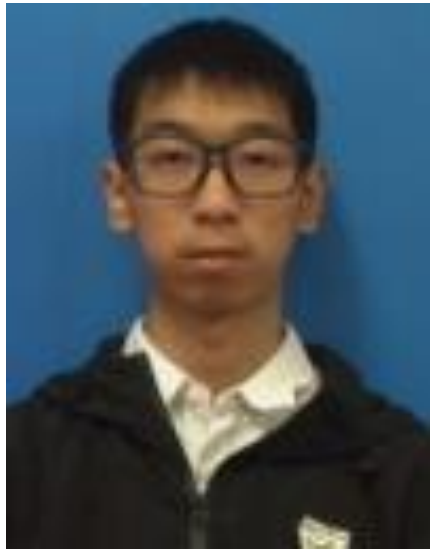
Bi-weekly Report
Voiceprint vs. Face Recognition

报告人：江昊宇、侯瑞海

2021.10.15

自我介绍

- 江昊宇



西北民族大学大学 信息院 研二

- 侯瑞海

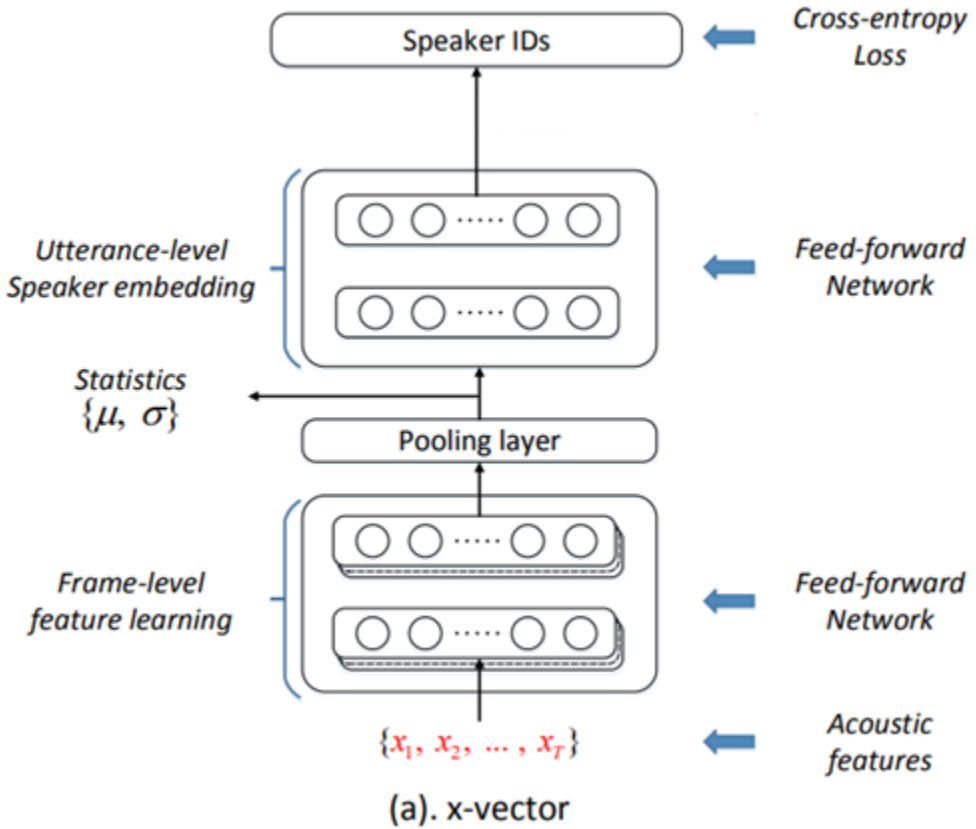
北京邮电大学 计算机学院 大三

出发点

- 说话人识别、人脸识别都有各自的 **局限性**
- 不同模态之间的 **信息互补**
- 单模态局限性可以通过其他模态进行修正
- 提高系统识别性能和顽健性

https://ieeexplore.ieee.org/abstract/document/9350195/

说话人识别



(a). x-vector

Table 1: *Extended TDNN x-vector architecture*

Layer	Layer Type	Context	Size
1	TDNN-ReLU	t-2:t+2	512
2	Dense-ReLU	t	512
3	TDNN-ReLU	t-2, t, t+2	512
4	Dense-ReLU	t	512
5	TDNN-ReLU	t-3, t, t+3	512
6	Dense-ReLU	t	512
7	TDNN-ReLU	t-4, t, t+4	512
8	Dense-ReLU	t	512
9	Dense-ReLU	t	1500
10	Pooling (mean+stddev)	Full-seq	3000
11	Dense(Embedding)-ReLU		512
12	Dense-ReLU		512
13	Dense-Softmax		Num. spks.

模型概要

训练模型使用的网络是Fast Resnet34，使用的损失函数是AM-softmax

$$\mathcal{L}_{AMS} = -\frac{1}{n}\sum_{i=1}^{n} log\frac{e^{s\cdot(cos\theta_{y_i}-m)}}{e^{s\cdot(cos\theta_{y_i}-m)}+\sum_{j=1,j\neq y_i}^{c} e^{s\cdot cos\theta_j}}$$
$$= -\frac{1}{n}\sum_{i=1}^{n} log\frac{e^{s\cdot(W_{y_i}^T f_i-m)}}{e^{s\cdot(W_{y_i}^T f_i-m)}+\sum_{j=1,j\neq y_i}^{c} e^{sW_j^T f_i}}. \quad (6)$$

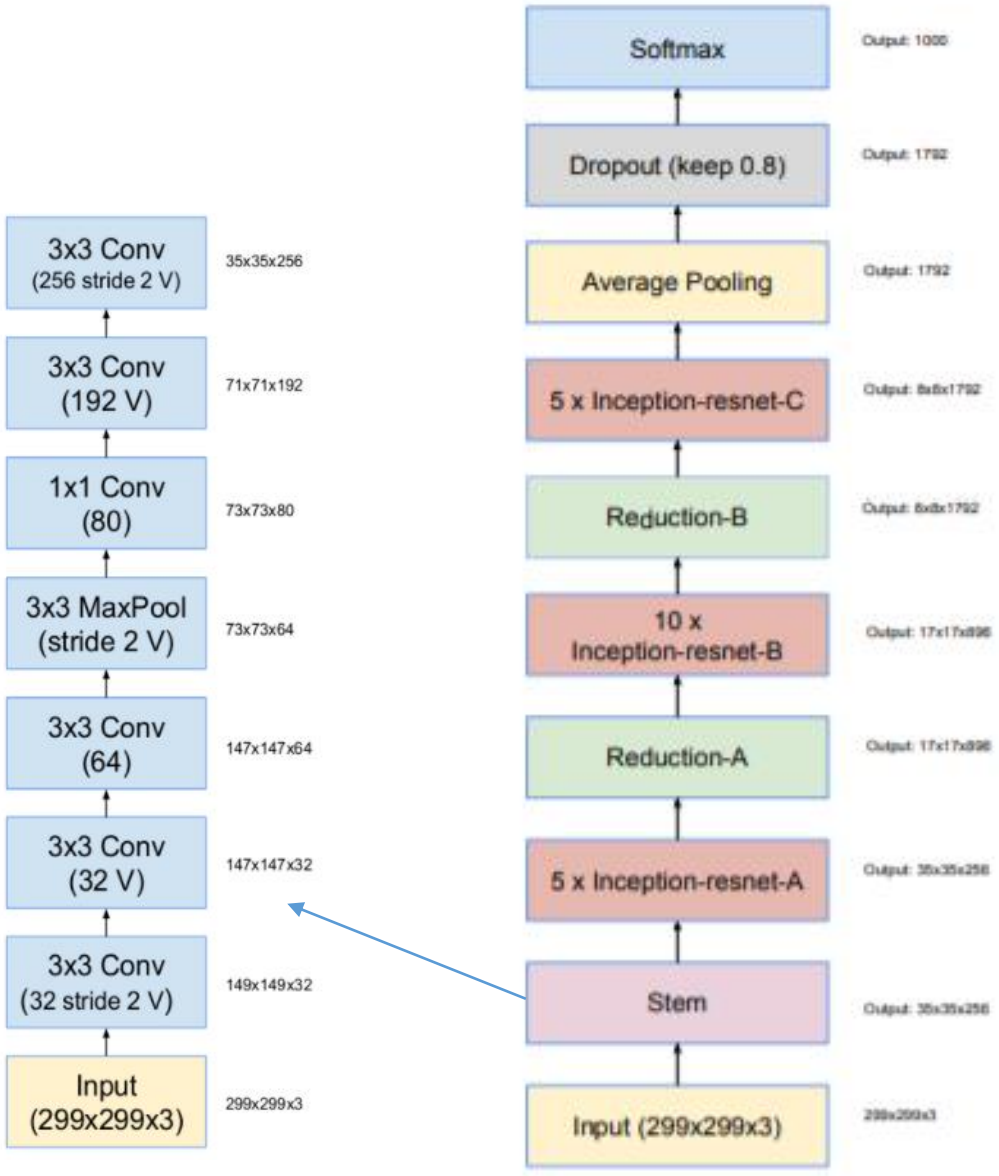
- AM-Softmax loss 能进一步增大类间差异，缩小类内差异，从而达到优化余弦距离的效果。
- Fast Resnet34性能与Thin ResNet34的两种模型不相上下，而计算成本却不到它们的一半。

https://arxiv.org/pdf/2003.11982.pdf

人脸识别模型

使用网络：Inception-ResNet V1

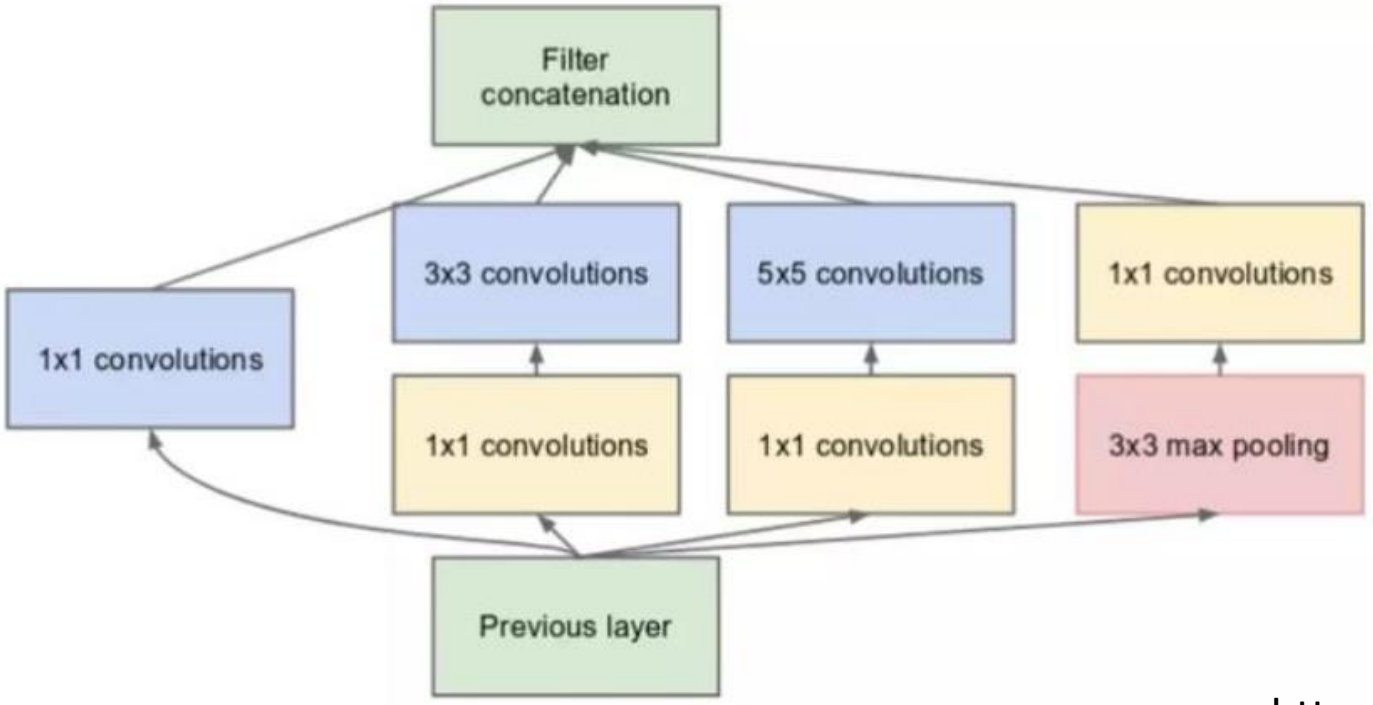
采用了Inception模块来减少参数的数量，同时增加了网络的宽度



https://arxiv.org/abs/1602.07261v2

Inception 模块

- 一个block中同时使用了1*1，3*3，5*5的卷积，增加了网络对尺度的适应性
- 1*1卷积用于降维，减少 weights大小 和 feature map维度



https://arxiv.org/abs/1409.4842

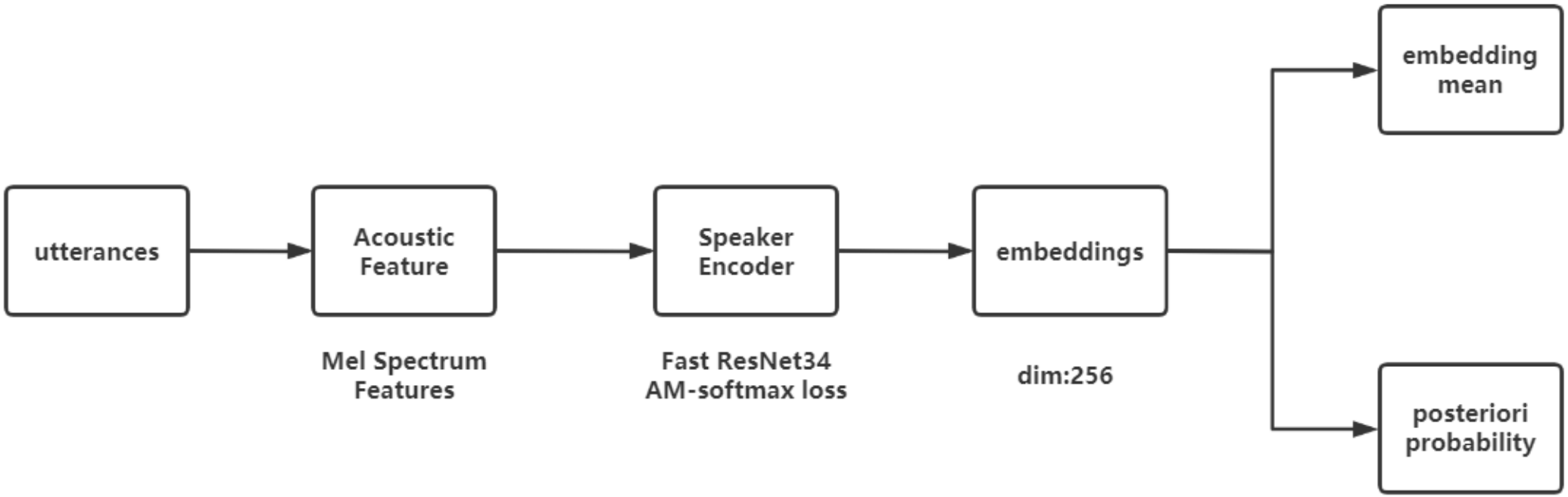
VoxCeleb 音视频多模态数据集

	dev	test
# of speakers	5,994	118
# of videos	145,569	4,911
# of utterances	1,092,009	36,237

数据分布

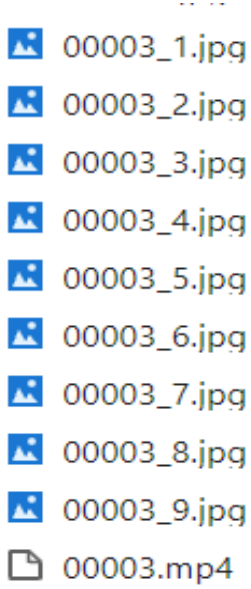
dataset	speaker_num	train	test	dataset	speaker_num	train	test
audio	50	9140	726	image	50	73002	5834
	100	16593	1458		100	122243	10370
	200	34092	2728		200	242955	18708
	1000	160036	13600		1000	1168935	94789

说话人识别流程


utterances
Acoustic Feature
Mel Spectrum Features
Speaker Encoder
Fast ResNet34 AM-softmax loss
embeddings
dim:256
embedding mean
posteriori probability

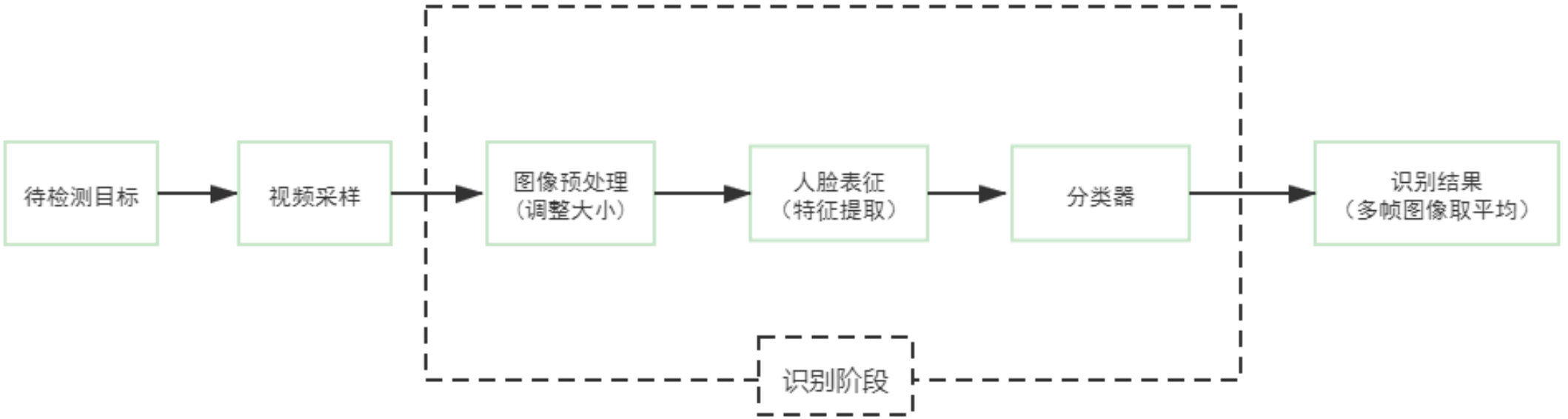
视频数据处理

使用cv2，读取每个视频的fps，以该值为分帧数值对视频进行采样分帧



同一个视频采样出的图片有相同的前缀

人脸识别流程



评测方案

- 基于表征向量的余弦打分

$$\text{embedding}_{speaker}=\frac{\sum_{\text{n}} embedding}{\text{n}}$$

$$similarity(A,B) = \frac{A \cdot B}{\|A\| \times \|B\|} = \frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n} A_i^2} \times \sqrt{\sum_{i=1}^{n} B_i^2}}$$

- 基于后验概率的打分排序

$$P(y|x) = \frac{e^{h(x,y_i)}}{\sum_{j=1}^{n} e^{h(x,y_i)}}$$

$$p_{speaker}=\frac{\sum_{\text{n}} \log \text{p}_{\text{ image}}}{\text{n}}$$

测试结果

基于表征向量的余弦打分

system	speaker_num	acc	system	speaker_num	acc
speaker identification	50	75.482%	face identification (pretrained model)	50	93.260%
	100	72.702%		100	92.857%
	200	79.802%		200	92.363%
	1000	84.250%		1000	90.744%

随着类的个数的增加，说话人识别的准确率是呈上升趋势

随着类的个数的增加，人脸识别的准确率在逐渐下降

注：人脸识别识别系统采用 Inception ResNet v1 预训练模型

测试结果

基于后验概率的打分

system	speaker_num	acc	system	speaker_num	acc
speaker identification	50	78.698%	face identification (pretrained +finetune)	50	65.471%
	100	75.037%		100	63.064%
	200	79.391%		200	61.362%
	1000	85.175%		1000	59.033%

说话人识别准确率呈上升趋势，后验概率总体效果要略好于表征向量方法

人脸识别准确率依旧在下降，且后验概率方式的总体准确率偏低

注：人脸识别识别系统采用 Inception ResNet v1 预训练模型 + Finetune

总结

- 随着类别数的增加，说话人识别的准确率总体是呈上升趋势，其潜在原因是随着训练数据增多，模型训练收敛更优，性能逐渐提升。
- 随着类别数的增加，人脸识别的准确率则逐渐下降，其原因是预训练模型收敛稳定，在此基础上，随着类别数增加，测试任务逐渐困难，性能自然下降。
- 总之，当前实验存在一个训练模型对训练数据的依赖和测试任务难易程度的权衡问题。当前实验设计将训练数据和测试任务两个变量混淆在一起，致使实验结果存在偏差，还需进一步探索。

下一步工作

- 增大测试数据集的规模，找到声纹识别拐点
- 简化inception resnet模型，重新训练人脸模型进行比较
- 进行加噪测试，对比两种模态识别性能的鲁棒性
- 探索特征、模型域的多模态融合方法

谢谢大家