Guarding Against Deepfake Audio with One-Class Softmax

Chih-Yi Lin

Motivation

- Current anti-spoofing detection methods face **generalization** issues: Binary classification assumes similar distributions between training and evaluation (test) data for both "real" and "fake" audios
- Not true for the fake audios: The techniques used in the training data may never catch up with the ones in the real world
- One-Class Softmax (Zhang et al., 2021): Compacting the real audios and pushing away the fake audios in embedding space through **angular margin injection**

Research Questions:

1. Can OC-Softmax outperform Softmax across diverse **systems**?
2. Does OC-Softmax exhibit superior generalization capabilities on **real-world** samples (In-The-Wild dataset) compared to Softmax?

Method

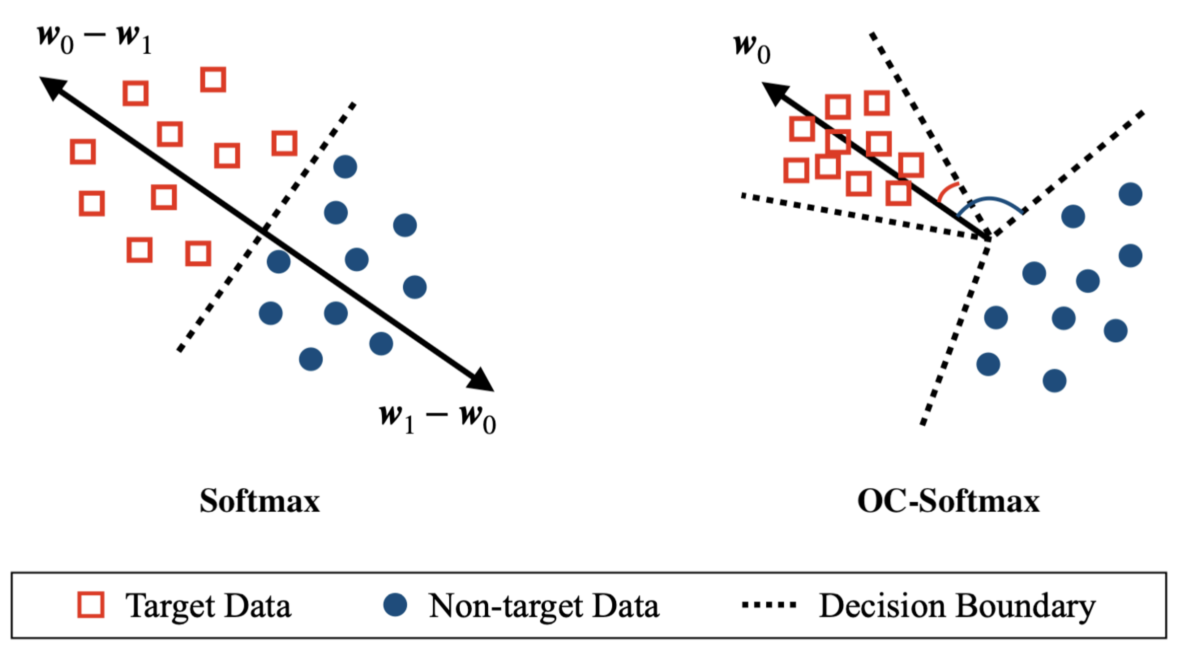

Figure 1: Softmax vs. OC-Softmax in 2D space[a]

$$L_{OCS} = \frac{1}{N}\sum_{i=1}^{N} log(1 + e^{\alpha(m_{y_i} - \hat{w}_0 \hat{x}_i)(-1)^{y_i}}) \quad (1)$$

- m: margin for cosine similarity to bound the angle between w_0 and x_i (θ_i), where
  - m_0: margin for the class "real", m_1: margin for the class "fake"
  - $m_0, m_1 \in [-1, 1], m_0 > m_1$
  - E.g., $y_i = 0$, m_0 force θ_i to be smaller than $\arccos m_0$
- w_0: optimization direction of the target class embeddings

[a] Figure from Zhang et al., 2021

Datasets

Datasets	Descr.	#Train (Real/Fake)	#Dev (Real/Fake)	#Eval
ASVspoof 2019-LA	Artifacts from TTS or VC	25,380 (2,580/22,800)	24,844 (2,548/22,296)	
ASVspoof 2021-DF[a]	Compressed audios	100,000	25,000	
In-The-Wild	Samples collected from the internet	–	–	31,779

Table 1: Summary of Train, Dev, Eval Datasets

[a] The class "real" is **oversampled** to balance two classes

Features and Experimental Details

Feature	Model	Train/Dev Set	Window Len.(ms)	Hop Len.(ms)	# Filters[a]	Time Frame	Embed Dim.[b]
LFCC	ResNet18	ASV19-LA	20	10	20 dim * 3	7.5s	256
LFCC	ResNet18	ASV21-DF					256
MFCC	ResNet18	ASV21-DF	400	160	128 dim * 3	15s	256
MFCC + Whisper[c]	MesoNet	ASV21-DF					1024

Table 2: Features

[a] LFCC/MFCC, delta, double delta
[b] Input for OC-Softmax
[c] Whisper: A Transformer-based encoder-decoder ASR system. Utilizing its encoder as feature extractor

- Hyper-param. of OC-Softmax: α=20, m_0=0.9 and m_1=0.2
- Models were trained for 20 epochs, selected with the lowest EER on Dev
  - Whisper: Trained with the first 5 epochs frozen, followed by 15 epochs unfrozen

Results

Feature	Model	Train/Dev Set	Loss	Dev EER	Eval-In-The-Wild
LFCC	ResNet18	ASV19-LA	Softmax	0.354%	**34.39%**
			OC-Softmax	0.279%	39.397%
LFCC	ResNet18 (100 epochs)	ASV19-LA	Softmax	0.274%	34.30%
			OC-Softmax	0.201%	**28.166%**
LFCC	ResNet18	ASV21-DF	Softmax	1.843%	60.176%
			OC-Softmax	2.196%	**47.157%**
MFCC	ResNet18	ASV21-DF	Softmax	1.464%	47.655%
			OC-Softmax	1.956%	**42.909%**
MFCC + Whisper	MesoNet	ASV21-DF	Softmax	0.444%	37.00%
			OC-Softmax	0.71%	**29.86%**

Table 3: Experiment results

- **MFCC-Whisper-MesoNet-ASV21-OC-Softmax** achieves superior performance with equivalent training epochs
- OC-Softmax outperforms Softmax after 100 epochs in the first configuration, suggesting potential performance gains with **more training epochs** across all configurations
- Models trained on ASV21-DF do **not** consistently outperform those trained on ASV19-LA: Higher Dev EER, may need more training epochs
- Significant EER disparity observed between the Dev and Eval sets

Visualization of Learned Embeddings

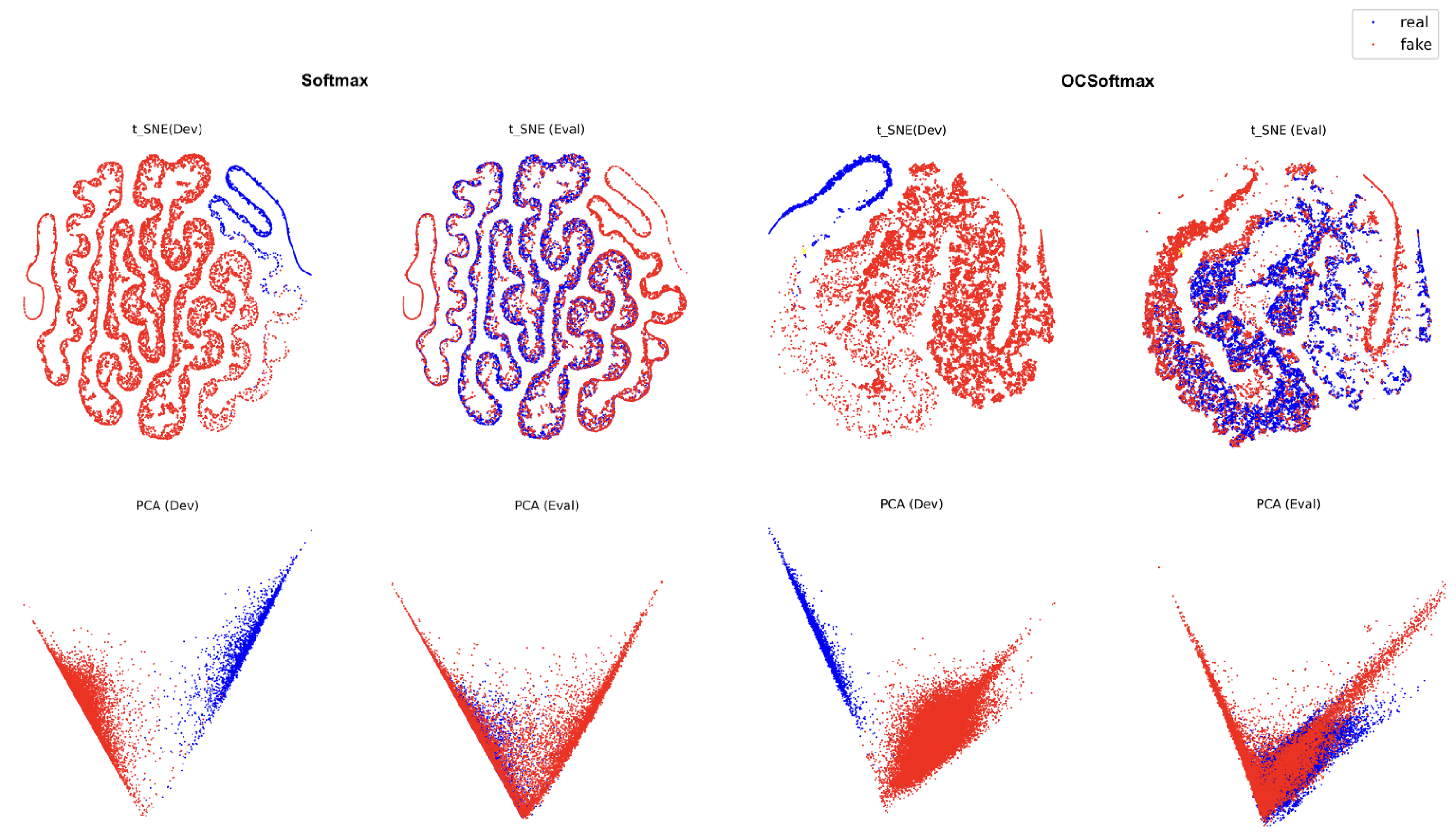

Figure 2: Distributions of the embeddings learned by LFCC-ResNet18-ASV19-100ep based on gold labels

- Dev vs. Eval: Different distributions for **both classes**
- Softmax vs. OC-Softmax: While OC-Softmax produces a more **compact** cluster for real audios in both Dev and Eval sets, it presents entanglement in the latter, posing a greater challenge for models

Conclusions

- OC-Softmax **outperforms** Softmax for all settings on real-world samples, except LFCC-ResNet-ASV19-20ep
- Utilizing **Whisper's encoder** as a feature extractor alongside conventional features demonstrates significant potential
- Future works: One-class classification methods, Continuous learning

References

[1] Y. Zhang, F. Jiang, and Z. Duan, "One-class learning towards synthetic voice spoofing detection," *IEEE Signal Processing Letters*, vol. 28, pp. 937–941, 2021.

[2] P. Kawa, M. Plata, M. Czuba, P. Szymański, and P. Syga, "Improved deepfake detection using whisper features," 2023.