Self-attention Networks for Speaker Identification with Negative-Focused Triplet Loss

J Mo¹, L Xu*
¹College of Electrical Engineering Zhejiang University, Hangzhou 310027, China
*Zhejiang University Robotics Institute, Yuyao, Zhejiang 315400, China

Abstract. In this paper, a model based on self-attention for text-independent speaker identification task is proposed. Traditional CNNs are expert at capturing local features but are inefficient when capturing long-range dependencies. The self-attention helps to capture long-range dependency more efficiently and requires less parameters. Besides, based on Triplet Loss and inspired by Weighted Triplet Loss, we propose a novel loss function for speaker identification task, named Negative-Focused Triplet Loss, which makes the training process more efficient and effective. To evaluate the performance of our model, the experiments are conducted on Voxceleb dataset, and we achieve Top-1 accuracy of 90.3% and Top-5 accuracy of 96.9%, which are competitive with the previous state-of-the-art performance. Without any hard sample selecting operation, we achieve better result than the baseline method that uses Cluster-Range Loss or Triplet Loss, which demonstrates the high efficiency of our proposed approach.

1. Introduction
Speaker recognition (SR), including speaker verification (SV) and speaker identification (SI), and this paper focuses on the latter. The speaker identification aims to find out the speaker identity in the enrollment list corresponding to the current given test speaker utterance, and thus the task could be treated as a multi-class classification process. Both speaker verification and speaker identification could be further divided into text-dependent and text-independent. In this work, we will deal with the text-independent situation, which is more practical and meaningful.

For speaker identification, there have been a lot of approaches proposed in the previous literature. The traditional methods include hidden markov models (HMMs)[1][2], vector quantization (VQ)[3], Gaussian mixture models (GMMs)[1] and Gaussian mixture model-Universal background models (GMM-UBM)[4]. Support vector machines (SVMs)[5] and the famous i-vector[6] model were also explored in this field. It is worth mentioning that the i-vector is still the dominated method for speaker identification, due to its great success in a lot of previous research.

Recently, with the rapid development of deep learning, deep neural networks (DNNs), convolutional neural networks (CNNs) and recurrent neural networks (RNNs) have been explored in the field of speaker recognition. To achieve better performance, CNNs are usually designed to be very deep or very wide, such as ResNet[7] and Inception[8][9]. In order to obtain long-range dependencies, the CNNs are required to stack a lot of layers. However, the attention mechanism obtains long-range dependencies in a relatively more direct way. Recently the attention mechanism has been successfully applied in the field of natural language processing [10] and other deep learning fields. Besides, except for the neural networks, a well-design loss function is also worthy noticing[11][12].
In this work, we follow the experience in [13][14] and also insert self-attention block in our networks. Furthermore, to train the model more effectively and more efficiently, we propose a novel loss function – Negative-Focused Triplet Loss, which is based on normal Triplet Loss and inspired by the Weighted Triplet Loss[15].

The rest of this paper is organized as follows. The background of some basic blocks for this work in given in Section 2. Then the proposed methods are presented in Section 3. The experimental setup is shown in Section 4 and the results of the experiments will be reported in Section 5. Finally, the conclusions are given in Section 6.

2. Background

In this section, a simple description of self-attention and triplet-loss will be given. Based on these blocks, we propose our framework.

2.1. Self-Attention

In recent literature, attention mechanism has had a high reputation from researchers and evoked strong repercussions in some fields such as natural language learning (NLP) and computer vision (CV). Traditional CNNs are not efficient in capturing long-range dependencies, while this is exactly what attention mechanism experts in. In the mechanism, there are often three vectors, i.e., \( Q, K \) and \( V \). The formulation below is the most commonly used attention computation form.

\[
\text{Attention}(Q, K, V) = \text{softmax}(QK^T)V
\]  

Self-attention is a special type of attention mechanism, where the \( Q, K \) and \( V \) are all from a given input. The employing of self-attention mechanism helps to quickly catch the global relationships of the whole feature map. Given an input \( x \in \mathbb{R}^{N \times W \times C} \) of the previous layer, it is firstly converted to three feature spaces. Here we name the three feature spaces \( f, g, h \) respectively. The conversion could be realized by a dense layer or a \( 1 \times 1 \) convolutional layer. Then, we have \( f(x) = W_f x \), \( g(x) = W_g x \) and \( h(x) = W_h x \). To calculate the attention map, we have:

\[
\alpha_{i,j} = f(s_{i,j}) = \frac{e^{s_{i,j}}}{\sum_{j=1}^{N} e^{s_{i,j}}}
\]

where \( f(s_{i,j}) \) is a soft-max function, and \( s_{i,j} = f_i^T g \). To obtain the output of self-attention, we have:

\[
o_i = \sum_{j=1}^{N} \alpha_{i,j} h_j
\]

2.2. Triplet Loss

Triplet loss was originally used in FaceNet [16] and from then on has been widely applied to face recognition and speaker recognition, along with which came various variants of triplet loss and they also achieve satisfactory performance.

In this work, we follow the form in [17], where the triplet loss is of cosine similarity metric. To construct the triplet loss, three kinds of embedding vectors are required, i.e., anchors, positives, and negatives. When selecting an embedding vector as the anchor, the positives refer to those embedding vectors from the same speaker as the anchor, while the negatives refer to those embedding vectors of other speakers. We hope the embeddings of same speaker get as close as possible, whilst the embeddings of different speakers get far away from each other. And thus we hope that,

\[
s^a + \alpha < s^p, \quad \forall (a, p, n) \in T
\]

where \( \alpha \) is used to restrict the distance between positive and negative embedding vectors. \( T \) denotes the set of all possible constructed triplets of the dataset. \( s^a \) means the cosine similarity of anchor-positive (a-p) pair, and \( s^a \) means the cosine similarity of anchor-negative (a-n) pair. Formally, for \( N \) triplets we have:
where \([\cdot]_+\) means the function max \((x, 0)\).

3. Proposed Methods
In this work the ResNet is combined with self-attention mechanism. Besides, inspired by weighted triplet loss, a novel loss function of more compact form is proposed to train our model.

3.1. Input Feature
To conduct the experiment, we need to split the dataset into training set and evaluation set. Before extracting acoustic feature, we apply voice activity detection (VAD) to each utterance so as to remove the silent segment. After that, we truncate 800 frames for each utterance. Next, 64-dimensional f-bank feature is extracted for each utterance. Finally, when constructing a batch we again randomly truncate 320 frames from the pre-truncated frames for each utterance.

3.2. Self-Attention Block
The self-attention employed in this work follows the experience in [18]. The output of self-attention has the same size as the input. However, like [18], the output of self-attention is not directly added to the original input but is first multiplied with a trainable parameter \(\gamma\), which is initialized with 0. Such operation makes the networks depends on the feature learned by the former convolutional layers in the early stage of training. As the number of training epochs gets larger, the value of \(\gamma\) also become higher. During the late stages of training, the network focuses more on the output of self-attention block.

\[
z_i = \gamma o_i + x_i
\]  

(6)

3.3. Negative-Focused Triplet Loss
In [15], a variant of triplet loss named Weighted Triplet Loss is proposed. In Weighted Triplet Loss, the weights imposed in the terms of loss function are obtained by computation and keep changing as the training process. In this work, we propose a loss function based on triplet loss, and also adopt the idea of weighting. However, the weight in our loss function is a hyper-parameter, and is only imposed in front of \(s^{(n)}\). Formally, given a batch, we have:

\[
L = \sum_{i=1}^{N_a} \left[ (1+\eta) s^{(n)} - s^{(p)} + \alpha \right]_+
\]  

(7)

where \(\eta\) is a small number. In this work, we found 0.0005 is the best value for \(\eta\). When \(\eta\) is 0, the Negative-Focused Triplet Loss degrades into normal Triplet Loss.

To explain the original purpose of designing the Negative-Focused Triplet Loss, we show it in Figure 1. As mentioned before, the speaker identification can be treated as a multi-class classification task. We obtain the output class by observing the scores generated by the classifier for each class, and the class with the highest score will be chosen as the class of the given input. However, even the highest score might not be very high for those hard samples, but it does not affect the final result as long as the highest score points to the correct class. On the other hand, when the highest score points to a wrong class with high certainty, the accuracy of the classification also gets lower. Take Figure 1 for example, \(b\) is tended to classified to the class of \(c\), with a high score because the similarity between \(b\) and \(c\) is relatively high, even though that will be wrong. However, \(b\) could be also classified to the correct class with a low score, because the similarity between \(b\) and \(a\) is relatively low. Both increasing \(S_1\) and decreasing \(S_2\) will help to correctly classify \(b\). However, \(b\) and \(c\) are essentially different for they are from different speakers. We argue that it is easier to decrease \(S_2\) than increasing \(S_1\). Therefore, we impose a weight which is a little larger than 1 on \(s^{(n)}\) so as to slightly speed up the decrease of \(S_2\). The imposed weight should not be too large, because there is a delicate balance
relationship between $s^{a,r}$ and $s^{a,p}$. Provided that $\eta$ is set too large, the model will focus too much on the decrease of $s^{a,r}$, and the increasing of $s^{a,p}$ will be relatively ignored.

Figure 1. Learning process of Negative-Focused Triplet Loss

4. Experimental Setup

4.1. Dataset and Evaluation
Experiments in this paper are all performed on Voxceleb-1 dataset. The utterances in this dataset are extracted from videos in Youtube, and they are from 1251 speakers, with a balanced sex ratio. For speaker identification, the most used evaluation criterion is Top-1 accuracy and Top-5 accuracy. The report of Top-1 accuracy and Top-5 accuracy of our model tested on Voxceleb-1 dataset will be given in Section 5.

4.2. Training Method
To train the model in this work, we employ SGD with a momentum of 0.99 as the optimizer. The initial learning rate is set to 0.005 and the learning rate ends with 0.0005. During the last 10 epochs the learning rate is set to 0.0001. Each test model will be trained for 100 epochs. The major difference between this work and [14] is the employed loss function. To compare the performance of Cluster-Range Loss[14] and our proposed Negative-Focused Triplet Loss, both loss functions are tested in the experiments. To explore the best value for $\eta$, we also set $\eta$ to different values. Both Cluster-Range Loss and Negative-Focused Triplet Loss are employed to train the speaker embeddings, while classifier is trained by the softmax cross entropy function. Such operation is achieved by cutting the gradient from the fully-connected layer into the embeddings. In [14], the jointly training method is reported to be better than simply employing softmax cross entropy as the loss function. The architecture of the neural networks in this work is illustrated as Figure 2.

Figure 2. Architecture of the neural networks in this work
5. Results
As is presented in Table 1, the best value for $\eta$ is 0.0005 and all the results with our loss is better than that of Triplet Loss ($\eta = 0$). As is shown in Figure 3, by utilizing our proposed Negative-Focused Triplet Loss, the entire accuracy curves of both Top-1 accuracy and Top-5 accuracy lies above that of Cluster-Range Loss, from the very beginning to the end. Our model achieves the Top-1 accuracy of 90.3% and Top-5 accuracy of 96.9%. Our Top-1 accuracy is also competitive compared with the best results given in [13], where Top-1 accuracy is 90.8% and Top-5 accuracy is 96.5%, which are the highest results in published paper that we can find. Besides, our Top-5 accuracy is higher than the result in [13]. The results strongly demonstrate the effectiveness of our approach.

Table 1. Results for different value of $\eta$

| $\eta$     | Top-1 accuracy | Top-5 accuracy |
|------------|----------------|----------------|
| 0.0004     | 87.9           | 95.7           |
| 0.0005     | **90.3**       | **96.9**       |
| 0.0006     | 87.8           | 95.8           |
| 0 (Triplet Loss) | 86.7         | 95.4           |

Figure 3. ACC curves of corresponding loss function.

5.1. Evaluating Self-attention Mechanism
The framework of this paper is based on the work of [14], and the effectiveness of self-attention has been reported. Due to the limitation of article length, a list for comparison is not given here. As is reported in [13][14], the Res-SA architecture achieves much better results than the methods using i-vector, and also better than those CNNs without self-attention. The significant performance strongly demonstrates the effectiveness of self-attention mechanism.

5.2. Evaluating Negative-Focused Triplet Loss
Compared with normal Triplet Loss, the Negative-Focused Triplet Loss treats $s^{+\cdot}$ and $s^{+\cdot}$ in the loss function differently. By multiplying $s^{+\cdot}$ with a hyper-parameter which is a little bit larger than 1, the identification performance is significantly improved. Compared with Res-SA + Triplet Loss, our Res-SA + NFTL architecture improves the Top-1 accuracy by 3.6%. When comparing with Res-SA + CRL architecture, our results surpass its Top-1 accuracy by 1.2% and Top-5 accuracy of 1.1%, which demonstrates that our proposed Negative Focused Triplet Loss is more efficient than Cluster-Range Loss when dealing with speaker identification task.

6. Conclusions
In this paper, a self-attention based residual network is employed and a novel loss is proposed for the speaker identification task. With the proposed Negative-Focused Triplet Loss, our model achieves competitive results compared with previous state-of-the-art results. No complex operation is used, our
proposed loss could be applied to any work dealing with speaker identification using triplet loss. Besides, the effect of hard samples selecting is not yet explored in this work. Future work will focus on further consideration of employing hard sample selection, which is very important in previous literature.

References
[1] Reynolds D A and Rose R C 1995 Robust text-independent speaker identification using Gaussian mixture speaker models. *IEEE transactions on speech and audio processing*. 3 72-83
[2] Rabiner L R, Levinson S E and Sondhi M M 1983 On the application of vector quantization and hidden Markov models to speaker-independent, isolated word recognition. *Bell System Technical Journal*. 62 1075-1105
[3] Soong F K, Rosenberg A E, Juang B H and Rabiner L R 1987 Report: A vector quantization approach to speaker recognition *AT&T technical journal*, 66 14-26
[4] Reynolds D A, Quatieri T F and Dunn R B 2000 Speaker verification using adapted Gaussian mixture models *Digital signal processing*. 10 19-41
[5] Campbell W M, Campbell J P, Reynolds D A, Singer E and Torres-Carrasquillo P A 2006 Support vector machines for speaker and language recognition *Computer Speech & Language* 20 210-29
[6] Dehak N, Kenny P J, Dehak R, Dumouchel P and Ouellet P 2010 Front-end factor analysis for speaker verification *IEEE Transactions on Audio, Speech, and Language Processing*. 19 788-798
[7] He K, Zhang X, Ren S and Sun J 2016 Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* pp 770-778
[8] Szegedy C, Ioffe S, Vanhoucke V and Alemi A A 2017 Inception-v4, inception-resnet and the impact of residual connections on learning *Thirty-first AAAI conference on artificial intelligence*.
[9] Szegedy C, Vanhoucke V, Ioffe S, Shlens J and Wojna Z 2016 Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition* pp 2818-2826
[10] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez A N and Polosukhin I 2017 Attention is all you need *Advances in neural information processing systems* pp 5998-6008
[11] Wan L, Wang Q, Papir A and Moreno I L 2018 Generalized end-to-end loss for speaker verification. In 2018 *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* pp 4879-4883
[12] Cheng D, Gong Y, Zhou S, Wang J and Zheng N 2016 Person re-identification by multi-channel parts-based cnn with improved triplet loss function. In *Proceedings of the iEEE conference on computer vision and pattern recognition* pp 1335-1344
[13] An N N, Thanh N Q and Liu Y 2019 Deep CNNs with self-attention for speaker identification *IEEE Access*, 7, 85327-85337.
[14] Bian T, Chen F and Xu L 2019 Self-attention based speaker recognition using Cluster-Range Loss *Neurocomputing*. 368 59-68
[15] Yu J, Zhu C, Zhang J, Huang Q and Tao D 2019 Spatial pyramid-enhanced NetVLAD with weighted triplet loss for place recognition *IEEE transactions on neural networks and learning systems*.
[16] Schroff F, Kalenichenko D and Philbin J 2015 Facenet: A unified embedding for face recognition and clustering In *Proceedings of the IEEE conference on computer vision and pattern recognition* pp 815-823
[17] Li C, Ma X, Jiang B, Li X, Zhang X, Liu X and Zhu Z 2017 Deep speaker: an end-to-end neural speaker embedding system arXiv preprint arXiv:1705.02304.
[18] Zhang H, Goodfellow I, Metaxas D and Odena A 2018 Self-attention generative adversarial networks arXiv preprint arXiv:1805.08318