Text-Independent Speaker Verification with Adversarial Learning on Short Utterances

Authors: Kai Liu, Huan Zhou

Presented by: Kai Liu, Huan Zhou
Content

1. Introduction
2. Related Works
3. Proposed Approach
4. Experiments and Results
5. Conclusion and Future work
Introduction

Speaker Verification System
• i-vector
• X-vector
• D-vector
• G-vector
... 

Short-utterance Speaker Verification
• Performance decline dramatically
  e.g. (NIST-SRE 2010) i-vector/PLDA EER : 
    2.48%(full) → 24.78%(5 seconds)
Introduction

Improvement

• feature extraction techniques, intermediate parameter estimation, speaker model generation, score normalization
• teacher-student framework & scoring scheme calibration
• duration robust speaker embeddings
  • NN architectures: Inception Net, Inception-ResNet, ResCNN, GANs, ...
  • Losses: triplet loss, am-softmax, ...
Related Works

Figure 3: Training of the generator network $G$ and its application in the testing stage.

cite: Ivector transformation using conditional generative adversarial networks for short utterance speaker verification

Table 1: The speaker verification results in terms of EER (%) on all the three conditions of the SRE08 “short2-10sec” male trial list.

<table>
<thead>
<tr>
<th>System</th>
<th>Cond. 6</th>
<th>Cond. 7</th>
<th>Cond. 8</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td>a) Baseline</td>
<td>7.28</td>
<td>6.15</td>
<td>6.06</td>
<td>6.50</td>
</tr>
<tr>
<td>b) Single G</td>
<td>10.04</td>
<td>8.85</td>
<td>8.33</td>
<td>9.07</td>
</tr>
<tr>
<td>c) a + b</td>
<td>7.28</td>
<td>5.77</td>
<td>6.06</td>
<td>6.37</td>
</tr>
<tr>
<td>d) D-WCGAN</td>
<td>9.45</td>
<td>8.08</td>
<td>8.33</td>
<td>8.62</td>
</tr>
<tr>
<td>e) a + d</td>
<td>6.89</td>
<td>5.77</td>
<td>5.30</td>
<td>5.99</td>
</tr>
</tbody>
</table>

Table 2: The speaker verification results in terms of EER (%) on all the three conditions of the SRE08 “10sec-10sec” male trial list.

<table>
<thead>
<tr>
<th>System</th>
<th>Cond. 6</th>
<th>Cond. 7</th>
<th>Cond. 8</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td>a) Baseline</td>
<td>11.97</td>
<td>10.32</td>
<td>9.60</td>
<td>10.63</td>
</tr>
<tr>
<td>b) Single G</td>
<td>15.32</td>
<td>13.89</td>
<td>12.00</td>
<td>13.77</td>
</tr>
<tr>
<td>c) a + b</td>
<td>11.16</td>
<td>10.71</td>
<td>9.60</td>
<td>10.49</td>
</tr>
<tr>
<td>d) D-WCGAN</td>
<td>15.42</td>
<td>13.89</td>
<td>13.60</td>
<td>14.30</td>
</tr>
<tr>
<td>e) a + d</td>
<td>10.75</td>
<td>8.73</td>
<td>8.80</td>
<td>9.43</td>
</tr>
</tbody>
</table>
Proposed Approach

Fig. 1.1. Framework of our proposed system
Proposed Approach

Fig. 1.2. Generator network structure
Proposed Approach

Discriminator-Related Loss Functions

• conditional wasserstein distance loss

\[
\min_{G_f} \max_{D_w} L_{cw}(D_w, G_f) = \\
E_y[D_w(y; x)] + E_x[D_w(G_f(x); x)]
\]

• Fréchet Inception Distance (fid) loss

\[
L_{fid} = |\mu_y - \mu_g|^2 + tr\left(\frac{C_y + C_g - 2(C_yC_g)^{1/2}}{2}\right)
\]
Proposed Approach

Generator-Related Loss Functions

- softmax loss

\[
L_{\text{class}} = \frac{1}{N} \sum_{i=1}^{N} \log \frac{e^{W_{z_i}^T g_i + b_{z_i}}}{\sum_{j=1}^{c} e^{W_{j}^T g_i + b_j}}
\]

- triplet loss

\[
L_{\text{triplet}} = \sum_{\gamma \in \Gamma} \max \left( \| g_a - g_p \|_2^2 - \| g_a - g_n \|_2^2 + \Psi, 0 \right)
\]

- center loss

\[
L_{\text{center}} = \frac{1}{2} \sum_{i=1}^{m} \| x_i - c_{y_i} \|_2^2
\]

- cosine loss

\[
L_{\text{cos}} = 1 - \bar{g}^* \bar{y}
\]
Proposed Approach

Total Loss Functions

• Discriminator

\[ L_W = \frac{L_w}{L_{cw}} + \lambda L_{fid} \]

• Generator

\[ L_G = \frac{L_w}{L_{cw}} + \alpha L_{class} + \beta L_{cos} + L_{center} + \epsilon L_{triplet} \]
Dataset

Train Set
- subset of voxceleb2
- 1,057 speakers
- 164,716 utterances (randomly cut to 2 seconds vs. original wav)

Test Set
- subset of voxceleb1
- 40 speakers
- 13,265 utterance pairs (randomly cut to 2 seconds and 1 second)
## Experiments

### Table 1. System descriptions

<table>
<thead>
<tr>
<th>system</th>
<th>$L_c$</th>
<th>$L_{cos}$</th>
<th>$L_t$</th>
<th>$L_{class}$</th>
<th>$L_{cw}$</th>
<th>$L_{fid}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>v1</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>v2</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>v3</td>
<td></td>
<td>✓</td>
<td>✓ a</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>v4</td>
<td></td>
<td></td>
<td>✓ a</td>
<td>✓</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>v6</td>
<td>✓</td>
<td></td>
<td>✓ b</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>v5</td>
<td></td>
<td></td>
<td>✓ a</td>
<td>✓</td>
<td></td>
<td>✓</td>
</tr>
<tr>
<td>v7</td>
<td>✓</td>
<td></td>
<td>✓ b</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>v8</td>
<td>✓</td>
<td></td>
<td>✓ b</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
</tbody>
</table>

$L_t$: a means that inputs are sampled from both $y$ and $g$ and $b$ means from $g$ only
Experiments

Fig. 2. DET performances for different systems

$ps$ : we compute EER by compare embedding cosine distance
Experiments

- FID loss has positive effect (v1 vs. v2);
- Conditional WGAN outperforms WGAN (v3 vs. v4);
- Triplet loss is preferred (v7 vs. v2);
- Triplet a greatly outperforms triplet b (v3 vs. v8);
- softmax has positive effect (v3 vs. v5);
- Center loss has negative effect (v6 vs. v7);
- Cosine loss has significant positive effect (v6 vs. v8).

<table>
<thead>
<tr>
<th>system</th>
<th>$L_c$</th>
<th>$L_{cos}$</th>
<th>$L_t$</th>
<th>$L_{class}$</th>
<th>$L_{cw}$</th>
<th>$L_{fid}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>v1</td>
<td>√</td>
<td></td>
<td>√</td>
<td></td>
<td>√</td>
<td>√</td>
</tr>
<tr>
<td>v2</td>
<td>√</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>v3</td>
<td>√</td>
<td>a</td>
<td>√</td>
<td>√</td>
<td></td>
<td></td>
</tr>
<tr>
<td>v4</td>
<td></td>
<td>a</td>
<td>√</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>v6</td>
<td>√</td>
<td></td>
<td>√</td>
<td>b</td>
<td>√</td>
<td></td>
</tr>
<tr>
<td>v5</td>
<td></td>
<td></td>
<td>√</td>
<td>b</td>
<td></td>
<td></td>
</tr>
<tr>
<td>v7</td>
<td>√</td>
<td></td>
<td>√</td>
<td>b</td>
<td>√</td>
<td></td>
</tr>
<tr>
<td>v8</td>
<td></td>
<td>b</td>
<td></td>
<td></td>
<td>√</td>
<td></td>
</tr>
</tbody>
</table>
Experiments

Table 2. Comparison with the baseline system

<table>
<thead>
<tr>
<th>system</th>
<th>2s-2s</th>
<th>1s-1s</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>EER(%)</td>
<td>minDCF</td>
</tr>
<tr>
<td>G-vector</td>
<td>7.557</td>
<td>0.8170</td>
</tr>
<tr>
<td>ours</td>
<td>7.237</td>
<td>0.7578</td>
</tr>
<tr>
<td>fusion</td>
<td>7.168</td>
<td>0.7734</td>
</tr>
</tbody>
</table>
Conclusion

• proposed enhanced embedding for short-utterance speaker verification with Wasserstein Conditional GAN

• validated the effectiveness of a bunch of loss criteria on the GAN training
Future work

- better GAN structure
- more data
- how to describe distribution similarity in a better way
- GAN inside embedding extraction network
- more training tricks
Thank you.

把数字世界带入每个人、每个家庭、每个组织，构建万物互联的智能世界。
Bring digital to every person, home, and organization for a fully connected, intelligent world.

Copyright©2018 Huawei Technologies Co., Ltd. All Rights Reserved.

The information in this document may contain predictive statements including, without limitation, statements regarding the future financial and operating results, future product portfolio, new technology, etc. There are a number of factors that could cause actual results and developments to differ materially from those expressed or implied in the predictive statements. Therefore, such information is provided for reference purpose only and constitutes neither an offer nor an acceptance. Huawei may change the information at any time without notice.