Generalizing Voice Presentation Attack Detection to Unseen Synthetic Attacks

You (Neil) Zhang
University of Rochester

Feb 06, 2023
Outline

One-Class Learning Towards Synthetic Voice Spoofing Detection (SPL’21)

SAMO: Speaker Attractor Multi-Center One-Class Learning for Voice Anti-Spoofing (ICASSP’23)
A voice anti-spoofing system is desired to distinguish presentation attacks from bona fide speech.
Research question

Motivation:

- The fast development of speech synthesis are posing increasingly more threat.
- The distribution mismatch between the training set and test set for the spoofing attacks class.

How can the anti-spoofing system defend against unseen spoofing attacks?

Generalization ability!
One-Class Learning Towards Synthetic Voice Spoofing Detection

You Zhang, Fei Jiang, Zhiyao Duan

University of Rochester, NY, USA

Definition of one-class

- “One-class classification (OCC) algorithms aim to build classification models when the negative class is either absent, poorly sampled or not well defined.”

- “In one-class classification, one of the classes (referred to as the positive class or target class) is well characterized by instances in the training data. For the other class (nontarget), it has either no instances at all, very few of them, or they do not form a statistically-representative sample of the negative concept.”

Illustration of comparison

One-class learning

- Compact the bona fide speech representation
- Isolate the spoofing attacks

Training: OC-Softmax loss (Proposed)

\[
\mathcal{L}_{OCS} = \frac{1}{N} \sum_{i=1}^{N} \log \left( 1 + e^{\alpha (m_{y_i} - \hat{w}_0 \hat{x}_i)(-1)^{y_i}} \right).
\]

Inference: cosine similarity

\[
S_{OCS} = \hat{w}_0 \hat{x}_i.
\]
Comparing OC-Softmax with binary classification

(a) Original Softmax  (b) AM-Softmax  (c) OC-Softmax (Proposed)
Comparing OC-Softmax with binary classification

Softmax:

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

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

AM-Softmax:

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

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

OC-Softmax:

$$\mathcal{L}_{OCS} = \frac{1}{N} \sum_{i=1}^{N} \log \left(1 + e^{\alpha (m_{y_i} - \hat{\mathbf{w}}_0 \hat{\mathbf{x}}_i)(-1)^{y_i}}\right).$$
## Dataset

ASVspoof 2019 Logical Access (TTS + VC)

- Bona fide speech (VCTK dataset)
- 6 known attacks (appear in the training set)
- 13 unknown attacks (only appear in the evaluation set)

<table>
<thead>
<tr>
<th></th>
<th>Bona fide</th>
<th>Spoofed</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td># utterance</td>
<td># utterance</td>
</tr>
<tr>
<td>Training</td>
<td>2,580</td>
<td>22,800</td>
</tr>
<tr>
<td>Development</td>
<td>2,548</td>
<td>22,296</td>
</tr>
<tr>
<td>Evaluation</td>
<td>7,533</td>
<td>63,882</td>
</tr>
</tbody>
</table>
Evaluation of OC-Softmax

Results on the development and evaluation sets of ASVspoof 2019 LA using different losses

<table>
<thead>
<tr>
<th>Loss</th>
<th>Dev Set</th>
<th>Eval Set</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>EER (%)</td>
<td>min t-DCF</td>
</tr>
<tr>
<td>Softmax</td>
<td>0.35</td>
<td>0.010</td>
</tr>
<tr>
<td>AM-Softmax</td>
<td>0.43</td>
<td>0.013</td>
</tr>
<tr>
<td>OC-Softmax</td>
<td>0.20</td>
<td>0.006</td>
</tr>
</tbody>
</table>

- OC-Softmax performs the best on unseen attacks.
- Achieved the state-of-the-art single-system performance.

Feature Embedding Visualization (red: bona fide, green: A17 attack, blue: spoofing attacks)
Comparison with single systems

Achieved the state-of-the-art performance

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

Euclidean distance-based one-class loss (isolate loss, single-center loss)

Cosine similarity-based one-class loss (OC-Softmax, angular isolate loss)

Takeaways

● One-class learning aims to **compact the target** class representation in the embedding space, set a tight classification boundary around it, and push away non-target.

● The proposed OC-Softmax could improve the **generalization ability** of anti-spoofing system against **unseen spoofing attacks**.
SAMO: Speaker Attractor Multi-Center One-Class Learning for Voice Anti-Spoofing

Siwen Ding¹, You Zhang², Zhiyao Duan²

¹Columbia University, NY, USA
²University of Rochester, NY, USA

Research question

Motivation:
● In our previous work, we compact the embedding space of the bona fide speech into one cluster.
● However, due to the variety of timbre and speaking traits of different speakers, the bona fide speech of different speakers naturally forms multiple clusters in the embedding space.

How to improve the generalization ability while maintaining the variation of bona fide speech?
Related work

Deep Multi-sphere Support Vector (SDM’20)

If data is naturally multi-cluster, merging them into one cluster could be harmful for detecting anomaly.

(Figure 2 in Ghafoori et al.)

Speaker attractor multi-center one-class learning

Model speaker diversity while maintaining the generalization ability brought by one-class learning

- Discriminate bona fide vs. spoofing attacks
- Cluster bona fide speech according to speakers
Speaker attractors

Define: a speaker-specific anchor in the embedding space

Compute: average the embeddings of each speaker’s bona fide speech

- **Training**: attract bona fide speech embeddings of the same speaker
- **Inference**: cosine similarity between test utterance and enrolled utterance or attractors of training speakers

\[
S_{SAMO} = \begin{cases} 
\hat{w}_s \hat{x}_i & \text{if } s_i \text{ is enrolled} \\
\max_s (\hat{w}_s \hat{x}_i), s \in S_{train} & \text{otherwise}
\end{cases}
\]
Loss function for multi-center one-class learning

- Compact the bona fide speech representation belonging to the same speaker
- Push away the spoofing attacks from all speaker attractors

\[ \mathcal{L}_{SAMO} = \frac{1}{N} \sum_{i=1}^{N} \log \left( 1 + e^{\alpha (m y_i - d_i)} (-1)^{y_i} \right), \]

where \( d_i \) is calculated by

\[ d_i = \begin{cases} \hat{w}_{s_i} \hat{x}_i & \text{if } y_i = 0 \\ \max_s (\hat{w}_s \hat{x}_i), s \in S_{\text{train}} & \text{if } y_i = 1 \end{cases}. \]
SAMO Training algorithm

- Compact the bona fide utterances spoken by the same speaker
- Push away spoofing utterances from all speaker attractors

**Algorithm 1: SAMO Training Algorithm**

Require: $T$: Total number of epochs  
          $M$: speaker attractor update interval (# epochs)

1. Initialize network $F$ with random weights
2. Initialize speaker attractors $w_s$ as one-hot vectors
3. for $i \leftarrow 1$ to $T$ do
   4. if $i \mod M = 0$ then
      5. Update $w_s$ as the average bona fide embedding for each speaker $s \in S_{train}$
   6. end if
   7. Update $F$ by $L_{SAMO}$ with mini-batches $\triangleright$ Eq. (3)
4. end for
5. return Optimized network $F$ and speaker attractors $w_s$
Dataset

ASVspoof2019 LA target-only portion
- Same train/dev/eval splits
- Keep only the target speakers in dev/eval sets

<table>
<thead>
<tr>
<th>Partition</th>
<th># Speakers</th>
<th># Enrollment Utts</th>
<th>Bona Fide # Utts</th>
<th>Spoofing Attacks # Utts</th>
<th>Attack Types</th>
</tr>
</thead>
<tbody>
<tr>
<td>Train</td>
<td>20</td>
<td>-</td>
<td>2580</td>
<td>22800</td>
<td>A01~A06</td>
</tr>
<tr>
<td>Dev</td>
<td>10</td>
<td>142</td>
<td>1484</td>
<td>22296</td>
<td>A01~A06</td>
</tr>
<tr>
<td>Eval</td>
<td>48</td>
<td>696</td>
<td>5370</td>
<td>63882</td>
<td>A07~A19</td>
</tr>
</tbody>
</table>
Comparison with state-of-the-art methods

Table 2. Comparison of our proposed SAMO with Softmax and OC-Softmax on the target-only portion of the ASVspoof2019 LA evaluation set. All the systems use AASIST [9] backbone. The average (best) results across 3 trials with random training seeds are shown.

<table>
<thead>
<tr>
<th>Method</th>
<th>EER(%)</th>
<th>min t-DCF</th>
</tr>
</thead>
<tbody>
<tr>
<td>Softmax</td>
<td>1.74 (1.25)</td>
<td>0.0583 (0.0425)</td>
</tr>
<tr>
<td>OC-Softmax</td>
<td>1.25 (1.17)</td>
<td>0.0415 (0.0393)</td>
</tr>
<tr>
<td>SAMO (test w/o enrollment)</td>
<td>1.09 (0.91)</td>
<td>0.0363 (0.0306)</td>
</tr>
<tr>
<td>SAMO (test w/ enrollment)</td>
<td>1.08 (0.88)</td>
<td>0.0356 (0.0291)</td>
</tr>
</tbody>
</table>

SAMO further improves the performance, indicating the advantage brought by the multi-center modeling of bona fide speech.
Embedding visualization

2D t-SNE visualization of SAMO feature embeddings of bona fide and spoofed speech of three speakers

- Bona fide utterances are grouped in a small region.
- Utterances of the three speakers are generally clustered according to speaker identity.
Ablation studies

Effects of leveraging bona fide speech data for speaker attractors

Effects of the speaker attractor update interval $M$ (# epochs)

Table 3. Ablation experiments for SAMO. Results of test scenarios without and with enrollment data are both presented.

<table>
<thead>
<tr>
<th>Setup</th>
<th>Configuration</th>
<th>Test w/o enroll (w/ enroll)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>EER(%)</td>
</tr>
<tr>
<td>1</td>
<td>SAMO</td>
<td><strong>1.09 (1.08)</strong></td>
</tr>
<tr>
<td>2</td>
<td>one-hot and fixed attractors</td>
<td>47.93 (49.50)</td>
</tr>
<tr>
<td>3</td>
<td>w/o speaker attractor update</td>
<td>1.54 (1.55)</td>
</tr>
<tr>
<td>4</td>
<td>update every epoch ($M=1$)</td>
<td>1.33 (1.33)</td>
</tr>
<tr>
<td>5</td>
<td>update every 10 epochs ($M=10$)</td>
<td>2.36 (2.77)</td>
</tr>
</tbody>
</table>
Future work

With a larger variety of speakers in the training set, the benefit of SAMO could be demonstrated even more since the speaker attractors will better represent the bona fide embedding space.

Extend the SAMO idea to model other speech attributes, such as device and codec variations.
Other works on anti-spoofing

- **Channel Robustness**

- **Joint Optimization with ASV**
Take-aways

One-class learning could improve the generalization ability of anti-spoofing system against unseen spoofing attacks.

It aims to compact the bona fide speech representation in the embedding space, and push away spoofing attacks by a certain margin.

Speaker attacker multi-center one-class learning could further improve the generalization ability while clustering bona fide speech around a number of speaker attractors and pushing away spoofing attacks from all the attractors.

Thank you!  Questions?