MULTI-RESOLUTION MULTI-HEAD ATTENTION IN DEEP SPEAKER EMBEDDING

Zhiming Wang, Kaisheng Yao, Xiaolong Li, Shuo Fang
Ant Financial Services Group, Hangzhou, China
Outlines

- Deep Speaker Embedding Framework
- Pooling: an Overview
- the Proposed Pooling Methods
- Experiments and Results
- Conclusions
Deep Speaker Embedding Framework

- Deep Speaker Embedding (x-vector)
  - DNN: cepstral acoustic features → a sequence of encoded vector
  - A pooling layer: a segment-level representation (embedding); today’s topic
  - A classifier (softmax or fully connected network): project to speaker ids
  - Outputs at reciprocal a certain layer: embedding feature
Outline

- Deep Speaker Embedding Framework
- Pooling: an Overview
- the Proposed Pooling Methods
- Experiments and Results
- Conclusions
Pooling(1): an Overview

✓ Statistics pooling
✓ Self attentive pooling
✓ Multi-head attentive pooling (for increased discriminative information)
✓ Multi-resolution multi-head attentive pooling (for encouraging diversity from multiple heads)

Pooling is Essential

Zhiming Wang, Kaisheng Yao, Xiaolong Li, and Shuo Fang
Pooling(2): an Overview

- **Statistics pooling**
  - mean + std: capture *overall information* and *dynamical variability*

- **Self attentive pooling**
  - compute importance of each frame

- **Attentive statistics pooling**
  - mean + std: using self attentive weight $\alpha_t$, NOT in average with $\alpha_t = 1/N$

---

$\alpha_t = \frac{e^{s_d(h_t)}}{\sum_{j=1}^{N} e^{s_d(h_j)}}$

$e = \sum_{t=1}^{N} \alpha_t h_t$

$s_i^{(i)}(x) = V_i^T f(W_i x + g_i) + b_i$,

where $f(\cdot)$ is a non-linear activation function, $W_i \in \mathbb{R}^{d \times l}$, $g_i \in \mathbb{R}^l$, $V_i \in \mathbb{R}^d$ and $b_i \in \mathbb{R}$ are parameters to learn.
Pooling(3): an Overview

- Multi-head attentive pooling ([1])
  - split the encoded frame into non-overlapping homogeneous sub-vectors
  - apply attentive pooling on frame sequence of sub-vectors
  - may ignore possible correlations among different sub-vectors, especially for $h_t$ with small dimensional size

$$h_t = [h_t^{(1)}, h_t^{(2)}, \ldots, h_t^{(K)}]$$

$$\alpha_t^{(i)} = \frac{e^{(s_t^{(i)}(x))}}{\sum_{j=1}^{N} e^{(s_t^{(j)}(x))}}$$

$$e^{(i)} = \sum_{t=1}^{N} \alpha_t^{(i)} h_t^{(i)}$$

$$e = [e^{(1)}, e^{(2)}, \ldots, e^{(K)}]$$

$$s_t^{(i)}(x) = V_t^T x,$$

$$V_t \in \mathbb{R}^{i}$$

Zhiming Wang, Kaisheng Yao, Xiaolong Li, and Shuo Fang
Outline

- the Framework of Deep Speaker Embedding
- Pooling: an Overview
- The Proposed Pooling Methods
- Experiments and Results
- Conclusions
The Proposed Pooling Methods (1)

- Global multi-head attentive pooling
  - Apply K-head attention over the entire encoded sequence

\[
\begin{align*}
s^{(i)}_t(x) &= V^T_i f(W_s x + g_i) + b_i \\
\alpha^{(i)}_t &= \frac{e^{(s^{(i)}_t(h_t))}}{\sum_{j=1}^{N} e^{(s^{(i)}_t(h_j))}} \\
e^{(i)} &= \sum_{t=1}^{N} \alpha^{(i)}_t h_t \\
e &= (e^{(1)}, e^{(2)}, \ldots, e^{(K)})
\end{align*}
\]
The Proposed Pooling Methods (2)

- Multi-resolution multi-head attentive pooling (for diversity)
  - Increasing $T$ will make $\alpha_T(z_i)$ less sharper, thus with lower resolution
  - As $T \rightarrow \infty$, $\alpha_T(z_i) = 1/N$, average pooling: bridge between attentive and statistical pooling

![Graphs showing the function curves of $\alpha_T(z_i)$ as $T$ varies.]

Fig. 1. The function curves of $\alpha_T(z_i) = \frac{e^{(z_i/T)}}{\sum_j e^{(z_j/T)}}$ as $T$ varies.

\[
\begin{align*}
    s^{(i)}(x) &= V_T^T f(W_i x + g_i) + b_i \\
    \alpha^{(i)}_t &= \frac{e^{(s^{(i)}_d(h_t)/T_i)}}{\sum_{j=1}^N e^{(s^{(i)}_d(n_j)/T_i)}} \quad T_i = 1: \text{global multi-head attention} \\
    e^{(t)} &= \sum_{t=1}^N \alpha^{(i)}_t h_t \\
    e &= (e^{(1)}, e^{(2)}, \ldots, e^{(K)})
\end{align*}
\]

Zhiming Wang, Kaisheng Yao, Xiaolong Li, and Shuo Fang
The Proposed Pooling Methods(3)

- Compared with Povey’s method[6], using a penalty regularization in training objective to encourage diversity
  - attentive weights from different heads are orthogonal: a stronger requirement
  - not guarantee the approximate orthogonality of attentive weights during prediction
  - our method: achieve diversity of extracted speech characteristics through the learned multi-resolution attentive model

\[
A = \text{softmax}(g(H^T W_1) W_2) 
\]

(1)

\[
E = HA 
\]

(2)

\[
P = \|(A^T A - I)\|_F^2 
\]

(3)

Zhiming Wang, Kaisheng Yao, Xiaolong Li, and Shuo Fang
Outline

- the Framework of Deep Speaker Embedding
- Pooling: an Overview
- the Proposed Pooling Methods
- Experiments and Results
- Conclusions
Experiments and Results(1)

1. dataset: VoxCeleb1
2. input features: 40-dimensional log-Mel filter banks
3. network architecture: 34-layer convolution ResNet
4. loss function: additive cosine margin softmax
5. optimizer: RAdam

\[
\mathcal{L}_{\text{CosAMS}} = -\frac{1}{B} \sum_{u=1}^{B} \log \frac{e^{\eta (\cos(\theta_{x,u,w_y,u}) - m)}}{Z_{x,u}},
\]

\[
Z_{x,u} = e^{\eta (\cos(\theta_{x,u,w_y,u}) - m)} + \sum_{j \neq y_u} e^{\eta \cos(\theta_{x,u,w_j})},
\]

<table>
<thead>
<tr>
<th>Layer</th>
<th>Configuration</th>
</tr>
</thead>
<tbody>
<tr>
<td>Conv1</td>
<td>(3 × 3, 64), stride (1 × 1)</td>
</tr>
<tr>
<td>Res1</td>
<td>[(3 × 3, 64), 2] × 3</td>
</tr>
<tr>
<td>Res2</td>
<td>[(3 × 3, 128), 2] × 4</td>
</tr>
<tr>
<td>Res3</td>
<td>[(3 × 3, 256), 2] × 6</td>
</tr>
<tr>
<td>Res4</td>
<td>[(3 × 3, 512), 2] × 3</td>
</tr>
<tr>
<td>Conv2</td>
<td>(3 × 3, 512), stride (1 × 3)</td>
</tr>
<tr>
<td>Pooling</td>
<td>pooling as represented in Sec.2</td>
</tr>
<tr>
<td>Linear1</td>
<td>output-dimension-of-pooling × 512</td>
</tr>
<tr>
<td>Linear2</td>
<td>512 × 512</td>
</tr>
<tr>
<td>Classifier</td>
<td>512 × C, C = 1211</td>
</tr>
<tr>
<td>Front-end</td>
<td>Approaches</td>
</tr>
<tr>
<td>--------------</td>
<td>---------------------------------------------</td>
</tr>
<tr>
<td>i-vector</td>
<td>PLDA</td>
</tr>
<tr>
<td>VGG-M</td>
<td>adaptive average pooling</td>
</tr>
<tr>
<td>VGG-M</td>
<td>adaptive average pooling</td>
</tr>
<tr>
<td>W. Cai et al. [20]</td>
<td></td>
</tr>
<tr>
<td>ResNet-34</td>
<td>temporal average pooling</td>
</tr>
<tr>
<td>ResNet-34</td>
<td>self-attentive pooling</td>
</tr>
<tr>
<td>ResNet-34</td>
<td>learnable dictionary encoding(LDE)</td>
</tr>
<tr>
<td>Our baselines</td>
<td></td>
</tr>
<tr>
<td>ResNet-34</td>
<td>attentive statistics pooling</td>
</tr>
<tr>
<td>ResNet-34</td>
<td>multi-head attention($K = 4$), with Eq.(1)</td>
</tr>
<tr>
<td>ResNet-34</td>
<td>multi-head attention($K = 4$), with Eq.(5)</td>
</tr>
<tr>
<td>Our proposals</td>
<td></td>
</tr>
<tr>
<td>ResNet-34</td>
<td>global multi-head attention($K = 4$)</td>
</tr>
<tr>
<td>ResNet-34</td>
<td>multi-resolution multi-head attention($K = 4$)</td>
</tr>
<tr>
<td>ResNet-34</td>
<td>global multi-head attention($K = 5$)</td>
</tr>
<tr>
<td>ResNet-34</td>
<td>multi-resolution multi-head attention($K = 5$)</td>
</tr>
<tr>
<td>ResNet-34</td>
<td>global multi-head attention($K = 6$)</td>
</tr>
<tr>
<td>ResNet-34</td>
<td>multi-resolution multi-head attention($K = 6$)</td>
</tr>
</tbody>
</table>

Table 3. Results for verification on the test set of VoxCeleb1, all using the development set of VoxCeleb1 for training. “ASofmax” represents angular softmax loss.
Experiments and Results (3)

- multi-resolution multi-head attention
- capture different views of speech characteristics
- less uncertain (lower entropy) in achieving discriminative information, representative of being more regularized

Fig. 2. The self-attentive weights from single-head vs. multi-resolution multi-head attention along the temporal axis, given the same test speech; much more attention is paid to the frames of higher weight scores. $E(= - \sum \alpha_i \log(\alpha_i))$ is the entropy.
Outline

- Deep Speaker Embedding Framework
- Pooling: an Overview
- The Proposed Pooling Methods
- Experiments and Results
- Conclusions
Conclusions

- proposed global and multi-resolution multi-head attention
- consistent improvement on top of that achieved with increased number of attention heads

Why
- analyzing speech segments as a whole
- multiple views from different attention heads with various resolutions
- improved certainty on each head
Thank you!

Q & A