

**数据挖掘实习一**

**中国地质大学（武汉）地理与信息工程学院**

**2021年 6 月 18 日**

# Lab\_1 情感数据集的标注和分析

## 一、实习目的

1、了解7种基本面部表情及3种复合表情的基本知识和标注方法；

2、理论结合实践，对人脸数据进行标注并对标注结果进行分析；

3、掌握解决数据集中类别不均衡问题的方法。

## 二、实习题目

视频情感数据集的标注和分析

#### 【实验数据】

未标记的视频样本数据。

#### 【实验要求】

1、以2-3个人为一组进行数据集标注，其中每一组分配约450个视频样本；

2、对标注后的数据进行分析，例如分析标注后的数据整体的类别分布情况；

3、将数据集划分为训练集和测试集，分析并解决训练集中类别分布不均衡的问题。

#### 【过程描述】

1、数据处理

首先配置环境，下载安装打标软件，成员按照的自己的想法同时对相同的数据文件进行打标处理，生成标签文件，获得打标文件汇总。

先将程序导出的txt文件做预处理，提取到每一个视频号对应的表情，然后把三个人的数据作对比，有两个人及以上的人标注一样的表情，就输出到sameData，不一样的保留三个人的标注数据输出到diffData。

对于三名成员标注均不同的标签，我们又从原数据中把diffDate中的数据提取了出来，进行了第二轮集体标注，用标注后的新标签覆盖三人原来的标签，然后再跑一遍上文的投票决定标签的程序，所有数据都会在sameData中，不会产生diffDate，sameData即为最终标注数据。

2、类别分布

最终标注结果的情感类别分布如下，其中happiness、helplessness、sadness数量偏多，disgust、disappointment、fear偏少。结合个人标注与集体标注的经验分析：

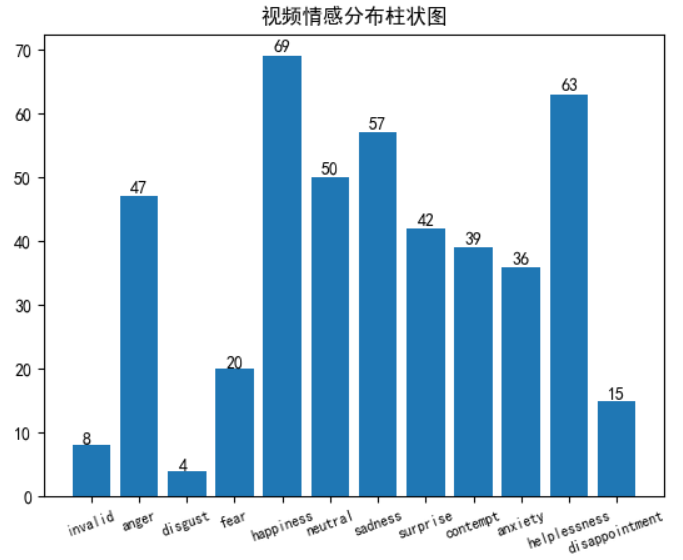
①happiness偏多：有一些场景中，人物的表情确实笑了，但是ta可能出于礼貌或者还带有其他情感，而我们没有过多考虑，都一律标注成了开心。

②helplessness过多：我们把无助与无奈都标成了helplessness，我们的理解中无助与无奈有一些差别，无助的表情失落没有过多表情与动作，无奈是接连摇头或摊手。

③sadness偏多而disappointment偏少：在人物感情两者兼具的时候，可能我们标注的时候更倾向于标成sadness。

④disgust偏少：在disgust与contempt同时出现的时候，更倾向于标disgust。

⑤fear偏少：样本数量中原本带有fear感情的数据偏少。



图一 类别分布图

1. 数据增强

由于视频后续需要进行抠脸操作，为了保证识别效果，我们在数据增强方面主要采用了视频旋转、镜像、翻转、高斯噪声等操作，不对视频进行阴影明暗和遮挡剪裁之类的处理。

首先由于invalid类是无效的，所以按照标签先筛除了invalid类。然后视频在实践过程中按帧进行操作，最后生成新的视频，同时根据原未增强视频重写生成新的标签文件。我们将每个类别的视频数量增加到60。第一轮循环对30个以下的视频进行镜像处理，数量翻倍；第二个循环对30个以上，不满60个的视频按照需要补充的个数进行翻转处理，补齐60个。此时仅剩disgust类别不满60，因此对类别进行单独处理，增加高斯噪声，使达到60个数量。

1. 划分数据集

按照4:1的比例划分测试集和训练集。即从每一个类别中抽取4/5视频和对应的标签移入训练集文件夹，剩余视频和标签移入测试集文件夹。

1. 人像抠图

使用助教给的代码，另加修改，将目标视频都移动到同一文件夹下，扫描文件夹下的视频，每一个视频在save文件夹中生成对应的新的文件夹，然后使用助教的裁剪代码对视频裁剪，获取到对应的人脸截图。这里运行程序比较烧显卡，由于缺帧有时候会出一些bug，修改了代码之后需要重新运行。

#### 【改进设想】

增加标注程序里标签的种类，如绝望的、无奈的、痛苦的、担忧的、激动的等。

## 三、本次实习小结

本次实习中，我的主要任务是从总数据中挑出三人标注均不同的数据，以及对预处理好的标签进行情感类别统计与分析。在标注与分析中，我感受到了本次实习难度很大，因为在我们标注的过程中存在比较大的分歧，产生分析的原因是视频中人物的情感本身就很复杂，很难在给定的标签中选择一个或几个来确定ta的情感，因此可能会对后续训练的准确性产生较大的影响。

=======================================================

我的联系方式：

电话：

邮箱：