Multi-View Self-Attention based Transformer for Speaker Recognition

ICASSP 2022

Rui Wang\(^1\)*, Junyi Ao\(^2,3\)*, Long Zhou\(^4\), Shujie Liu\(^4\), Zhihua Wei\(^1\), Tom Ko\(^2\), Qing Li\(^3\), Yu Zhang\(^2\)

\(^1\)Department of Computer Science and Technology, Tongji University
\(^2\)Department of Computer Science and Engineering, Southern University of Science and Technology
\(^3\)Department of Computing, The Hong Kong Polytechnic University
\(^4\)Microsoft Research Asia

*Equal contribution. Work done during internship at Microsoft Research Asia.
Speaker Recognition

Speaker recognition aims to identify the voice of the specific targets.

- Convolutional architectures remain dominant, such as residual network (ResNet) and time delay neural network (TDNN).
- The applications of Transformer to speaker recognition are limited, e.g., combining CNN-like architectures with self-attention by either replacing utterance-level pooling layers or frame-level convolutional blocks.
Motivation & Challenges

Applying Transformer to speaker tasks has two challenges:

• Transformer is hard to be scaled efficiently since acoustic features sequences are much longer than text sentences.

• Transformer is deficient in some of the inductive biases inherent to CNNs, such as translation equivalence and locality.
Multi-View Self-Attention

To enhance the Transformer’s capabilities of capturing global dependencies while modeling the locality, a multi-view self-attention is proposed to employ windows with different sizes surrounding each token in a head-wise manner.
Multi-View Self-Attention

Specifically, given a fixed window size $w$, each token attends to $\frac{1}{2}w$ tokens on both sides.

The sliding window for the $i$-th head at the $l$-th layer to explicitly model different ranges of receptive fields by setting them as

$$w^l_i = \begin{cases} 2^i + 1, & \text{if } i \geq 1 \\ 1, & \text{if } i = 0 \end{cases}$$
Transformer Variants

We study five Transformer variants.

a. **First Decoder Token.** Multi-layer multi-head attentive pooling.

b. **Last Decoder Token.** Input-related pooling.

c. **Average Encoder Token.** Temporal Average pooling.

d. **First Encoder Token.** Use of a single token.

e. **Pooling Encoder Tokens.** Like X-vector.
Experimental Setup

• Datasets
  • Speaker Identification:
    • VoxCeleb1 development set: over 1,000,000 utterances from 1,251 celebrities
    • VoxCeleb1 test set: over 8,251 utterances from 1,251 celebrities
  • Speaker Verification:
    • VoxCeleb1/VoxCeleb2 development set: over 100,000/1,000,000 utterances from 1,211/5,994 celebrities
    • VoxCeleb1 test set: 37,720 pairs of trials and over 4,715 utterances from 40 celebrities

• Acoustic Features: 80-d mel-filter banks with the 64ms windows and 16ms shifts
• Identification Metrics: Top-1 accuracy (ACC)
• Verification Metrics: Equal error rate (EER)
Experiment Results

We compare these five variants with or without multi-view self-attention (MV) and report the performance on VoxCeleb1 test set.

- Multi-view self-attention achieves improvement in most settings.
- Multi-view self-attention can generate various token representation.

| Variant | Architecture Details | ACC (%) ↑ | EER (%) ↓ | EER (%) ↓ |
|---------|----------------------|-----------|-----------|-----------|
|         |                      | Vox1 + MV | Vox1 +MV  | Vox2 +MV  |
| a       | Attentive Pooling    | 94.33     | 5.33      | 2.72      |
|         |                      | 94.36     | 5.45      | 2.56      |
| b       | Input-related Pooling| 93.61     | 5.89      | 2.92      |
|         |                      | 94.09     | 5.40      | 2.68      |
| c       | Average Pooling      | 92.96     | 6.33      | 3.60      |
|         |                      | 91.81     | 6.13      | 3.23      |
| d       | Single Token         | 92.29     | 5.96      | 3.32      |
|         |                      | 88.16     | 7.37      | 3.96      |
| e       | Like X-vector        | 95.04     | 4.77      | 2.89      |
|         |                      | 96.38     | 4.35      | 2.68      |
Experiment Results

We compare the proposed method with VGG, TDNN, ResNet, and Transformer.

- We boost the Transformer to be competitive or superior to VGG, TDNN, and ResNet-like networks.
- Compared to previous Transformers, we achieve significant improvement.
- Our Transformer classification model achieves the state-of-the-art performance.

| Training on VoxCeleb1 development | Implementaion | Extractor  | ACC (%) | EER (%) |
|-----------------------------------|---------------|------------|---------|---------|
| VGG-M                             | VGG           | 80.5       | 7.8     |
| X-vector                          | TDNN          | -          | 7.83    |
| Atten. Stats.                     | TDNN          | -          | **3.85**|
| Cai et al.                        | ResNet        | 89.9       | 4.46    |
| Chung et al.                      | ResNet        | 89.0       | 5.26    |
| SAEP                              | Transformer   | -          | 7.13    |
| S-vectors                         | Transformer   | -          | 5.50    |
| Our work (e)                      | CNN+Transformer | **96.38** | 4.35    |

| Training on VoxCeleb2 development | Implementaion | Extractor  | EER (%) |
|-----------------------------------|---------------|------------|---------|
| MHA                               | VGG           | 3.19       |
| Atten. Stats.                     | TDNN          | 2.59       |
| Xie et al.                        | ResNet        | 3.22       |
| SAEP                              | Transformer   | 5.44       |
| S-vectors                         | Transformer   | 2.67       |
| Our work (a)                      | CNN+Transformer | **2.56**  |
| Our work (e)                      | CNN+Transformer | 2.68      |
Conclusion

- We propose a multi-view self-attention mechanism for Transformer-based speaker networks, which enable to capture global dependencies and model the locality.

- We study the proposed multi-view self-attention mechanism in five different Transformer variants with different network architectures, embedding locations, and pooling methods.

- Our method achieves 96.38% top-1 accuracy for speaker identification task on Voxceleb1 and 4.35% and 2.56% EER on VoxCeleb1 and VoxCeleb2, respectively, for speaker verification task.
Multi-View Self-Attention based Transformer for Speaker Recognition

ICASSP 2022

Rui Wang\textsuperscript{1}\textsuperscript{*}, Junyi Ao\textsuperscript{2,3}\textsuperscript{*}, Long Zhou\textsuperscript{4}, Shujie Liu\textsuperscript{4}, Zhihua Wei\textsuperscript{1}, Tom Ko\textsuperscript{2}, Qing Li\textsuperscript{3}, Yu Zhang\textsuperscript{2}

\textsuperscript{1}Department of Computer Science and Technology, Tongji University
\textsuperscript{2}Department of Computer Science and Engineering, Southern University of Science and Technology
\textsuperscript{3}Department of Computing, The Hong Kong Polytechnic University
\textsuperscript{4}Microsoft Research Asia

\textsuperscript{*}Equal contribution. Work done during internship at Microsoft Research Asia.