Deep Discriminative Embedding with Ranked Weight for Speaker Verification

Dao Zhou[1], Longbiao Wang[1(✉)], Kong Aik Lee[2], Meng Liu[1], and Jianwu Dang[1,3]

[1] Tianjin Key Laboratory of Cognitive Computing and Application, College of Intelligence and Computing, Tianjin University, Tianjin, China
{zhoudao,longbiao_wang,liumeng2017}@tju.edu.cn
[2] Institute for Infocomm Research, A*STAR, Singapore, Singapore
kongaik.lee@gmail.com
[3] Japan Advanced Institute of Science and Technology, Nomi, Ishikawa, Japan
jdang@jaist.ac.jp

Abstract. Deep speaker-embedding neural network trained with a discriminative loss function is widely known to be effective for speaker verification task. Notably, angular margin softmax loss, and its variants, were proposed to promote intra-class compactness. However, it is worth noticing that these methods are not effective enough in enhancing inter-class separability. In this paper, we present a ranked weight loss which explicitly encourages intra-class compactness and enhances inter-class separability simultaneously. During the neural network training process, the most attention is given to the target speaker in order to encourage intra-class compactness. Next, its nearest neighbor who has the greatest impact on the correct classification gets the second most attention while the least attention is paid to its farthest neighbor. Experimental results on VoxCeleb1, CN-Celeb and the Speakers in the Wild (SITW) core-core condition show that the proposed ranked weight loss achieves state-of-the-art performance.

Keywords: Speaker verification · Speaker embedding · Intra-class compactness · Inter-class separability

1 Introduction

Automatic speaker verification (ASV) is the process of automatically validating a claimed identity by analyzing the spoken utterance from the speaker. Speaker verification technology has been found important in various applications, such as, public security, anti-terrorism, justice, and telephone banking. Over the past few years, i-vector based representation [1], used in conjunction with a Probabilistic Linear Discriminant Analysis [2] backend, has been the state-of-the-art technique and has been deployed in most implementations.

H. Yang et al. (Eds.): ICONIP 2020, CCIS 1333, pp. 79–86, 2020.
https://doi.org/10.1007/978-3-030-63823-8_10

With the advancement in deep learning, performance of ASV has been greatly improved due to the large learning capacity of the deep neural network. In [3], a system based on convolutional neural network (CNN) outperforms the i-vectors/PLDA system. The aim of a speaker-embedding neural network [4] is to obtain a true representation (i.e., minimum intra-class differences) of a speaker's voice that is sufficiently different from other speakers (i.e., maximum inter-class differences). Apart from the neural network architecture, the loss function plays an important role to achieve this goal. In this regard, two types of loss functions, namely, metric-learning loss and classification loss, have shown to be effective for training speaker-embedding neural networks. Metric-learning loss [5,6] use pairwise or tripletwise training samples to learn discriminative features based on distance metric learning. Triplet loss [6] achieves both optimizations simultaneously by taking three training samples as input at one time.

The use of classification loss for speaker embedding neural network has been a major topic of interest. Classification loss, which includes softmax loss and its variants like angular softmax (A-Softmax) loss [7], additive margin softmax (AM-Softmax) loss [8] and additive angular margin softmax (Arc-Softmax) loss [9] have been proposed. A-Softmax loss and AM-Softmax loss introduce angular margin and cosine margin into the softmax loss respectively that tend to focus on encouraging intra-class compactness. Arc-Softmax loss incorporates the angular margin in an additive manner, that is different from the multiplicative angular margin in A-Softmax loss, to learn highly discriminative feature. These methods are not only simpler to implement compared to the triplet loss, but also give impressive performance in ASV tasks [10]. However, these variants of softmax loss have not paid special attention to inter-class separability. Very recently, [11] proposed the exclusive regularization to encourage inter-class separability for face verification and achieved outstanding performance.

In this paper, we present a ranked weight loss which explicitly encourages intra-class compactness and enhances inter-class separability simultaneously, which are achieved by paying the most attention and less attention to the target speaker and the speaker who is farther to the target speaker respectively during the neural network training process. The efficacy of the proposed ranked weight loss is validated on VoxCeleb1, CN-Celeb and SITW [3,12,13] corpora.

2 Prior Works

The proposed ranked weight loss is mainly inspired by the work of AM-Softmax loss [8] and exclusive regularization [11], we start with the definition of them.

2.1 AM-Softmax Loss

The traditional softmax loss is presented as:

$$L_S = \frac{1}{N}\sum_{i=1}^{N} -log\frac{e^{\mathbf{W}_{y_i}^T \mathbf{x}_i + b_{y_i}}}{\sum_{j=1}^{n} e^{\mathbf{W}_j^T \mathbf{x}_i + b_j}} \tag{1}$$

where N and n denote the batch size and the number of classes respectively, $\mathbf{x}_i$ is the feature vector of the i-th sample that belongs to class y_i and b is the bias term. $\mathbf{W}_j$ is the weight vector of class j. Here, $\mathbf{W}_{y_i}^T \mathbf{x}_i$ can be reformulated as $\|\mathbf{W}_{y_i}\| \|\mathbf{x}_i\| cos(\theta_{y_i,i})$.

However, the learned embeddings are not discriminative enough under the supervision of softmax loss. To address this issue, AM-Softmax loss [8,10] introduced an additive cosine margin into softmax loss to minimize intra-class distance. In AM-Softmax loss, the bias term b is discarded, weight $\mathbf{W}$ and feature x are normalized, and a hyperparameter s is introduced to scale the cosine values. AM-Softmax loss is given by:

$$L_{AM} = \frac{1}{N} \sum_{i=1}^{N} -log \frac{e^{s \cdot (cos(\theta_{y_i,i}) - m)}}{e^{s \cdot (cos(\theta_{y_i,i}) - m)} + \sum\limits_{j=1; j \neq y_i}^{n} e^{s \cdot cos(\theta_{j,i})}} \tag{2}$$

in which m is a factor used to control the cosine margin.

2.2 Exclusive Regularization

Recently, [11] proposed the exclusive regularization to enlarge inter-class distance by penalizing the angle between a target class and its nearest neighbor. The formulation of exclusive regularization is defined as:

$$\begin{aligned} L_R &= \frac{1}{n} \sum_{y_i=1}^{n} \max_{j \neq y_i} cos(\psi_{y_i,j}) \\ &= \frac{1}{n} \sum_{y_i=1}^{n} \max_{j \neq y_i} \frac{\mathbf{W}_{y_i} \mathbf{W}_j}{\|\mathbf{W}_{y_i}\| \|\mathbf{W}_j\|} \end{aligned} \tag{3}$$

where $\mathbf{W}_{y_i} \in \mathbb{R}^d$ is the y_i-th column of the weight matrix $\mathbf{W} \in \mathbb{R}^{d \times n}$, d represents the dimension of feature vectors and n is the number of classes. $\mathbf{W}_{y_i}$ can be regarded as the cluster center of class y_i, and $\psi_{y_i,j}$ is the angle between $\mathbf{W}_{y_i}$ and $\mathbf{W}_j$.

When applying the exclusive regularization to cooperate with AM-Softmax loss to supervise the model, the overall loss function (AME-Softmax loss) can be represented as:

$$L_{AME} = (1 - \lambda) L_{AM} + \lambda L_R \tag{4}$$

in which λ is the tradeoff between the AM-Softmax loss and the exclusive regularization.

3 Ranked Weight Loss

As stated in Sect. 2, AM-Softmax loss [8] achieves the promising performance in ASV tasks by encouraging intra-class compactness, and [11] further proposed the exclusive regularization to enhance inter-class separability. Nevertheless, the

exclusive regularization only considers the impact of the target speaker's nearest neighbor and ignores the impact from other speakers.

We propose a ranked weight loss to encourage intra-class compactness and enlarge inter-class separability more fully. During the neural network training process, our method pays the most attention to the target speaker, while the corresponding attention is paid to its neighbors according to the inter-class distance. $\mathbf{W}_{y_i}$ is regarded as the cluster center of speaker y_i in our proposed method. For a training set with m speakers, the inter-class distance can be defined as:

$$d(y_i, j) = 1 - \frac{\mathbf{W}_{y_i}\mathbf{W}_j}{\|\mathbf{W}_{y_i}\|\,\|\mathbf{W}_j\|}, y_i \neq j \tag{5}$$

In addition, we set $d(y_i, y_i) = 0$, and rank the distances in accending order as follows:

$$d(y_i, y_i) < ... \leq d(y_i, j) \leq ... \leq d(y_i, m) \tag{6}$$

The ranked weights are then given by:

$$w_j^{'} = \frac{1 - d(y_i, j)}{\sum_{k=1}^{m}(1 - d(y_i, k))^2} \tag{7}$$

In order to ensure that the ranked weights are positive so as not to affect the gradient direction of the loss function, we perform an exponential operation on the ranked weight function, the ranked weight function is redefined as:

$$w_j = e^{w_j^{'}} \tag{8}$$

Therefore, the ranked weights are sorted in descending order as:

$$w_{y_i} > ... \geq w_j \geq ... \geq w_m \tag{9}$$

From Eq. (6) and Eq. (9), we observe that the target speaker y_i gets the most weight, while the smaller the inter-class distance, the greater the corresponding class weight. Finally, we introduce the ranked weight into AM-Softmax loss, leading to the ranked weight loss (RAM-Softmax loss), which is given by:

$$L_{RAM} = \frac{1}{N}\sum_{i=1}^{N} -log\frac{e^{w_{y_i}s\cdot(cos(\theta_{y_i,i})-m)}}{e^{w_{y_i}s\cdot(cos(\theta_{y_i,i})-m)} + \sum\limits_{j=1;j\neq y_i}^{n} e^{w_j s\cdot cos(\theta_{j,i})}} \tag{10}$$

We see that the most weight and the corresponding weight are paid to increase the value of $cos(\theta_{y_i,i})$ and reduce the value of $cos(\theta_{j,i})$ respectively, which are inversely related to intra-class distance (i.e., cosine distance of feature-to-center) and inter-class distance, respectively.

4 Experimental Settings

4.1 Dataset

In our experiments, we validate the effectiveness of our proposed method on VoxCeleb1 [3] and CN-Celeb [12] datasets that are collected 'in the wild'. CN-Celeb

contains more genres and is more challenging than VoxCeleb1. Each utterance in VoxCeleb1 is no less than 3 s long, while more than 30% of the utterances are less than 2 s long in CN-Celeb. Both corpora are divided into development set and test set, respectively. The development set of VoxCeleb1, contains 148,642 utterances from 1,211 speakers, is used for model training, and the test set of VoxCeleb1 involves 4,874 utterances from 40 speakers. In CN-Celeb, 111,260 utterances from 800 speakers make up the first part CN-Celeb(T) (training set), and the second part CN-Celeb(E) (test set) consists of 18,849 utterances from 200 speakers. In addition, the SITW core-core condition is also used to evaluate the performance of the proposed ranked weight loss.

4.2 Implementation Details

In our experiments, the residual CNN (ResCNN) is used to training our model. The ResCNN contains 4 residual modules, which is simliar to the architecture in [6], but the depth of them are 2 instead of 3, then followed by an adaptive average pooling of size 4×1. The 1024-dimensional speaker embeddings are extracted from the fully connected layer.

During training, we randomly sample 3-s segments from each utterance to generate the input spectrograms through a sliding hamming window, window width and step are 20 ms and 10 ms respectively. The model is trained with a mini-batch size of 64. We used the standard stochastic gradient descent (SGD) as the optimizer. The initial learning rate is set to 0.1. And the additive cosine margin m is 0.2, while the angular margin terms are set to 2 and 0.2 in A-Softmax loss and Arc-Softmax loss respectively. Cosine similarity and equal error rate (EER) are used for back-end scoring method and performance evaluation metric, respectively.

5 Experimental Results

5.1 Exclusive Regularization Vs. Ranked Weight

To compare the performance of exclusive regularization and our proposed ranked weight, we conduct a series of experiments on CN-Celeb and VoxCeleb1. We follow [8] to set the scale factor s to 30, and the balance factor λ in AME-Softmax loss is set to different values (i.e., 0.1, 0.3 and 0.5) to show the effectiveness of exclusive regularization comprehensively. AM-Softmax loss is used as baseline method. Table 1 shows results of the above three methods. The ranked weight loss (RAM-Softmax loss) achieves the lowest EER compared with AM-Softmax loss and AME-Softmax loss. The relative reduction in EER amounts to 10.37% and 9.05% on VoxCeleb1 while 13.95% and 12.83% on CN-Celeb, respectively.

To explore the distribution of embeddings under different loss functions. We randomly sample 40 utterances for each speaker in the test set of VoxCeleb1, the distance information of these embeddings are illustrated in Table 2. Here, the intra-class distance refers to the average Euclidean distance from each sample to

Table 1. Verification performance of different loss function based systems on VoxCeleb1 and CN-Celeb. λ refers to the balance factor in AME-Softmax.

Dataset	Loss	λ	EER (%)
VoxCeleb1	AM-Softmax	–	4.82
	AME-Softmax	0.1	4.77
	AME-Softmax	0.3	4.75
	AME-Softmax	0.5	4.84
	RAM-Softmax	–	**4.32**
CN-Celeb	AM-Softmax	–	16.42
	AME-Softmax	0.1	16.21
	AME-Softmax	0.3	16.73
	AME-Softmax	0.5	16.33
	RAM-Softmax	–	**14.13**

the center of the corresponding class, while the inter-class distance refers to the average Euclidean distance between the centers of the classes. The center of class is computed by the average position of samples from that class. The proposed ranked weight loss achieves the largest inter-class distance and the smallest intra-class distance compared with AM-Softmax loss and AME-Softmax loss.

Table 2. The distance statistics under different loss functions. 'Intra' refers to intra-class distance, and 'Inter (Top-k)' refers to average distance between the target speaker and its k nearest neighbors.

	AM-Softmax	AME-Softmax	RAM-Softmax
Intra	7.94	7.70	**5.00**
Inter (Top-1)	13.68	14.12	**14.85**
Inter (Top-2)	15.78	15.96	**16.64**
Inter (Top-3)	17.38	17.35	**18.28**
Inter (Top-5)	20.25	20.27	**21.32**
Inter (Top-10)	26.31	26.72	**27.95**

Furthermore, we explore the performance of the above three methods on SITW core-core test set. Note that the speakers that overlap with the VoxCeleb1 *dev* (development set) were removed. Verification results are presented in Table 3. It is apparent that the proposed ranked weight loss achieves better performance than AM-Softmax loss and AME-Softmax loss.

Table 3. Verification performance on SITW core-core test set.

Training set	Loss	EER (%)
VoxCeleb1 *dev*	AM-Softmax	10.95
	AME-Softmax	10.72
	RAM-Softmax	**9.92**
CN-Celeb(T)	AM-Softmax	26.67
	AME-Softmax	26.81
	RAM-Softmax	**22.78**

5.2 Compared with Other Methods

Finally, we compare the performance of our ranked weight loss with the other state-of-the-art methods in Table 4. It is worth noticing that deep speaker systems perform better than traditional i-vector system on VoxCeleb1 while the i-vector system shows promising performance on CN-Celeb, which demonstrates that CN-Celeb is significantly different from VoxCeleb1. However, our proposed ranked weight loss achieves the best result on both datasets, it indicates that our method is more robust than i-vector and other loss functions.

Table 4. Verification performance of different ASV systems.

Dataset	Front model	Loss	Dims	Back-end scoring	EER (%)
VoxCeleb1	i-vector [3]	–	–	PLDA	8.80
	ResNet-34 [14]	A-Softmax	128	PLDA	4.46
	ResCNN	Softmax	1024	Cosine	6.44
	ResCNN	Arc-Softmax	1024	Cosine	4.64
	ResCNN	A-Softmax	1024	Cosine	4.43
	ResCNN	RAM-Softmax	1024	Cosine	**4.32**
CN-Celeb	i-vector [12]	–	150	LDA-PLDA	14.24
	TDNN [12]	Softmax	150	LDA-PLDA	14.78
	ResCNN	Softmax	1024	Cosine	16.85
	ResCNN	A-Softmax	1024	Cosine	15.91
	ResCNN	Arc-Softmax	1024	Cosine	15.41
	ResCNN	RAM-Softmax	1024	Cosine	**14.13**

6 Conclusions

We have presented a ranked weight loss, which explicitly enhances intra-class compactness and inter-class discrepancy simultaneously. This is achieved by paying the most attention to the target speaker and less attention to the speaker

further apart from the target speaker during the neural network training process. Extensive experiments on VoxCeleb1, CN-Celeb and SITW Core condition corpora showed that the proposed ranked weight loss achieved the competitive performance compared with the current state-of-the-art methods.

Acknowledgments. This work was supported in part by the National Natural Science Foundation of China under Grant 61771333, the Tianjin Municipal Science and Technology Project under Grant 18ZXZNGX00330.

References

1. Dehak, N., Kenny, P., Dehak, R., Dumouchel, P., Ouellet, P.: Front-end factor analysis for speaker verification. IEEE Trans. Audio Speech Lang. Process. **19**, 788–798 (2011). https://doi.org/10.1109/TASL.2010.2064307
2. Prince, S., Elder, J.H.: Probabilistic linear discriminant analysis for inferences about identity. In: 2007 IEEE 11th International Conference on Computer Vision, pp. 1–8. IEEE (2007)
3. Nagrani, A., Chung, J.S., Zisserman, A.: VoxCeleb: a large-scale speaker identification dataset. In: Interspeech, pp. 2616–2620 (2017)
4. Snyder, D., Garcia-Romero, D., Sell, G., Povey, D., Khudanpur, S.: X-vectors: robust DNN embeddings for speaker recognition. In: 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, pp. 5329–5333 (2018). https://doi.org/10.1109/ICASSP.2018.8461375
5. Chung, J.S., Nagrani, A., Zisserman, A.: VoxCeleb2: deep speaker recognition. In: Interspeech, pp. 1086–1090 (2018)
6. Li, C., et al.: Deep speaker: an end-to-end neural speaker embedding system (2017)
7. Liu, W., Wen, Y., Yu, Z., Li, M., Raj, B., Song, L.: SphereFace: deep hypersphere embedding for face recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 212–220 (2017)
8. Wang, F., Cheng, J., Liu, W., Liu, H.: Additive margin softmax for face verification. IEEE Signal Process. Lett. **25**, 926–930 (2018)
9. Deng, J., Guo, J., Zafeiriou, S.: ArcFace: additive angular margin loss for deep face recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4690–4699 (2019)
10. Liu, Y., He, L., Liu, J.: Large margin softmax loss for speaker verification. In: Interspeech, pp. 2873–2877 (2019)
11. Zhao, K., Xu, J., Cheng, M.M.: RegularFace: deep face recognition via exclusive regularization. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1136–1144 (2019)
12. Fan, Y., et al.: CN-CELEB: a challenging Chinese speaker recognition dataset (2020)
13. McLaren, M., Ferrer, L., Castan, D., Lawson, A.: The speakers in the wild (SITW) speaker recognition database. In: Interspeech, pp. 818–822 (2016)
14. Cai, W., Chen, J., Li, M.: Exploring the encoding layer and loss function in end-to-end speaker and language recognition system. In: Odyssey 2018 The Speaker and Language Recognition Workshop, pp. 74–81 (2018)