1021-1027进展

1.几篇论文阅读概述

参考的几篇论文都来自于之前提供的文档，主要看了有代码的几个，包括：

Feature Fusion-based Forgery Detection类别下的：**Betray oneself**: A novel audio deepfake detection model via mono-to-stereo conversion

Hybrid Feature-based Forgery Detection类别下的：**One-class learning** towards synthetic voice spoofing detection

End-to-end Forgery Detection类别下的：End-to-end anti-spoofing with **RawNet2**(2021)

这些论文都有相关的参考代码，且都基于了ASVSpoof2019数据集开展了相关实验。

但是代码方面也有一些问题：

1.**Betray oneself**提供的代码不完整

2.**One-class**提供的代码需要matlab预处理数据，linear frequency cepstral coefficient (**LFCC**)

Betray oneself

研究背景概括：

任务: Audio Deepfake Detection，

fake audio的具体类型: (1)TTS：text-to-speech，(2)VC：voice conversion，(3)replay 等

主要特点：

首先使用要给预训练好的立体声生成器，将单通道音频生成为立体声，再基于此立体声完成检测任务；通过融入立体声信息能够提高ADD（听觉场景描述）性能。

本质可以理解为一种数据增强。

模型架构：

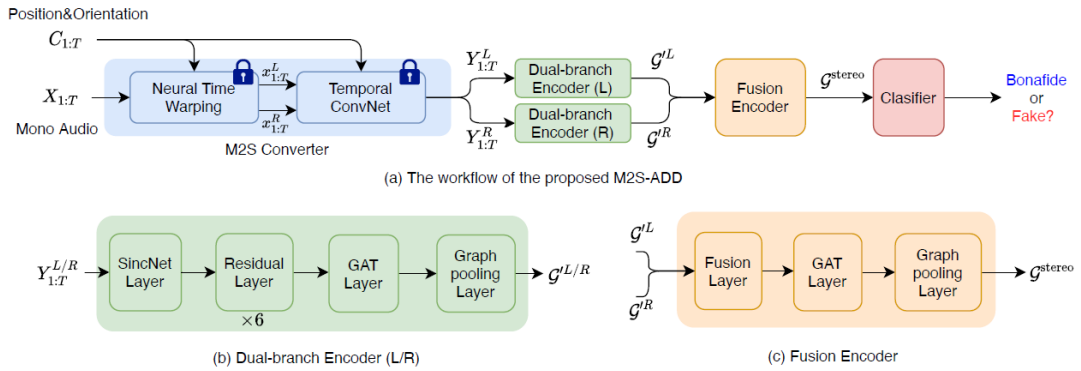


Table 1: *The configuration for our M2S-ADD model.*

Layer	Input:64600 samples	Output shape
	Dual-branch Encoder(L/R)	
SincNet Layer	Conv-1D(129,1,70) Maxpool-2D(3) BN&SeLU	(70,64472) (1,23,21490)
Residual Layer	{ Conv-2D ((2,3), 1,32) BN & SeLU Conv-2D ((2,3), 1,32) Maxpool-2D ((1,3)) } × 2	(32,23,2387)
	{ Conv-2D ((2,3), 1,64) BN & SeLU Conv-2D ((2,3), 1,64) Maxpool-2D ((1,3)) } × 4	(64,23,29)
GAT Layer	GAT Layer	$left : (32, 23)$ $right : (32, 29)$
Graph Pooling Layer(1)	Graph pooling	$left : (32, 14)$ $right : (32, 23)$
	Projection	(32,12)
	Fusion Encoder	
Fusion Layer	Element-wise multiplication	(32,12)
GAT Layer	GAT Layer	(16,12)
Graph Pooling Layer(2)	Graph Pooling	(16,7)
	Projection	(1,7)
	Classifier	
Classifier	FC(2)	2

主要组成包括：

M2S Converter --- 单通道到双通道转换

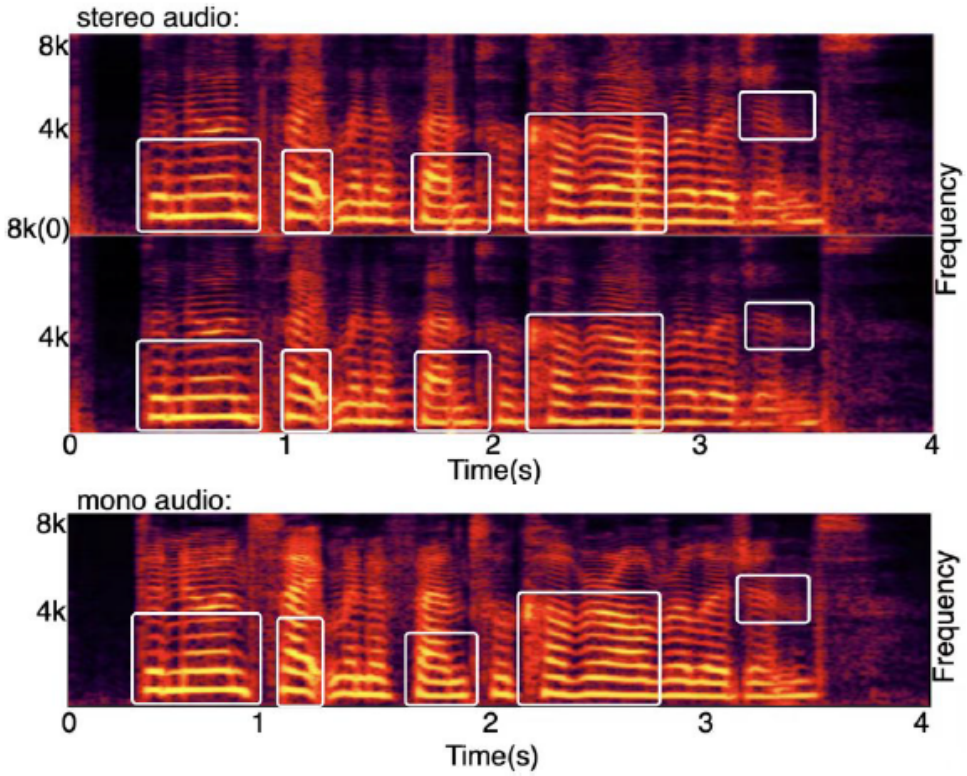
Dual-branch Encoder -- 分别对每个通道进行特征提取。结构上是SincNet + Residual layers(一组卷积层) + GAT + Graph pooling.

其中GAT是图注意力网络，汇聚前面卷积层的输出

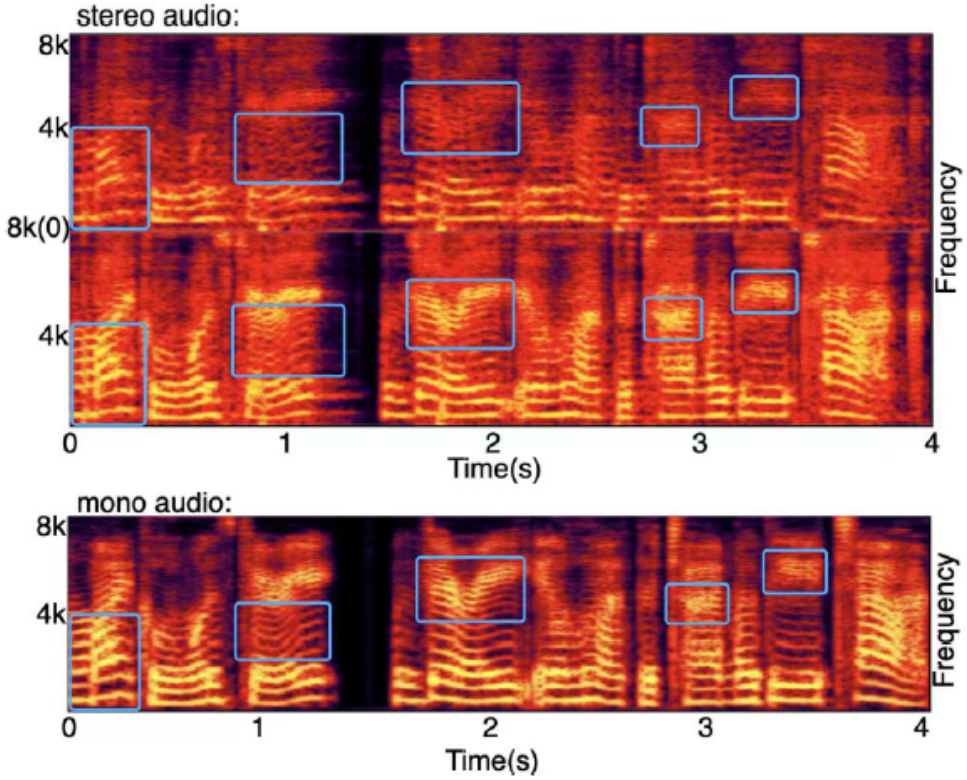
Graph pooling是图池化层

Fusion Encoder and Classifier -- 将两个通道的输出混合到一起，并进行分类处理。

多通道的主要效果：



(a) Bonafide audio



(b) Fake audio

Figure 2: *Comparison of visualization analysis in the spectral domain. (a) shows the spectral details of the mono and stereo signals for bonafide audio; (b) shows the spectral details for fake audio. Unlike the white boxes, the blue boxes show that the spectral artifacts are particularly exposed when the mono fake audio is converted to stereo audio.*

（a）展示了真实音频的单声道和立体声信号的频谱细节；（b）展示了伪造音频的频谱细节。与白色方框不同，蓝色方框表明当单声道伪造音频转换为立体声音频时，频谱伪迹尤为明显。

One-class learning

主要特点

one-class：训练时只有正常数据，测试时包含异常数据。

核心内容：

在否定训练集和测试集分布类似的假设的基础上，基于单类训练方法，提升了语音检测模型泛化性：对于那些未曾在训练集中见过的合成语音类型，都能取得较好的检测。---追求能够检测训练数据中未出现的攻击的通用型欺骗对策，在野外环境中，欺骗攻击的性质无法预测并将持续演变

主要参考价值：

one-class训练场景和loss设计，使用t-SNE和PCA的数据可视化(应该可用)。可以理解为基于其他语音检测主干模型实现的一种改进。

相关基础架构(在RawNet2中能查找到介绍)：

LFCC：It uses 70 linearly-spaced filters with
conventional cepstral analysis and a GMM back-end classifier.(简单快速)

RawNet2

主要特点：

end-to-end模型；认为基于手工构建的特征再实现处理，不如直接将数据映射到特征空间的特征；

模型架构：

初步编码 + 卷积神经网络(主干网络采用ResBlock) + GRU + 分类头

Layer	Input≈**64000** samples	Output shape
Fixed Sinc filters	Conv(**129**,1,128) Maxpooling(3) BN & LeakyReLU	(**21290**,128)
Res block	{ BN & LeakyReLU Conv(3,1,128) BN & LeakyReLU Conv(3,1,128) Maxpooling(3) FMS } × 2	(**2365**,128)
Res block	{ BN & LeakyReLU Conv(3,1, **512**) BN & LeakyReLU Conv(3,1, **512**) Maxpooling(3) FMS } × 4	(**29**,512)
GRU	GRU(1024)	(1024)
FC	1024	(1024)
Output	**1024**	**2**

一些改进：

输入处理：我们将原始波形输入的持续时间固定为约4秒（64000个样本），通过裁剪长语音或拼接短语音来实现。

2.代码复现情况

目前成功运行了这个RawNet2的模型训练，看起来不是很快，项目代码具体理解和改进下一步(10.28-11.3)再深入考察。

3.其他：数据集情况概述

AVSpoof2019

参考其提供的说明文件，

1)包含PA和LA两部分的数据集：

PA: physical access

LA: logical access

2)PA和LA

每个数据集下都包含了

train数据集，dev数据集，eval数据集，protocols，asv_protocols，asv_scores

3)数据格式(train, dev, eval)

都是16kHz的采样率, 16-bit的音频数据，后缀为flac

4)protocols(协议)

包括一组protocols，其中

a.LA

txt文件，如LA.cm.train.trb.txt；

其中内含包含文件名--标签类型

bonafide -- 正常

spoof -- 合成的

b.CMprotocol

合成数据还包括不同类型：

A01-A19

c.ASV protocols

格式为<1>.<2>.<3>.txt

数据集类别，性别，类型

d.baseline ASV scores

由基线ASV系统提供，用于t-DCF evalutation我 我

ADD2022

中文数据集，也包含多种异常类型

数据集包含train，dev，label；数据格式为.wav，

人耳确实可以听出来大部分的正常和合成语音的区别