EED 305: Digital Signal Processing

PROJECT PRESENTATION

Group 16

Suchit Reddi (2010110507)

Akshay Veeragandham (2010110963)

Koteswara Bezawada (2010111057)

ONE-CLASS LEARNING TOWARDS SYNTHETIC VOICE SPOOFING DETECTION

You Zhang, Fei Jiang, Zhiyao Duan

“The key idea of one-class classification is to capture the target class distribution and set a tight boundary around it, so that all non-target data would be placed outside the boundary.”

Important Links:

Paper: https://ieeexplore.ieee.org/document/9417604

Implementation: https://github.com/SuchitReddi/AIR-ASVspoof

Video: https://www.youtube.com/watch?v=pX9aq8CaIvk

Dataset: https://www.asvspoof.org/index2019.html

IEEE Signal Processing Letters, Vol. 28, 2021

ABSTRACT

- Human voices can be used to authenticate their identity. Automatic Speaker Verification (ASV) Systems are vulnerable to voice spoofing attacks. Voice spoofing attacks include impersonation, replay, text-to-speech, and voice conversion.
- Impersonation and replay attacks does not have valid databases, hence reducing the chances of research in these types of attacks.
- Researchers are developing anti-spoofing techniques to encounter known attacks. However, they face difficulties detecting unknown attacks, which have different statistical distributions from known attacks. The fast development of voice spoofing algorithms is resulting in increasingly powerful unknown synthetic attacks.
- This paper uses one-class learning to train an anti-spoofing system to detect unknown synthetic voice spoofing attacks. The idea is to capture the target bona fide class and set a tight boundary around it (using OC-softmax loss function), so that all non-target fake class objects will be placed outside the boundary.

Search or jump to... / Pull requests Issues Codespaces Marketplace Explore

CorentinJ / Real-Time-Voice-Cloning Public

Watch 854 Fork 6.9k Star 38k

Code Issues 96 Pull requests 4 Actions Wiki Security Insights

master 1 branch 0 tags Go to file Add file Code

CorentinJ Update README.md 98d0ca4 on Sep 8 295 commits

README.md

Real-Time Voice Cloning

This repository is an implementation of Transfer Learning from Speaker Verification to Multispeaker Text-To-Speech Synthesis (SV2TTS) with a vocoder that works in real-time. This was my master's thesis.

SV2TTS is a deep learning framework in three stages. In the first stage, one creates a digital representation of a voice from a few seconds of audio. In the second and third stages, this representation is used as reference to generate speech given arbitrary text.

About

Clone a voice in 5 seconds to generate arbitrary speech in real-time

CYBERSECURITY

Does your boss sound a little funny? It might be an audio deepfake

Voice deepfake attacks against enterprises, often aimed at tricking corporate employees into transferring money to the attackers, are on the rise. And at least in some cases, they're succeeding.

Audio deepfakes are a new spin on the impersonation tactics that have long been used in social engineering and phishing attacks, but most people aren't trained to disbelieve their ears. | Illustration: Christopher T. Fong/Protocol

By Kyle Alspach | August 18, 2022

As a cyberattack investigator, Nick Giacopuzzi's work now includes responding to growing attacks against businesses that involve deepfaked voices — and has ultimately left him convinced that in today's world, "we need to question everything."

Most Popular

Sponsored Content

Is there a paved road toward cloud native resiliency?

Bulletins

Voice phishing attacks reach all-time high

by Brian Stone in Security
on May 24, 2022, 12:03 PM PDT

A study conducted by Agari and PhishLabs found a five-times increase in attempted vishing attacks from the beginning of 2021 to Q1 of 2022.

Cases of voice phishing or vishing have been reported to have risen a whopping 550% over the past 12 months alone, according to the Quarterly Threat Trends & Intelligence Report co-authored by Agari and PhishLabs. In March 2022, the amount of vishing attacks experienced by organizations reached its highest level ever reported, passing the previous record set in September of 2021.

TYPES OF ATTACKS

Some of the spoofing attacks against ASV systems are:

1)Impersonation: Voice samples from twins or professional mimics.

2)Replay: Pre-recorded audio of target speaker is reused.

3)Text-to-Speech (TTS): Converting text to spoken words with speech analysis.

4)Voice Conversion (VC): Converting voice of source speaker to target speaker using AI.

Replay attacks are the most successful attacks. But there are no databases available on Impersonation and Replay attacks. So, it is difficult to research on those specific types of attacks.

SIMILAR PREVIOUS MODELS

- **Zhang et al.** – Combining Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) improved system robustness.
- **Monteiro et al.** – Deep residual network (ResNet) with temporal pooling.
- **Chen et al.** – ResNet with large margin cosine loss function and applied frequency mask augmentation.
- **Gomez-Alanis et al.** – Light convolution gated RNN architecture to improve long-term dependency for spoofing attacks detection.
- **Wu et al.** – Feature genuinization based light CNN system outperforming all other single systems.
- **Aravind et al.** – Transfer learning approach with ResNet architecture.

Model fusion based on sub-band modelling was introduced, which increased performance, but also model complexity.

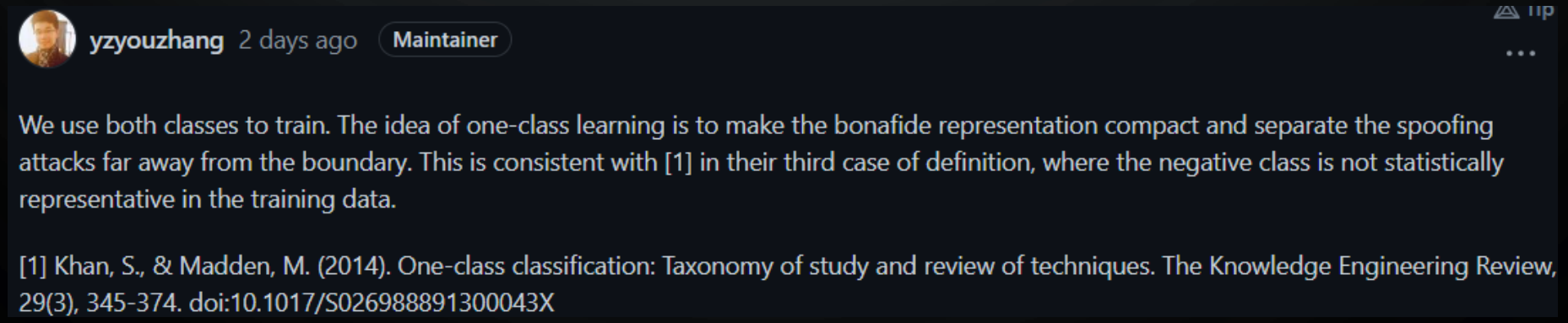
BINARY CLASSIFICATION

- Existing previous models suffered from generalization to unseen spoofing attacks.
- In the training stage, most models used binary classification. But when using binary classification, it is assumed that the distribution between training and test datasets for both target and non-target classes is similar.
- In practice, target or bona fide class has a big training set with diverse speakers.
- But non-target or fake class does not have a training set containing all the known synthetic attacks. Due to the development in speech synthesis techniques, the data in training set may never be able to catch up.
- This distribution mismatch between training and test sets for fake class makes this issue a good fit for one-class classification.
- One-class learning idea has been introduced in image forgery detection successfully.

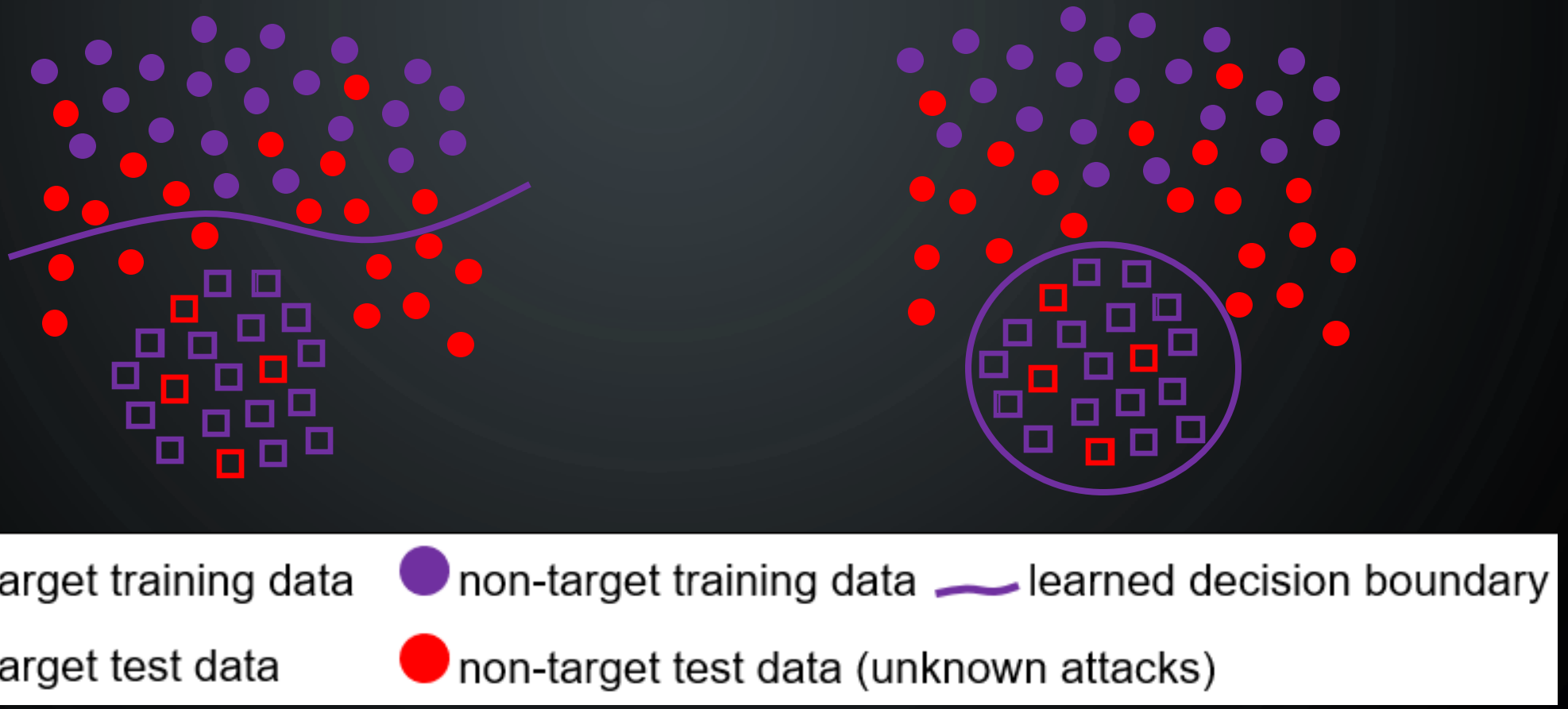
ONE-CLASS CLASSIFICATION

- In one-class classification, one of the classes (positive or target class) is well characterized by instances in the training data. For the other class (negative or non-target), it has either few or no instances at all, or they do not form a statistically-representative sample of the negative class.
- In this paper, speaker verification anti-spoofing problem is formulated as one-class feature learning to improve generalization ability.
- A loss function called one-class softmax (OC-Softmax) is proposed to learn a feature space in which bona fide speech embeddings have a compact boundary while spoofing data is kept away by a certain margin.
- The target class refers to bona fide speech and non-target class refers to spoofing attacks.

➢ **Alegre et al.** employed one-class support vector machine (OC-SVM), trained only on bona fide speech to classify local binary patterns of speech cepstrograms, showing the potential of one-class classification approach.



BINARY VS ONE-CLASS



(a) Binary classification

(b) One-class classification

- In deep learning-based voice spoofing detection models, an embedding vector is calculated for the input speech, which is then scored as bona fide or spoof by the trained model.
- The objective of training this model is to learn an embedding space in which the bona fide and spoof voices can be discriminated. This embedded space is further used for scoring the input speech.
- Previous systems used binary classification loss functions in training their systems to learn the embedding space.
- Widely used binary classification loss functions are Softmax and Angular Margin-Softmax (AM-Softmax).
- The loss function proposed in this paper is One-class Softmax (OC-Softmax).

METHOD

A. Binary Classification Loss Functions:

The original Softmax loss function for binary classification can be formulated as

- Training (Loss):

$$\mathcal{L}_S = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{e^{\boldsymbol{w}_{y_i}^T\boldsymbol{x}_i}}{e^{\boldsymbol{w}_{y_i}^T\boldsymbol{x}_i}+e^{\boldsymbol{w}_{1-y_i}^T\boldsymbol{x}_i}} = \frac{1}{N}\sum_{i=1}^{N}\log\left(1+e^{(\boldsymbol{w}_{1-y_i}-\boldsymbol{w}_{y_i})^T\boldsymbol{x}_i}\right)$$

- Inference (Score):

$$\mathcal{S}_S = \frac{e^{\boldsymbol{w}_0^T\boldsymbol{x}_i}}{e^{\boldsymbol{w}_0^T\boldsymbol{x}_i}+e^{\boldsymbol{w}_1^T\boldsymbol{x}_i}}.$$

- xi ∈ R^D and yi ∈ {0, 1} are embedding vector and label of the i-th sample, respectively. w0, w1 ∈ R^D are the weight vectors of the two classes. N - no.of samples in a mini-batch.

The AM-Softmax loss function can be formulated as

- Training (Loss):

$$\mathcal{L}_{AMS} = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{e^{\alpha(\hat{\boldsymbol{w}}_{y_i}^T\hat{\boldsymbol{x}}_i-m)}}{e^{\alpha(\hat{\boldsymbol{w}}_{y_i}^T\hat{\boldsymbol{x}}_i-m)}+e^{\alpha\hat{\boldsymbol{w}}_{1-y_i}^T\hat{\boldsymbol{x}}_i}} = \frac{1}{N}\sum_{i=1}^{N}\log\left(1+e^{\alpha\left(m-(\hat{\boldsymbol{w}}_{y_i}-\hat{\boldsymbol{w}}_{1-y_i})^T\hat{\boldsymbol{x}}_i\right)}\right)$$

- Inference (Score):

$$\mathcal{S}_{AMS} = (\hat{\boldsymbol{w}}_0-\hat{\boldsymbol{w}}_1)^T\hat{\boldsymbol{x}}_i.$$

- α is a scale factor, m is the margin of cosine similarity.
- wˆ , xˆ are normalized w and x, respectively.
- AM-Softmax introduces an angular margin to make the embedding distributions of both classes more compact.
- The larger the margin m, the more compact the embeddings will be.

METHOD

B. One-class Classification Loss Function:

The OC-Softmax loss function can be formulated as $\mathcal{L}_{ocs} = \frac{1}{N}\sum_{i=1}^{N}\log\left(1+e^{\alpha(m_{y_i}-\hat{w}_0\hat{x}_i)(-1)^{y_i}}\right)$

- It is reasonable to train a compact embedding space for bona fide speech. But if we do the same for fake speech, it may exclude some known attacks.
- To address this issue, two different margins are introduced. A tight boundary is made around the bona fide speech only, isolating the spoofed speech.
- Only one weight vector w0 is used in this loss function for the bona fide class. w0 refers to the optimization direction of the target class vector.

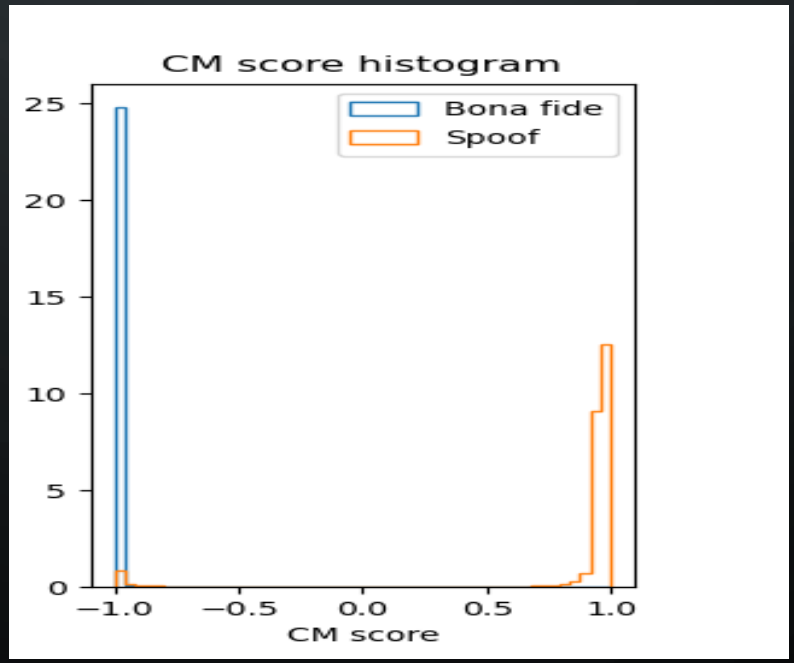
EXPERIMENT

Implementation: https://github.com/SuchitReddi/AIR-ASVspoof

The database is extracted from the ASVspoof2019 challenge LA dataset.

Training code was adopted based on deep residual network (ResNet-18).

Demo will be shown during presentation.



RESULT

One-class - 0.059 2.19

System	EER (%)	min t-DCF
CQCC + GMM [3]	9.57	0.237
LFCC + GMM [3]	8.09	0.212
Chettri et al. [22]	7.66	0.179
Monterio et al. [14]	6.38	0.142
Gomez-Alanis et al. [16]	6.28	-
Aravind et al. [18]	5.32	0.151
Lavrentyeva et al. [21]	4.53	0.103
ResNet + OC-SVM	4.44	0.115
Wu et al. [17]	4.07	0.102
Tak et al. [19]	3.50	0.090
Chen et al. [15]	3.49	0.092
Proposed	**2.19**	**0.059**

Comparison with single systems

ASVspoof 2019 LA scenario							
#	ID	t-DCF	EER	#	ID	t-DCF	EER
1	**T05**	0.0069	0.22	26	T57	0.2059	10.65
2	**T45**	0.0510	1.86	27	**T42**	0.2080	8.01
3	**T60**	0.0755	2.64	28	*B02*	0.2116	8.09
4	**T24**	0.0953	3.45	29	**T17**	0.2129	7.63
5	**T50**	0.1118	3.56	30	**T23**	0.2180	8.27
6	**T41**	0.1131	4.50	31	**T53**	0.2252	8.20
7	**T39**	0.1203	7.42	32	**T59**	0.2298	7.95
8	**T32**	0.1239	4.92	33	*B01*	0.2366	9.57
9	**T58**	0.1333	6.14	34	T52	0.2366	9.25
10	T04	0.1404	5.74	35	**T40**	0.2417	8.82
11	**T01**	0.1409	6.01	36	T55	0.2681	10.88
12	**T22**	0.1545	6.20	37	**T43**	0.2720	13.35
13	T02	0.1552	6.34	38	T31	0.2788	15.11
14	**T44**	0.1554	6.70	39	**T25**	0.3025	23.21
15	**T16**	0.1569	6.02	40	**T26**	0.3036	15.09
16	T08	0.1583	6.38	41	T47	0.3049	18.34
17	**T62**	0.1628	6.74	42	T46	0.3214	12.59
18	**T27**	0.1648	6.84	43	T21	0.3393	19.01
19	**T29**	0.1677	6.76	44	T61	0.3437	15.66
20	**T13**	0.1778	6.57	45	**T11**	0.3742	18.15
21	**T48**	0.1791	9.08	46	**T56**	0.3856	15.32
22	**T10**	0.1829	6.81	47	T12	0.4088	18.27
23	T54	0.1852	7.71	48	T14	0.4143	20.60
24	T38	0.1940	7.51	49	T20	1.0000	92.36
25	T33	0.1960	8.93	50	T30	1.0000	49.60

Leader board for ASVspoof2019 LA scenario

CONCLUSION

- This paper uses voice spoofing detection system based on one-class learning to enhance the robustness of automatic speaker verification systems.
- Using one-class classification instead of conventional binary classification makes the system stronger against unknown spoofing attacks.
- The bona fide speech lies in a tight boundary formed by angular margin using OC-softmax loss function, while the spoofed speech lies outside the boundary.
- The proposed system outperforms single systems using softmax, AM-softmax and many other methods and ranks 3rd among all participating systems in ASVspoof2019 challenge, even when the top systems use model fusion.