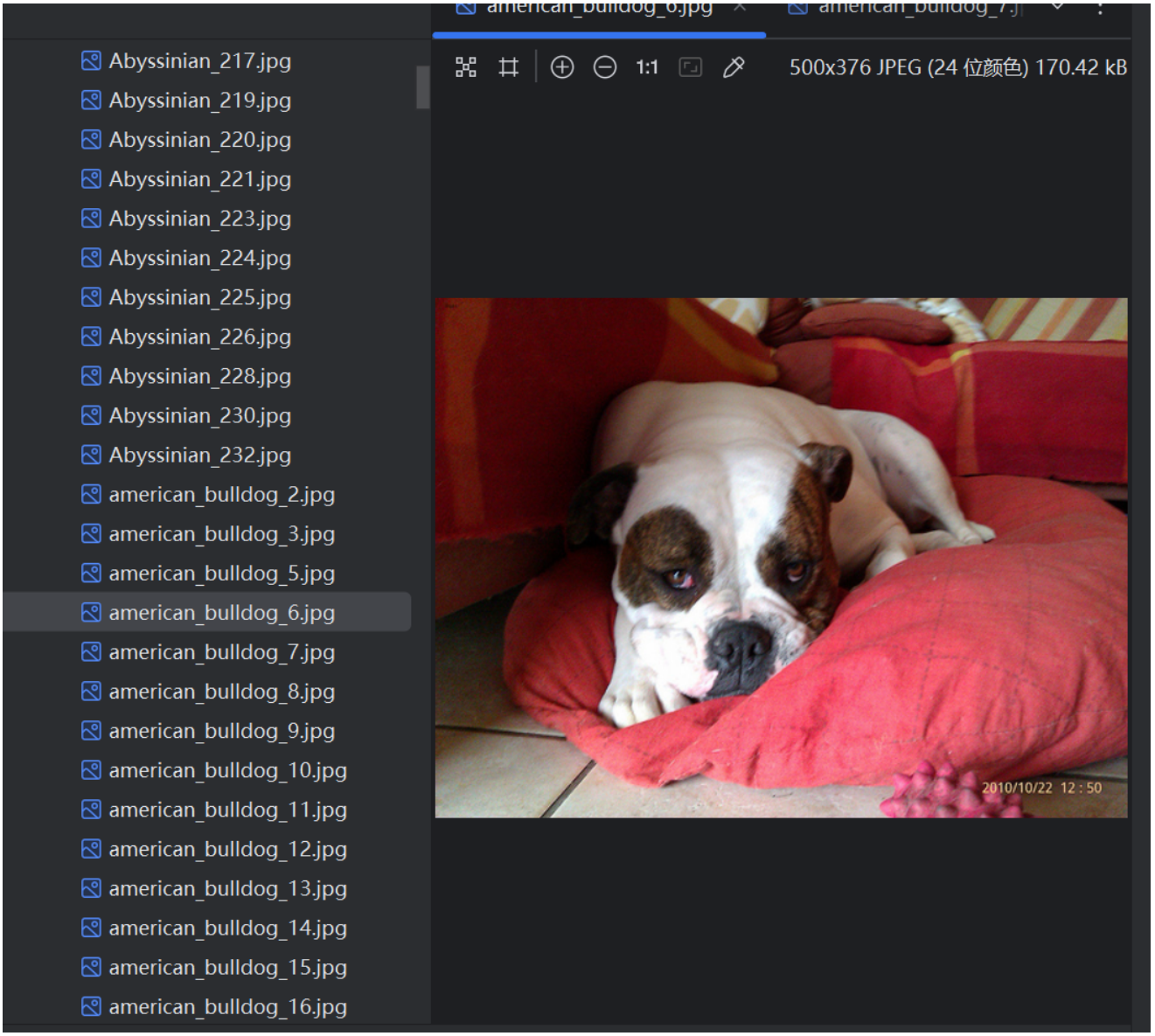
你有一个叫做王哲的好朋友，他是一位爱心人士，喜欢和他的动物朋友们一起玩耍。一天他突然发现了一些新朋友却不能弄清品种，让他非常苦恼。你为了帮助他解决苦恼，便决定用学到的知识，实现动物图像识别分类任务。

题目所用文件

请在招新群内的群文件获取。

3.2 目前情况

王哲认识了许多动物朋友，用相片记录了他们的外貌，并且为他们做好了标签，如下：



3.3 题目目标

自行写出训练文件（train.py可以在网上寻找框架并套用）,对图像进行分类训练，再在predict.py文件中调用自己的.pth文件（predict.py第十三行）实现图像的预测。

3.4 本题需提交准确率截图预测截图与pth文件

```
Epoch [2/10]: Loss: 0.6971
Train Accuracy: 87.08%
Test Accuracy: 86.60%
```

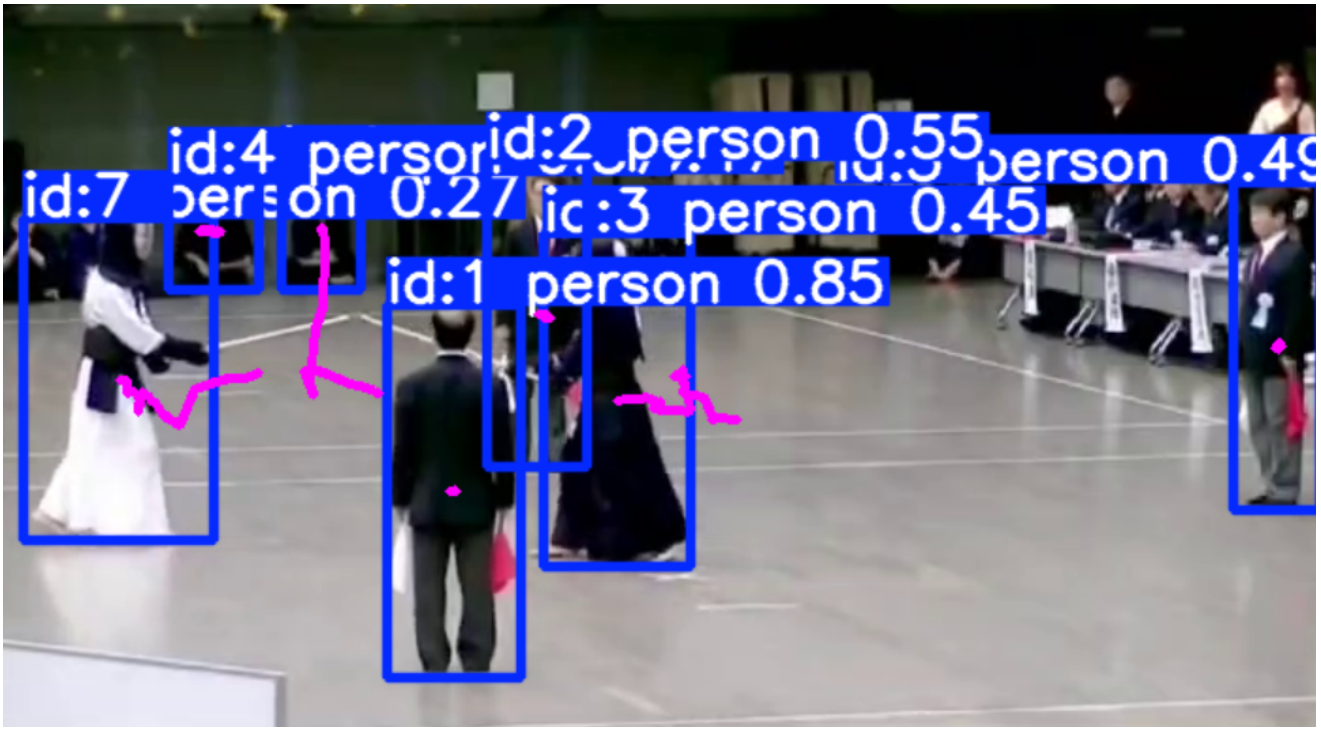
```
  warnings.warn(
Predicted label: ['Abyssinian']


进程已结束，退出代码为 0
```

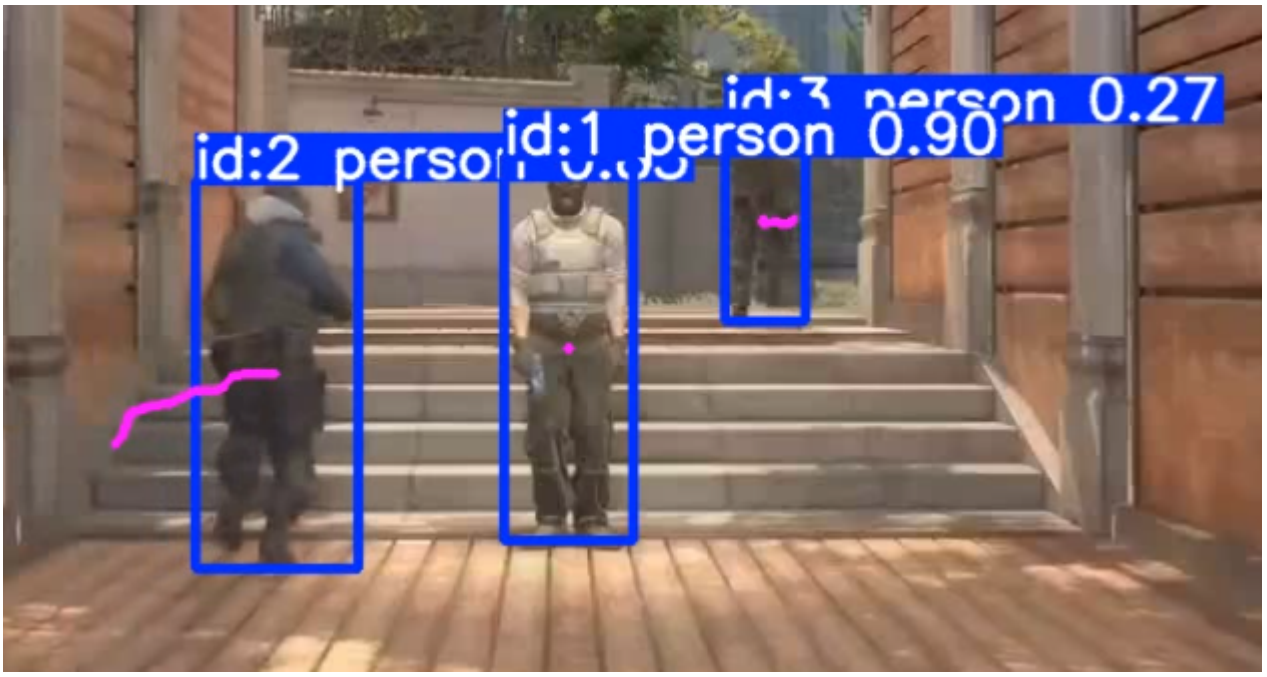
提示

虽然本题用了一些方法对目标文件进行了归类，但你仍然可以继续通过自己的方法重新归类数据集。
可以试着在GitHub等网站中搜索一些项目到本地复现的方式来学习理解。

4. OpenCV+YOLO综合实践（选做）

本题要求学习后自行选用合适的模型实现基础的视频的多目标跟踪。
下面是简单成果演示，大家可以挖掘更加有意思的功能：





本题目标实现不做限制，建议大家通过学习，合理利用网络来实现想要的功能。

在完成本题的基础上，大家也可以探究更多的应用，也可以研究算法的原理,。拓展方向按兴趣即可。如果不能在本地实现该项目，也能提交相关理解。

本题提交：

包含学习笔记（.md），相关截图，项目代码文件夹提交

5. Natural Language Processing（选做）

> *邓焯同学在网上冲浪的时候，忽然看到了某二字游戏OO的评论区，在看到两方阵营魔法对轰的同时，小邓同学也在想，能不能实现一个AI，使得它可以帮小邓看完这个群魔乱舞的评论区，并且为他总结一下评论区的情感倾向呢?*

简而言之

自行搜集资料，用你喜欢的方法，自行实现一个情感分析模型，用于判断评论的情感倾向。你可以选择使用任何一种机器学习或深度学习方法来完成这个任务。
没有完成这道题也没关系，重点在于你在解题的过程中学习到了什么。

⚠ 注意

可以先用senta等已实现的情感分析库走一遍流程，但建议要使用pytorch、tensorflow或Jieba等库辅助你搭建自己的模型哦！
没办法自己搭建模型也没关系，尝试用python的绘图库绘制你的情感分析结果（假设你实在无法搭建自己的模型，必须求助于开源的情感分析系统）

5.1 数据集

使用IMDb电影评论数据集，这是一个广泛用于情感分析的数据集，包含大量的正面和负面电影评论。数据集可以从以下链接下载：
Sentiment Analysis (stanford.edu)

5.2 基本要求

1. 清洗数据（请自行搜索数据有哪些清洗方式）
2. 将文本转换为模型可以处理的数值形式（文本可以转换成什么呢？）
3. 选择你喜欢的机器学习算法训练模型（哪些模型可以用于实现情感分析？）
4. 使用一个分数指标来评估你的所有模型（确保要理解你的分数指标为什么可以用于评估情感分析模型）

5.3 进阶要求

> **温馨提示**
>
> 完成了基本要求，我们其实就已经非常欢迎您来到我们工作室了。
> 除非您在10天以内完成了其余所有机器学习招新题内容，否则非常不建议着手攻克本进阶要求。

- 实现基于预训练模型（如BERT、RoBERTa）的情感分类器。
- 在情感分类之外，增加对情感强度（如情感极性分数）的分析。
- 阅读论文[1706.03762] Attention Is All You Need (arxiv.org)，并记录阅读笔记，如果你提交了阅读笔记，面试时将着重询问**Transformer**架构、**注意力机制**以及NLP的相关知识，**请谨慎提交。**
- 自行实现一个Transformer，完成以上的情感分析任务。

5.4 提交要求

- 你的所有源代码
- 你针对本题的所有学习笔记
- 任何文件应当规范命名，文件名应当有意义
- 如果你做了进阶要求里的内容，请将进阶要求的所有代码和文件与基本要求的文件分离，分别装在两个文件夹中提交

6. 最后

别往下翻了哥们，没有了，招新题就这些。