INTERSPEECH 2021
30 August – 3 September, 2021, Brno, Czechia

Variational Information Bottleneck based Regularization for Speaker Recognition

Dan Wang[1], *Yuanjie Dong*[1], *Yaxing Li*[1], *Yunfei Zi*[1], *Zhihui Zhang*[1], *Xiaoqi Li*[1], *Shengwu Xiong*[1,2⋆]

[1]School of Computer Science and Technology, Wuhan University of Technology, China
[2]Sanya Science and Education Innovation Park of Wuhan University of Technology, China

xiongsw@whut.edu.cn

Abstract

Speaker recognition (SR) is inevitably affected by noise in real-life scenarios, resulting in decreased recognition accuracy. In this paper, we introduce a novel regularization method, variable information bottleneck (VIB), in speaker recognition to extract robust speaker embeddings. VIB prompts the neural network to ignore as much speaker-identity irrelevant information as possible. We also propose a more effective network, VovNet with an ultra-lightweight subspace attention module (ULSAM), as a feature extractor. ULSAM infers different attention maps for each feature map subspace, enabling efficient learning of cross-channel information along with multi-scale and multi-frequency feature representation. The experimental results demonstrate that our proposed framework outperforms the ResNet-based baseline by 11.4% in terms of equal error rate (EER). The VIB regularization method gives a further performance boost with an 18.9% EER decrease.

Index Terms: speaker recognition, variational information bottleneck, deep speaker embeddings

1. Introduction

Prior to deep learning, speaker recognition based on i-vector [1] has been dominant. I-vectors are often used in conjunction with probabilistic linear discriminant analysis (PLDA) [2] as a baseline system in many studies. In recent years, deep learning has been pushing the state-of-the-art in many fields of research. Due to the availability of free large-scale datasets [3],[4],[5] and the superior learning ability of deep neural networks (DNNs), speaker recognition based on deep speaker embeddings including d-vector [6] and x-vector [7] shows great potential. Speaker recognition based on DNNs usually consists of three components: feature extraction network, feature aggregation layers, and training loss. In deep speaker embedding learning, convolutional neural networks (CNNs), such as time-delayed neural networks (TDNNs) [7],[8],[9] and ResNet [5],[10],[11],[12],[13], are commonly used to extract speaker information from acoustic features at the frame level. Subsequently, a low-dimensional vector, called deep speaker embedding or x-vector, is obtained by gathering frame-level information to form utterance-level representation. Methods such as statistical pooling [14], max pooling [15], attentive statistical pooling [8], multi-headed attentive statistical pooling [16] are popular choices. Pioneering work on speaker recognition using DNNs has learned speaker embeddings via the classification loss [7],[14]. Since then, the prevailing method has been to use softmax classifiers to train the embeddings [8],[17],[18]. A number of works have demonstrated that a series of variants, A-softmax [19], AM-softmax [20],[21], and AAM-softmax [22], outperform the vanilla softmax in terms of speaker recognition under certain conditions.

However, while speaker embeddings obtained from neural networks exhibit impressive performance, there are still some challenges. On the one hand, the neural network may lose information that helps to identify the speaker during the feature extraction process. On the other hand, a potential problem is that the speaker embeddings inferred from DNNs may contain irrelevant information since noise is bound to exist in practical application scenarios [23]. Therefore, it is necessary to extract more robust deep speaker embeddings that contain only speaker identity-related information to increase the speaker recognition accuracy.

In this paper, we introduce a novel regularization method, the variational information bottleneck (VIB), to extract more robust deep speaker embeddings. The variable information bottleneck forces the neural network to extract speaker embeddings with maximum retention of information relevant to the speaker's identity and removal of irrelevant information. We incorporate the VIB regularization strategy into a modified VovNetV2 feature extraction network that has previously been used for object detection and has shown excellent performance. We replace the effective squeeze and excitation (eSE) module in the original VovNetV2 with ULSAM to obtain more meaningful feature representations. Experiments show that our proposed enhanced VovNet with VIB regularization significantly improves the performance of the speaker recognition system.

2. Method

This section describes the speaker recognition framework used in our experiments and the proposed regularization method based on the variational information bottleneck.

2.1. Learning framework

An overview of our framework is given in Figure 1. We use an enhanced VovNetV2 [24] as a feature extraction network to obtain frame-level speaker representations. The original VovNetV2 is a powerful feature extraction network for object detection that outperforms the classical ResNet [25] and DenseNet [26]. We first apply it to speaker recognition with some modifications to accommodate acoustic feature inputs. The core components of the original VovNetV2 are the one-shot aggregation (OSA) module and the eSE attention mechanism. In our implementation, we choose the simpler yet more effective ULSAM [27] as an alternative to eSE. The specific structure of VovNet2 is described in Section 2.3.

For the sake of simplicity, we select temporal average pooling (TAP) to aggregate variable-length frame-level features into fixed-length utterance-level embeddings. DNNs have exten-

⋆Corresponding author.

http://dx.doi.org/10.21437/Interspeech.2021-482

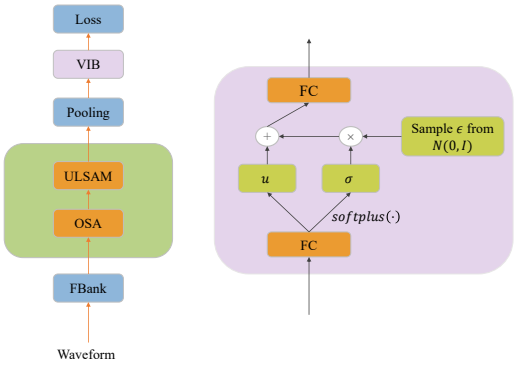

Figure 1: *Left: The overall architecture of the speaker recognition model. Right: The specific structure of the VIB. μ and σ denote the mean and variance, respectively.*

sively explored in the literature for the generation of speaker embedding with different objective functions. The classical loss function is softmax or one of his variants. Our experiments are performed on softmax and AM-softmax.

After the temporal aggregation layer, we add our proposed variational information bottleneck. It removes the information that is irrelevant to the speaker's identity, resulting in a more meaningful embedding. The principle of the variational information bottleneck and how it works in speaker recognition will be described in the next section.

2.2. Variational information bottlenecks

Suppose the input data is X and the desired output is Y. The goal of the information bottleneck (IB) is to learn an encoding Z that is maximally expressive about Y while being maximally compressive about X. Similarly, in speaker recognition, we expect to learn a low-dimensional speaker embedding z that maximizes the retention of information for predicting speaker identity y while removing as much irrelevant information from the input speech x as possible. Using mutual information (MI) to measure relevance and compression in information bottlenecks, Tishby et al. [28] propose the following optimization objective:

$$\max I\left(Z;Y\right)-\beta I\left(X;Z\right) \tag{1}$$

where $I(Z;Y)$ is the MI of Z and Y, $I\left(X;Z\right)$ denotes the MI between X and Z, and the Lagrange multiplier β balances the compression and correlation of the extracted features Z.

Speaker recognition is a supervised classification task, optimized by the following objective:

$$\min \mathbb{E}_{x,y\sim p(x,y)}[-\log q\left(y|x\right)] \tag{2}$$

where x and y correspond to the input speech and the predicted speaker identity, respectively. However, the models optimized by Eq.(2) do not consider what information is learnt aside from being able to predict the given labels. As a result, the obtained speaker embeddings may contain unrelated representations, making the speaker recognition model less robust.

Equivalently, we can maximize the objective function to fit the speaker recognition task:

$$\min -I\left(Z;Y\right)+\beta I\left(X;Z\right) \tag{3}$$

Intuitively, the first term is the objective function for the classification task, which can be replaced by Eq.(2). We consider the second term as an additional penalty to constrain the mutual information between the input speech and the compressed representation. Thus, the loss function of the speaker recognition model with the information bottleneck is given by:

$$\begin{aligned}\mathcal{L}_{IB}=\min \mathbb{E}_{x,y\sim p(x,y)}\left[\mathbb{E}_{z\sim\mathbb{E}(Z|X)}\left[-\log q(y|z)\right]\right.\\ \left.+\beta I\left(X;Z\right)\right]\end{aligned} \tag{4}$$

where z is the output of the information bottleneck layer and β controls the penalty strength. A large β indicates that Z contains less information about the speech input X and the model output Y contains fewer details about X. The resulting compressed encoding Z is noise-resistant to a certain extent.

Write out $I\left(X;Z\right)$ in full, this becomes

$$\begin{aligned}I\left(X;Z\right)=&\int \mathrm{d}z\mathrm{d}x p\left(x,z\right)\log p\left(z|x\right)\\ &-\int \mathrm{d}z\mathrm{d}x p\left(x,z\right)\log p\left(z\right)\end{aligned} \tag{5}$$

where $p\left(x,z\right)$ is the joint distribution and x,y,z are instances corresponding to X,Y,Z, respectively.

Obviously, $p(z)$ is the obtained compression result, which cannot be calculated directly, we let the variational approximation $r(z)$ as a substitute. Since Kullback-Leibler divergence is constantly positive, i. e., $KL[p(Z),r(Z)]\geq 0$, the KL of $p(z)$ and $r(z)$ are obtained as follows:

$$\int \mathrm{d}z p\left(z\right)\log p\left(z\right)\geq\int \mathrm{d}z p\left(z\right)\log r\left(z\right) \tag{6}$$

Combining equations (5) and (6), we have the following upper bound:

$$\begin{aligned}I\left(X;Z\right)&\leq\int \mathrm{d}z\mathrm{d}x p\left(x\right)p\left(z|x\right)\log\frac{p\left(z|x\right)}{r\left(z\right)}\\ &=\mathbb{E}_{x\sim p(x)}\left[KL\left[p\left(z|x\right)|r\left(z\right)\right]\right]\end{aligned} \tag{7}$$

In summary, the final loss function is:

$$\begin{aligned}\hat{\mathcal{L}}_{VIB}=\min \mathbb{E}_{x\sim p(x)}\left[\mathbb{E}_{z\sim\mathbb{E}(Z|X)}\left[-\log q\left(y|z\right)\right]\right.\\ \left.+\beta KL\left[p\left(z|x\right)|r\left(z\right)\right]\right]\end{aligned} \tag{8}$$

where $\hat{\mathcal{L}}_{VIB}\geq\mathcal{L}_{IB}$.

Overall, the variational information bottleneck balances data compression and information retention, forcing the neural network to forget useless or harmful information and retaining the most valid information to obtain more robust and discriminative speaker embeddings.

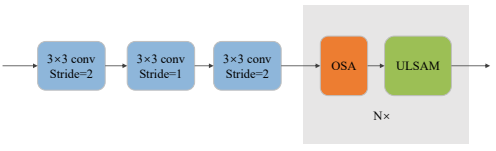

Figure 2: *The architecture of modified VovNetV2. It consists of three convolutional layers and N=4 stacked OSA and ULSAM modules.*

2.3. Feature extraction network

The modified VovNetV2 used in our implementation consists of a stem block including 3 convolutional layers, 4 stages of OSA modules with an output stride of 32, and ULSAM at the end, as shown in Figure 2.

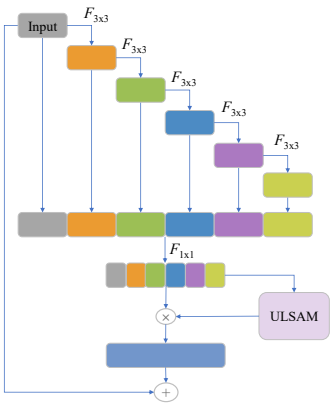

Figure 3: *The architecture of OSA. $F_{1\times1}$, $F_{3\times3}$ denote 1×1, 3×3 Conv layer respectively. The symbol ⊗ indicates element-wise multiplication and ⊕ denotes element-wise addition.*

2.3.1. OSA

Figure 3 depicts the specific structure of the OSA module. Every OSA module is comprised of 5 consecutive convolutional layers with the same input/output channels. Each convolutional layer is connected in two ways, one way to subsequent layers, and the other way to the final feature map. In other words, one-shot aggregation (OSA) aggregates all intermediate features in the last layer at once. This aggregation strategy circumvents feature redundancy while preserving the advantages of residual connection. Unlike ResNet [25], VovNet aggregates the features from shallower by concatenation rather than a summation. As mentioned in [29], the information carried by the earlier feature maps will be washed out when added with other feature maps. In contrast, by concatenation, the information will remain in its original form. As a result, VovNet can efficiently capture diverse receptive fields and is superior to ResNet in terms of accuracy.

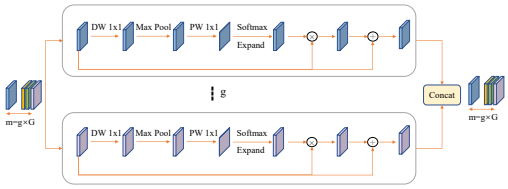

Figure 4: *The architecture of ULSAM. It divides the input feature maps into g groups where each group contains G feature maps. DW 1×1 is depthwise convolution with 1×1 kernel, PW 1×1 is pointwise convolution with only one filter. ⊗ denotes element-wise multiplication, ⊕ implies element-wise addition, and the final outputs are concatenated by "Concat".*

2.3.2. ULSAM

Vovnetv2 explicitly simulates the interdependency between the channels of the feature map through eSE to enhance its representation ability. The eSE employs a global average pooling for each channel independently to squeeze spatial dependency and then uses one fully connected (FC) layer to generate channel weights. The FC layer aims at dimensionality reduction as well as capturing nonlinear cross-channel interactions. However, dimensionality reduction inevitably leads to information loss and also brings side effects to channel attention prediction. Moreover, capturing dependency among all channels is unnecessary. For further boosting the performance of VoVNetV2, we propose a novel channel attention module, ultra-lightweight subspace attention mechanism (ULSAM).

ULSAM divides the input feature maps into g mutually exclusive groups, i. e., g feature subspaces, and each group contains G feature maps. ULSAM uses deep convolution in the initial step and 1×1 filter convolution in the subsequent steps. This approach increases the effective receptive size of the network and achieves multi-scale feature representation. Each feature subspace learns a different attention map. Different weights are assigned in different attention graphs to learn different importance, which is suitable for generating multi-frequency features. In contrary to eSE, ULSAM uses linear relationships in different feature map subspaces to integrate cross-channel information and capture complex cross-channel information interactions. Overall, ULSAM enables multi-scale, multi-frequency feature representation and better modeling of relationships between channels, making it ideal for fine-grained classification tasks such as speaker recognition. Figure 4 illustrates the specific structure of ULSAM.

3. Experiment

3.1. Dataset

We conduct training end-to-end on the VoxCeleb1 [4] dataset (only on the 'dev' partition, this contains speech from 1,211 speakers). The trained model is evaluated on the VoxCeleb1 test set with 40 unseen speakers.

Notably, the speakers in the VoxCeleb1 dataset are from different countries, with diverse ages, occupations, and accents. The data are completely real English speech with some real noise, not artificial white noise. The noise occurs at irregular points in time and the human voices are both large and small. The noise includes background human voices, laughter, echoes, room noise, and recording equipment noise.

3.2. Experimental setup

Implementation details. Our implementation is based on the PyTorch framework [30] and trained on NVIDIA 1080Ti with a batch size of 128. The network is optimized with Stochastic Gradient Descent (SGD) using an initial learning rate of 0.1. The learning rate decays every 30 epochs with a decay rate of 10%. The training is stopped after 100 epochs.

Baseline models. We train two baseline models: 1) Fast ResNet [31], a 34-layer ResNet network but the input dimensions are smaller and the strides are earlier in order to reduce computational requirements; 2) VovNetV2, consisting of OSA modules and eSE attention mechanism. According to [31], the hyperparameters m = 0.1 and s = 30 are chosen for the model trained with AM-softmax loss. In addition, the final embedding dimension we obtain is 512.

Input representations. The input settings follow [31]. We use a fixed-length 2-second temporal segment randomly extracted from each utterance. The spectrograms are extracted using a Hamming window of width 25ms and step 10ms. Both the baselines and our modified VovNet use 40-dimensional Mel filterbanks as input. Mean and variance normalization (MVN) is applied to the network inputs by instance normalization [32].

No voice activity detection (VAD) or data enhancement is used in the training.

3.3. Results

In this section, we verify the superior feature extraction capability of our modified VovNet with ULSAM, and then compare the performance of speaker recognition systems incorporating our proposed regularization method. The results on the VoxCeleb1 test set are reported in Table 1. The recognition accuracy was measured by equal error rate (EER).

Table 1: *EER results on VoxCeleb1 test set. For VIB regularization,* $\beta = 0.001$.

Front-end model	VIB	Loss	EER(%)
i-vector/PLDA		–	8.8000
VGG-M		softmax	10.2000
Fast ResNet		softmax	8.5737
Fast ResNet	✓	softmax	7.0520
VovNetV2		softmax	8.0064
VovNetV2	✓	softmax	6.3468
VovNetV2		AM-softmax	6.8452
VovNetV2	✓	AM-softmax	6.5429
VovNet/ULSAM		softmax	7.5928
VovNet/ULSAM	✓	softmax	6.1559
VovNet/ULSAM		AM-softmax	6.1294
VovNet/ULSAM	✓	AM-softmax	**5.8643**

We first examine our results when no VIB regularization is applied. With standard softmax loss and TAP aggregation, VovNetV2 outperforms the ResNet-based model (EER of 8.0064% vs 8.5737%). By replacing the eSE with ULSAM, a further performance gain is achieved (7.5928% EER). This suggests that VovNetV2 is a more compact feature extraction network than ResNet and has stronger feature expression capabilities. More importantly, our choice of ULSAM as an alternative to eSE has significant implications for the neural network to extract multiscale features that better characterize the speaker's identity.

Secondly, we pay attention to the variational information bottleneck regularization. As we can see from the table, the models trained with the proposed VIB regularization strategy consistently outperform those trained without. On the Fast ResNet and VovNetV2 models, the EER obtained a 17.75% and 20.73% improvement compared to no VIB regularization, respectively (7.0520% vs 8.5737% and 6.3468% vs 8.0064%). Our modified VovNet that integrates with VIB regularization yields further enhancements, obtaining an EER of 6.1559%. It is worth noting that on the modified VovNet (i. e., OSA+ULSAM) incorporating the VIB method, our model achieves the best results of 5.8643% EER. These results indicate that the variational information bottleneck performs a good regularization. It enhances the constraints on the neural network, further filtering the speaker identity irrelevant information while retaining only the most discriminative representation.

Figure 5 depicts the change process of the mutual information $I(X;Z)$and $I(Z;Y)$ during training. The three models are trained using softmax loss, with TAP as the pooling layer. We observe that both $I(X;Z)$ and $I(Z;Y)$ show an increasing trend at the beginning, as the neural network keeps learning

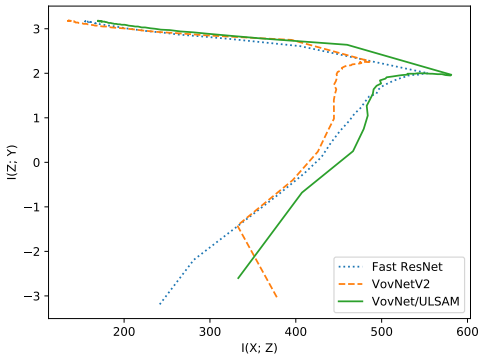

Figure 5: *Variation curves of the mutual information* $I(X;Z)$ *and* $I(Z;Y)$.

information from the input speech. Then, $I(Z;X)$ increases while $I(Z;Y)$ no longer changes, which means that the information learned by Z is not useful for predicting the identity of the speaker. For VovNet with ULSAM, the EER at this point is 7.5928%.The VIB forces the neural network to forget the information that is not relevant to the speaker's identity and preserve only the identity-related information. After that, as $I(X;Z)$ decreases, $I(Z;Y)$ increases, the model further eliminates the harmful information. As a result, the EER is further reduced to 6.1559%.

Table 2: *EER(%) results on different* β *values.*

Model	$\beta = 0.01$	$\beta = 0.001$	$\beta = 0.0001$
Fast ResNet	7.0732	7.0520	7.9109
VovNetV2	6.1771	6.3468	6.6119
VovNet/ULSAM	6.0286	6.1559	6.5854

The performance on various β is shown in Table 2. We train these three models with softmax loss and TAP layer. Despite small differences, we observe that a small β corresponds to a low penalty intensity, which will impair the performance because of insufficient noise removal.

4. Conclusions

In this paper, we introduce variational information bottlenecks as a regularization method for extracting robust deep speaker embeddings. The VIB forces the neural network to forget useless information and retain only the information related to the speaker identity. Also, we present an improved VovNetV2 framework, which enhances the feature representation by learning multi-scale and multi-frequency features. Our experiments demonstrate that the proposed feature extraction network and regularization method significantly help to improve speaker recognition accuracy.

5. Acknowledgements

This work was in part supported by the Major project of IoV, Technological Innovation Projects in Hubei Province (Grant No.2020AAA001, 2019AAA024) and Sanya Science and Education Innovation Park of Wuhan University of Technology (Grant No.2020KF0054).

6. References

[1] N. Dehak, P. J. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 788–798, 2011.

[2] S. Prince and J. Elder, "Probabilistic linear discriminant analysis for inferences about identity," in *2007 IEEE 11th International Conference on Computer Vision*, 2007, pp. 1–8.

[3] M. Mclaren, L. Ferrer, D. Castan, and A. Lawson, "The speakers in the wild (sitw) speaker recognition database," in *Interspeech 2016*, 2016.

[4] A. Nagrani, J. S. Chung, and A. Zisserman, "Voxceleb: A large-scale speaker identification dataset." in *Interspeech 2017*, 2017, pp. 2616–2620.

[5] J. S. Chung, A. Nagrani, and A. Zisserman, "Voxceleb2: Deep speaker recognition." in *Interspeech 2018*, 2018, pp. 1086–1090.

[6] E. Variani, X. Lei, E. Mcdermott, I. L. Moreno, and J. Gonzalez-Dominguez, "Deep neural networks for small footprint text-dependent speaker verification," in *IEEE International Conference on Acoustics*, 2014.

[7] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-vectors: Robust dnn embeddings for speaker recognition," in *ICASSP 2018 - 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018.

[8] K. Okabe, T. Koshinaka, and K. Shinoda, "Attentive statistics pooling for deep speaker embedding," in *Interspeech 2018*, 2018.

[9] Y. Tang, G. Ding, J. Huang, X. He, and B. Zhou, "Deep speaker embedding learning with multi-level pooling for text-independent speaker verification," in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019.

[10] C. Li, X. Ma, B. Jiang, X. Li, X. Zhang, X. Liu, Y. Cao, A. Kannan, and Z. Zhu, "Deep speaker: an end-to-end neural speaker embedding system," *arXiv*, 2017.

[11] W. Cai, J. Chen, and M. Li, "Exploring the encoding layer and loss function in end-to-end speaker and language recognition system," in *Odyssey 2018*, 2018.

[12] Z. Gao, Y. Song, I. Mcloughlin, P. Li, and L. R. Dai, "Improving aggregation and loss function for better embedding learning in end-to-end speaker verification system," in *Interspeech 2019*, 2019.

[13] S. Seo, D. J. Rim, M. Lim, D. Lee, and J. H. Kim, "Shortcut connections based deep speaker embeddings for end-to-end speaker verification system," in *Interspeech 2019*, 2019.

[14] D. Snyder, D. Garcia-Romero, D. Povey, and S. Khudanpur, "Deep neural network embeddings for text-independent speaker verification," in *Interspeech 2017*, 2017.

[15] S. Novoselov, A. Shulipa, I. Kremnev, A. Kozlov, and V. Shchemelinin, "On deep speaker embeddings for text-independent speaker recognition," in *Odyssey 2018 The Speaker and Language Recognition Workshop*, 2018.

[16] Y. Zhu, T. Ko, D. Snyder, B. Mak, and D. Povey, "Self-attentive speaker embeddings for text-independent speaker verification," in *Interspeech 2018*, 2018.

[17] M. Ravanelli and Y. Bengio, "Speaker recognition from raw waveform with sincnet," in *2018 IEEE Spoken Language Technology Workshop (SLT)*, 2018, pp. 1021–1028.

[18] ——, "Speaker recognition from raw waveform with sincnet," in *2018 IEEE Spoken Language Technology Workshop (SLT)*, 2018, pp. 1021–1028.

[19] W. Liu, Y. Wen, Z. Yu, M. Li, B. Raj, and L. Song, "Sphereface: Deep hypersphere embedding for face recognition," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 6738–6746.

[20] F. Wang, J. Cheng, W. Liu, and H. Liu, "Additive margin softmax for face verification," *IEEE Signal Processing Letters*, vol. 25, no. 7, pp. 926–930, 2018.

[21] H. Wang, Y. Wang, Z. Zhou, X. Ji, D. Gong, J. Zhou, Z. Li, and W. Liu, "Cosface: Large margin cosine loss for deep face recognition," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 5265–5274.

[22] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "Arcface: Additive angular margin loss for deep face recognition," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 4690–4699.

[23] J. S. Chung, J. Huh, and S. Mun, "Delving into voxceleb: Environment invariant speaker recognition," in *Odyssey 2020 The Speaker and Language Recognition Workshop*, 2020.

[24] Y. Lee and J. Park, "Centermask: Real-time anchor-free instance segmentation," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 13 906–13 915.

[25] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.

[26] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 2261–2269.

[27] R. Saini, N. K. Jha, B. Das, S. Mittal, and C. K. Mohan, "Ulsam: Ultra-lightweight subspace attention module for compact convolutional neural networks," in *2020 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2020, pp. 1627–1636.

[28] N. Tishby, F. C. Pereira, and W. Bialek, "The information bottleneck method," *University of Illinois*, vol. 411, no. 29-30, pp. 368–377, 2000.

[29] L. Zhu, R. Deng, M. Maire, Z. Deng, G. Mori, and P. Tan, "Sparsely aggregated convolutional networks," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 186–201.

[30] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala, "Pytorch: An imperative style, high-performance deep learning library," in *Advances in Neural Information Processing Systems*, vol. 32, 2019, pp. 8026–8037.

[31] J. S. Chung, J. Huh, S. Mun, M. Lee, H. S. Heo, S. Choe, C. Ham, S. Jung, B.-J. Lee, and I. Han, "In defence of metric learning for speaker recognition." in *Interspeech 2020*, 2020, pp. 2977–2981.

[32] D. Ulyanov, A. Vedaldi, and V. S. Lempitsky, "Instance normalization: The missing ingredient for fast stylization." *arXiv preprint arXiv:1607.08022*, 2016.