

爱奇艺多模态视频人物识别挑战赛

团队名称：emmm...

团队成员：侯昊迪

➤团队介绍

➤赛题背景

➤赛题分析

➤解决方案

团队介绍

团队名称：emmm...

最终成绩：0.8252

团队成员：侯昊迪

团队排名：5

所在组织：南京大学

赛题背景

深度学习

- 人脸识别
- 语音识别

现实挑战

- 多姿态
- 跨年龄
- 局部遮挡

应用需求

- 视频用户的个性化需求
- 视频广告投放效率

赛题分析

赛题任务：从开放视频数据中，识别出包含目标人物的视频

训练集


包含4934个目标
人物的视频片段

验证集/测试集

目标人物
视频片段

非目标人
物视频片
段

准确检测出包含目标人物的视
频片段，并识别目标人物ID



赛题分析

数据形式

视频片段

人脸特征

赛题难点

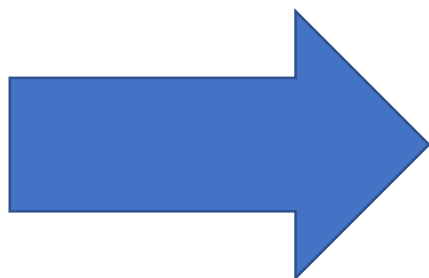
- 原始视频数据量巨大，包含的信息复杂多样；
- 如何对视频片段中的人物进行表示
- 如何对某一目标人物进行表示
- 如何提高模型对非目标人物的抗干扰能力

解决方案

- 原始视频数据量巨大，包含的信息复杂多样

数据解压耗时一天

人脸特征质量较高



战略放弃原始视频数据

解决方案

■ 如何对视频片段中的人物进行表示

视频人物表示难点

- 同一视频可能包含多个人物
- 不同视频帧的人脸质量不一

可能的解决方案

- 对整个视频特征取平均
- 利用质量分数等过滤部分特征后取平均
- 以质量分数为权重，对视频特征取加权平均

解决方案

■ 如何对某一目标人物进行表示

单特征表示

优势

- 查询方法简单直接
- 计算量小

劣势

- 难以表示同一人物的多种形态
- 查询阶段抗干扰能力低

多特征表示

- 能够表示同一人物的多种不同形态

- 对训练数据噪声容忍度低
- 查询方法比较复杂
- 计算量大

使用神经网络分类层各类的法向量作为对应人物的特征表示

- 可以进行端到端的训练
- 可以直接用分类概率作为查询置信度
- 训练和预测的过程一致

解决方案

■ 如何提高模型对非目标人物的抗干扰能力

利用验证集中的非目标人物数据训练模型

- 1、噪声数据扩充：通过拼接短视频以及随机截取长视频片段来扩充数据
- 2、噪声数据聚类：使用DBSCAN算法对噪声数据做聚类处理，将同一簇的数据作为同一个人物
- 3、将噪声数据添加到模型训练中

解决方案

■ 模型设计与训练方法

输入:

512维视频特征



中间层:

512维全连接层



512维embedding



中间层:

分类层

分类类别

4934目标人物

3691噪声聚类簇 共13559类

4934其他噪声

损失函数: ArcFace Loss[1]

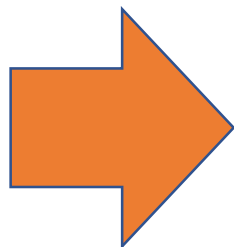
- 首先, 将不属于任何聚类簇的噪声随机初始化为4934个其他噪声类别中的某一类
- 对模型进行训练
- 每训练一轮, 用当前模型对“其他噪声”数据预测, 记预测类别为 l
- 更新“其他噪声”数据的类别标记, 若 $l < 4934$, 则类别更新为 $l + 8652$, 否则, 类别标记不变。

[1] Deng J, Guo J, Zafeiriou S. Arcface: Additive angular margin loss for deep face recognition[J]. arXiv preprint arXiv:1801.07698, 2018.

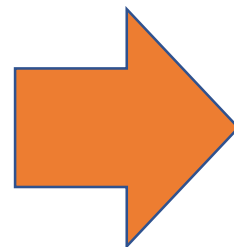
解决方案

■ 模型集成

将所有数据平均
划分为8份



用7份数据做训练，1份数据做验证，可得到8个模型



用8个模型对测试数据分别进行预测，并将预测概率相加作为最终结果

谢谢！
请各位专家批评指正