爱奇艺多模态人物识别挑战赛

队伍：炸天

郑杰鑫　香港科技大学
毛润泽　香港城市大学
杨金宇　香港科技大学
许晓淙　中山大学
罗逸轩　中山大学

目录
CONTENT

概述

Abstract

概述
Abstract

单模型

常规模型：**0.7464**

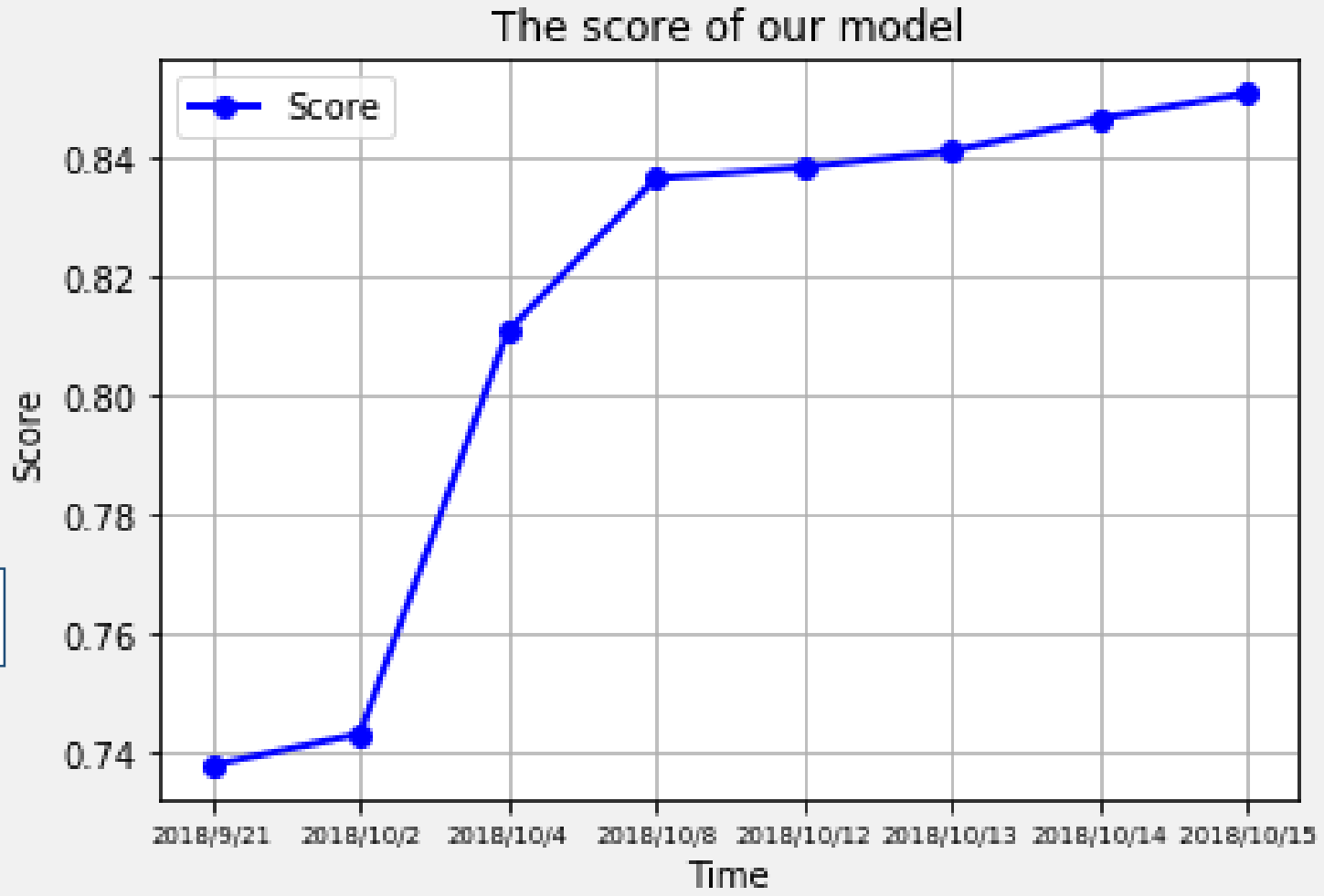
考虑噪声数据：**0.8106**

使用数据增强：**0.8259**

多模型

随机采样 + 选择图片质量：**0.8381**

场景识别 + 特征向量： **0.8505**



模型演进

Model procedure

模型演进

Model procedure

阶段1 自训练模型

对VGG2数据集做模糊处理， 使得预训练好的Arcface模型能够适用于视频人脸识别，再将得到的模型在比赛的训练集上微调。

Reference

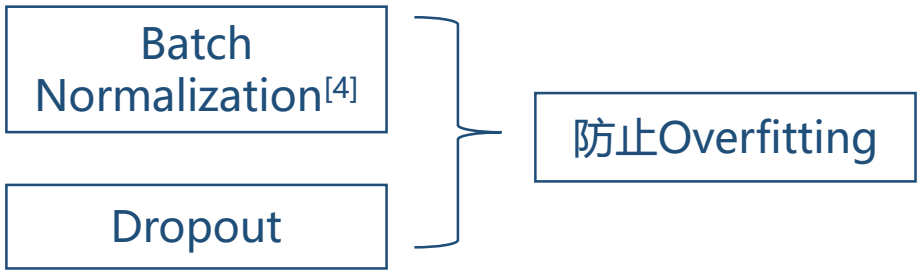
[1] Ding, C., & Tao, D. (2018). Trunk-branch ensemble convolutional neural networks for video-based face recognition. IEEE transactions on pattern analysis and machine intelligence, 40(4), 1002-1014.

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

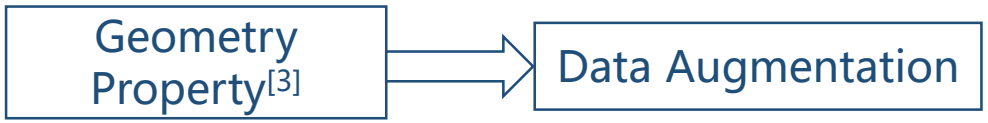
模型演进

Model procedure

阶段2 基于特征向量的建模



基于提供的特征向量，构建分类模型。使用Batch normalization、Dropout增加模型的表现。



基于提供特征的几何特性，对其进行数据增强，使得测试集数量增加一倍，降低预测结果的Variance。

Reference

[3] Schroff, F., Kalenichenko, D., & Philbin, J. (2015). Facenet: A unified embedding for face recognition and clustering. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 815-823).

[4] Ioffe, S., & Szegedy, C. (2015). Batch normalization: Accelerating deep network training by reducing internal covariate shift. arXiv preprint arXiv:1502.03167.

模型演进
Model procedure

阶段3 场景识别

SE-ResNext[5] → 45% accuracy in validation dataset

出于速度和准确率的考虑，使用SE-ResNext模型，对视频的原始数据进行训练。每个视频取第一和最后一帧训练。

阶段4 模型融合

特征向量建模
场景识别
→ 最终模型

将阶段2和阶段3得到的模型进行融合，得到最终结果。
对于人脸质量较差的情况，SE-ResNext模型的预测结果占比较高。
其占比随着人脸质量的提高而减小。

Reference
[5]Hu, J., Shen, L., & Sun, G. (2017). Squeeze-and-excitation networks. arXiv preprint arXiv:1709.01507, 7.

最终方案详解

Final model

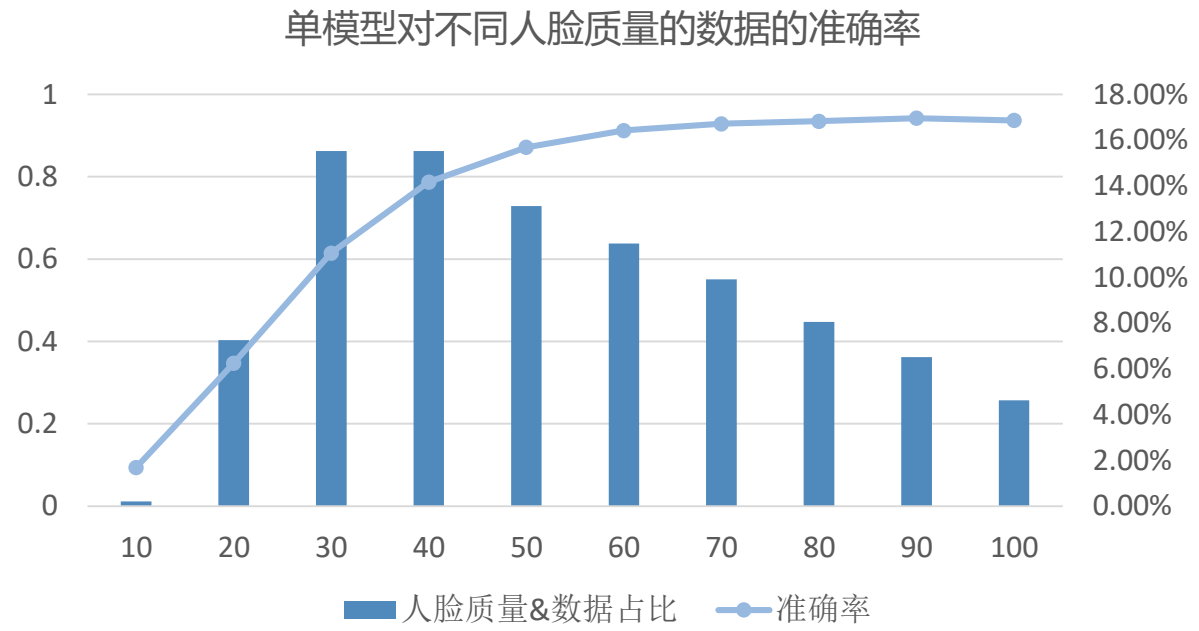
最终方案详解

Details of final model

基于特征向量的模型

训练：基于提供的人脸特征，构建简单的神经网络模型进行分类。

预测：对一个视频的所有特征向量进行预测，将得到的预测结果取平均，作为对应视频的预测结果。



多模型思路：

- 在每次训练中，随机选取80%的数据进行训练
- 取不同质量区间的数据进行训练，[0,200]、[20,200]、[0,80]...
- 取多个模型输出的均值

最终方案详解
Details of final model

基于特征向量的模型

Batch Normalization——加快训练速度，增加模型的鲁棒性；

Dropout——增加模型的鲁棒性；

Data Augmentation——对测试集的数据进行数据增强，减小输出结果的variance。

↓

原理：

同一个人的特征向量的集合是一个凸集（Convex），所以集合内任意向量的均值，仍然在范围之中。

操作：

随机取一个视频的2/3/4/5/6个特征向量，计算其均值后将其作为新的特征向量加入到数据集中。

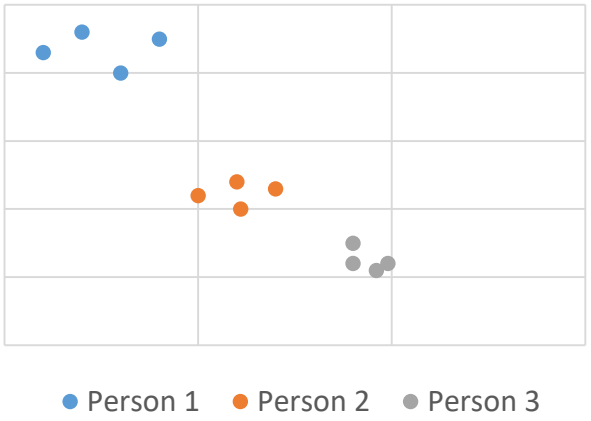
由于该方法不会改变数据集的边界和分布，对训练集进行增强的话，效果微小。

对测试集数据进行增强，可以降低最后预测结果的Variance。

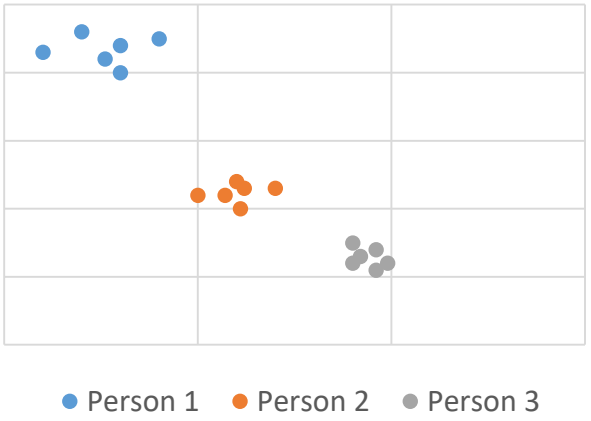
思考（赛后）：

对训练集的增强，应该选择同一个人的不同视频的特征向量进行。

增强前



增强后



最终方案详解

Details of final model

基于场景识别的模型(SE-ResNext)

训练：取每个视频选取第一帧和最后一帧图片进行训练（计算资源限制），大小为224*224

预测：对每个视频的所有帧进行预测，最终将所有预测结果平均，作为对应视频的预测结果

模型融合

视频情况	融合原则
视频未检测到人脸	$O_{final} = 0.8 * O_{ResNext}$
人脸平均质量∈[0,20)	$O_{final} = 0.5 * O_{ResNext} + 0.5 * O_{feature}$
人脸平均质量∈[20,30)	$O_{final} = 0.4 * O_{ResNext} + 0.6 * O_{feature}$
人脸平均质量∈[30,40)	$O_{final} = 0.3 * O_{ResNext} + 0.7 * O_{feature}$

思考

Reflection

思考
Reflection

比赛相关

- **噪声处理**——多模态数据应用空间广阔，但增加信息量的同时，额外引入的噪声数据较难处理。
- **对数据的理解**——对工业界的应用而言，除了识别人物之外，人物识别可以带来更多的价值。比如识别人物所处的环境（刘德华在综艺、刘德华在演唱会、刘德华演古装戏等），可以帮助更好的对视频进行分类和需要时的检索。 但目前的数据集仍欠缺较为细粒度的标签。

其他

- 将每个视频生成为特征向量，即video to vector，可以大大增加实际使用时的便利。在有新的样本或类别时，无需重新训练分类器。但目前仍然没有video to vector的方法，更无法生成具有几何特性的feature。
- 常规方法处理视频数据需要较多的计算资源。
- 数据集可以拓展为基于人物（明星）的视频理解（video-understanding）

Thanks & Discussion