Disentangled Representations for Arabic Dialect Identification based on Supervised Clustering with Triplet Loss

Zainab Alhakeem, Yoohwan Kwon and Hong-Goo Kang
DSP&AI Lab., Department of Electrical and Electronic Engineering
Yonsei University, Seoul, Korea
{zrkhakim, yhkwon}@dsp.yonsei.ac.kr, hgkang@yonsei.ac.kr

***Abstract*—In this paper, we propose a novel supervised clustering with triplet (SCT) loss that effectively learns disentangled representations for Arabic dialect identification (ADI). To improve the performance of ADI using latent representation based approaches, we need to extract embeddings that include only dialect related information by dissociating all the irrelevant information such as gender, channel, and speaker. In consideration of the embedding-level distribution, our proposed SCT loss minimizes intra-class variations and maximizes inter-class variations. Specifically, it uses the centroid of each dialect as a triplet component, thereby avoiding the issue of choosing an undesirable triplet component due to random sampling. Experimental results on the ADI-17 dataset show that our proposed method significantly outperforms conventional state-of-the-art methods in terms of the identification accuracy.**

Index Terms: Arabic Dialect Identification, Disentangled Representation, Supervised Clustering, Triplet Loss

I. Introduction

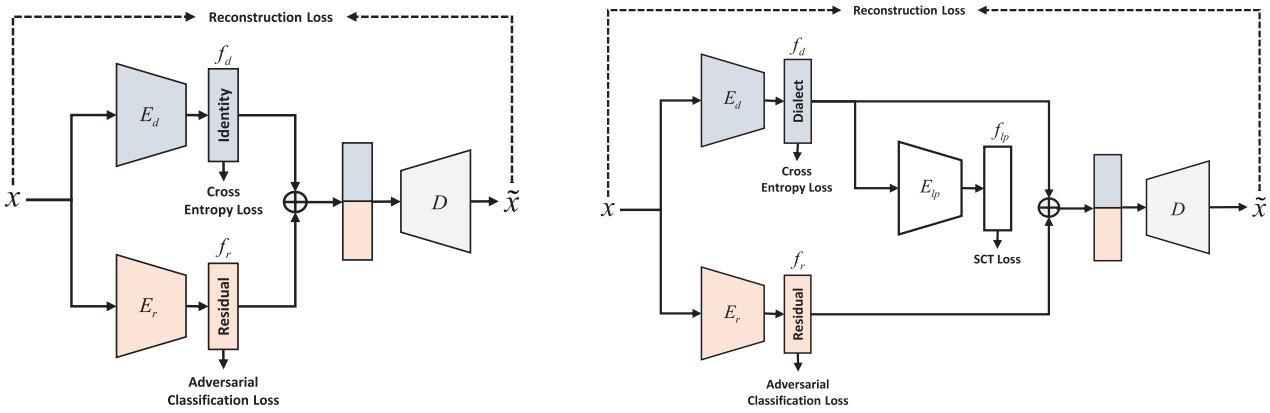
Arabic is the mother tongue of more than 350 million people in 22 countries [1]. Therefore, Arabic is frequently spoken in dialects rather than a standard form, and more than 30 different Arabic dialects exist [2]. Since Arabic dialects vary significantly both in semantics and phonetics [3], it is not easy to implement an automatic speech recognition system that can be applicable to all dialects. As a result, Arabic speech assistants are relatively under-resourced; for example, Google assistant only supports some services for modern standard Arabic and a few dialects (e.g., Egyptian and Saudi), and Alexa does not support Arabic at all [4], [5]. However, there is potential to further extend the capabilities of Arabic speech assistants if information on dialects is utilized.

In order to aid research in Arabic dialect identification (ADI), the fifth Multi-Genre Broadcast (MGB-5) challenge recently included the task of Arabic dialect identification for 17 different dialects (ADI-17) using about 3,000 hours of data that was collected from YouTube [6]. In [7], several models for the ADI task were investigated, such as ones based on i-vectors, x-vectors, end-to-end x-vectors with a softmax output layer, and an end-to-end CNN-based model with a global statistic pooling layer. The latter model was trained with different losses using softmax, tuplemax and additive margin softmax, and was used as a baseline in MGB-5 for the ADI task [6]. A transformer network model [8] was recently developed based on the self-attention technique to capture long range dependencies.

However, in [9], it was experimentally shown that representations from the baseline model in the ADI challenge contained not only dialect information, but also non-dialect information such as gender, channel, and speaker related representations. The presence of such superfluous information can cause difficulties in training that hurt a model's performance. In order to address this issue, an effective candidate approach is disentanglement representation learning, which has been explored thoroughly in computer vision and speech domains for dissociating target-specific feature representations from unrelated ones [10], [11], [12], [13], [14]. In [13], a disentanglement representation learning strategy was adopted for speaker identification to obtain speaker-related information using adversarial training with an autoencoder architecture. To effectively obtain speaker-specific embeddings by minimizing intra-class variation and mutual information with residual embeddings, [14] proposed identity change loss and mutual information loss criteria. However, they did not explicitly address any methods to maximize inter-class variations.

Triplet loss [15] is a suitable alternative for concurrently minimizing intra-class variations and maximizing inter-class variations. When we design triplet loss, it is crucial to select positive and negative inputs in dictionary pools [16]. Since finding all possible triplet combinations are computationally impractical, most existing methods utilize a random selection approach. However, this method is often not an optimal choice because there is no way to tell whether the selected examples are effective for training or not [16]. One selection method that aims to solve this problem is to design a class-dependent handcrafted dictionary pool, but this process is time-consuming [16], [17].

Clustering is a popular approach for vector quantization to group a large set of training samples and represent them as their centroids [18]. k-means clustering is one of the most famous clustering methods [19], which minimizes the square of intra-cluster sums. Interestingly, the minimization criterion is equivalent to the maximization of the square of inter-cluster sums. However, such clustering methods have to evaluate considerable pairs of training samples to successfully achieve clusters of high quality [19].



(a) The disentanglement model (b) The proposed model

Fig. 1. Overviews of the disentanglement model and the proposed model.

In this paper, we propose a method that incorporates a clustering-based contrastive loss to disentangle dialect-related representations learned for the ADI task. The specific contributions of our work are as follows: 1) we establish a connection between k-means clustering and the triplet loss; 2) we propose a novel supervised clustering with triplet (SCT) loss to concurrently minimize intra-class variations and maximize inter-class variations; 3) we obtain cluster means in a supervised manner and use them as centroids in order to avoid the construction of the dictionary pools of the positive and negative inputs in the triplet loss; 4) we demonstrate the effectiveness of the proposed SCT loss with the disentanglement network model for obtaining discriminative dialect embeddings from the ADI-17 data set.

II. Related Works

A. Disentangled representation

Representation disentanglement is a technique to estimate latent space embeddings by dissociating target-related characteristics from other unneeded ones so that the estimated target embeddings are more effective for downstream tasks [10]. Adversarial training with an autoencoder architecture is a popular method to achieve such dissociation goals [11], [12].

In [13] a disentanglement model for speaker recognition task was proposed. As shown in Fig. 1(a), two encoders, E_d and E_r, are utilized to embed speaker identity related information f_d and residual information f_r, respectively. The two output embeddings f_d and f_r are then concatenated and fed into the decoder, which performs reconstruction. The reconstructed output of the decoder and the original input are then evaluated. The disentanglement network model is trained with the following loss functions.

1) Identity classification loss: The target information f_d is obtained by the identity encoder E_d, which is trained using the following cross-entropy loss:

$$\mathcal{L}_{\text{d}} = -\sum_{i=1}^{C} y^i \log\left(\text{softmax}\left(f_d^i\right)\right), \tag{1}$$

where C is the total number of classes and y^i is the class label.

2) Adversarial classification loss: The residual information f_r is given by the adversarial encoder E_r, which is trained using a uniform distribution $\frac{1}{C}$ without any labels as follows:

$$\mathcal{L}_{\text{adv}} = \frac{1}{C}\sum_{i=1}^{C} \log\left(\text{softmax}\left(f_r^i\right)\right). \tag{2}$$

3) Reconstruction loss: Since the combination of the identity embedding f_d and the residual embedding f_r involves the original input information, the concatenated embedding of f_d and f_r is fed into the decoder D, which is trained using an L_2-norm distance between the reconstructed information and the original input information x as follows:

$$\mathcal{L}_{\text{rec}} = \frac{1}{2}\left\|D\left(f_d \oplus f_r\right) - x\right\|_2^2, \tag{3}$$

where $\oplus$ is the concatenation operation.

B. Triplet loss

Triplet loss has been extensively and successfully employed for many applications (e.g., face and speaker identification) [15], [20]. It uses three different training inputs (x_a, x_p, x_n), where x_a is an anchor, x_p is a positive input from the anchor class, and x_n is a negative input from a different class. The triplet loss naturally pushes the anchor close to the positive input, but far from the negative input as follows:

$$\mathcal{L}_{\text{T}} = \max\left(\left\|E\left(x_a\right) - E\left(x_p\right)\right\|_2^2 - \left\|E\left(x_a\right) - E\left(x_n\right)\right\|_2^2 + m, 0\right), \tag{4}$$

where $E\left(\cdot\right)$ represents the feature embedding function, and m indicates a margin between positive and negative pairs.

C. k-means clustering

Clustering has been extensively studied as a tool for unsupervised learning [21], [22]. k-means clustering [19] is one of the most well-known and popular algorithms, and its loss function is defined as:

$$\mathcal{L}_{\text{km}} = \sum_{i=1}^{k}\sum_{x\in S^i}\left\|x - \mu^i\right\|_2^2, \tag{5}$$

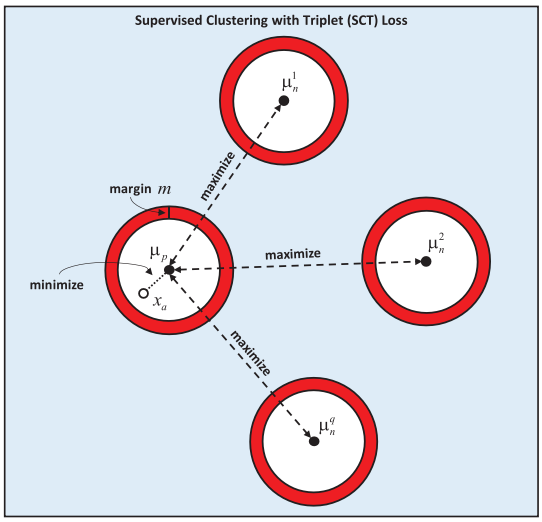


Fig. 2. The proposed supervised clustering with triplet (SCT) loss.

which aims to partition the N observations $\{x^1, x^2, \ldots, x^N\}$ into sets $S = \{S^1, S^2, \ldots, S^i, \ldots, S^k\}$. μ^i is the mean of each class in S^i. In the assignment step of k-means clustering, each observation is assigned to the nearest mean as follows:

$$S^i = \left\{ x_a : \left\| x_a - \mu^i \right\|_2^2 \leq \left\| x_a - \mu^j \right\|_2^2 \forall j, 1 \leq j \leq k \right\}, \quad (6)$$

where x_a is an observation which is assigned to S^i, and each mean at time t is updated as follows:

$$\mu_t^i = \frac{1}{\left| S_{t-1}^i \right|} \sum_{x^j \in S_{t-1}^i} x^j. \quad (7)$$

III. Proposed Method

A. Overview

In this section, we present a novel supervised clustering with triplet (SCT) loss to minimize intra-class variations and maximize inter-class variations in a disentanglement model (see Fig. 1(b)). The proposed SCT loss is illustrated in Fig. 2, where the distance between the anchor x_a and a positive class mean μ_p is encouraged to be minimized, while the distances between it and negative class means μ_n are forced to be maximized.

Similar to the disentanglement model in Fig. 1(a), the two encoders E_d and E_r are used to embed the dialect related information f_d and the residual information f_r, respectively. The two embeddings f_d and f_r are then concatenated and fed into the decoder D to reconstruct the original input information. In our proposed disentanglement model, we compute the SCT loss on the output of the dialect encoder E_d to address the dialect-related information. The dialect embedding f_d is fed through a linear projection layer E_{lp} for dimensionality reduction of the high dimensional data, resulting in a lower dimension embedding f_{lp}.

B. Supervised clustering with triplet (SCT) loss

Given a set of observations in a mini-batch $\{x^i, y^i\}_{i=1}^B$, where x^i is a training sample, y^i is its corresponding label, and B is the batch size, our supervised clustering with triplet (SCT) loss is as follows:

$$\mathcal{L}_{\text{SCT}} = \max\left(\left\| f_{lp,a}^i - \mu_p \right\|_2^2 - \frac{1}{|S_n|} \sum_{\mu_n \in S_n} \left\| f_{lp,a}^i - \mu_n \right\|_2^2 + m, 0 \right), \quad (8)$$

Here, $f_{lp,a}^i = E_{lp}\left(E_d\left(x_a^i\right)\right)$ indicates the representation for the anchor input x_a^i, μ_p is the mean of the positive class, and μ_n is the mean of the negative class. We define a subset S_n of top-q negative class means as follows:

$$S_n = \left\{ \mu_n^j : closest\left(f_{lp,a}^i, \mu_n^j, q\right), 1 \leq j \leq C, j \neq p \right\}, \quad (9)$$

which includes the top-q closest negative classes to the anchor input based on the L_2-norm distances. The purpose of selecting the top-q negative classes is to concentrate on maximizing the inter-class variations near the anchor class. This selection process offers partial compensation for any noisy labels in the maximization problem of the inter-class variations.

In order to avoid the recalculation of each mean in equation (7) and the memory allocation of all the training samples, the corresponding class mean of the anchor class is incrementally updated every batch as follows:

$$\mu_t^i = \lambda \mu_{t-1}^i + \frac{f_{lp,a}^i - \mu_{t-1}^i}{z_t^i}, \quad (10)$$

where λ is a forgetting factor and $z_t^i = z_{t-1}^i + 1$ is the number of averaged samples at time t. Since the clustering means for the positive and negative inputs in the triplet loss are assigned by considering the class labels, this achieves a supervised approach to clustering.

1) Relationship between k-means clustering and triplet loss: In equation (6), k-means clustering assigns the target sample x_a to the corresponding clustering set when only the condition $\left\| x_a - \mu^i \right\|_2^2 - \left\| x_a - \mu^j \right\|_2^2 \leq 0$ is satisfied. However, contrary to the acceptance condition, the rejection condition is given by: $\left\| x_a - \mu^i \right\|_2^2 - \left\| x_a - \mu^j \right\|_2^2 > 0$, which can be connected to the triplet loss function (see equation (4)) as follows:

$$\mathcal{L}_{\text{CT}} = \max\left(\left\| E\left(x_a\right) - E\left(\mu^i\right) \right\|_2^2 - \left\| E\left(x_a\right) - E\left(\mu^j\right) \right\|_2^2 + m, 0 \right). \quad (11)$$

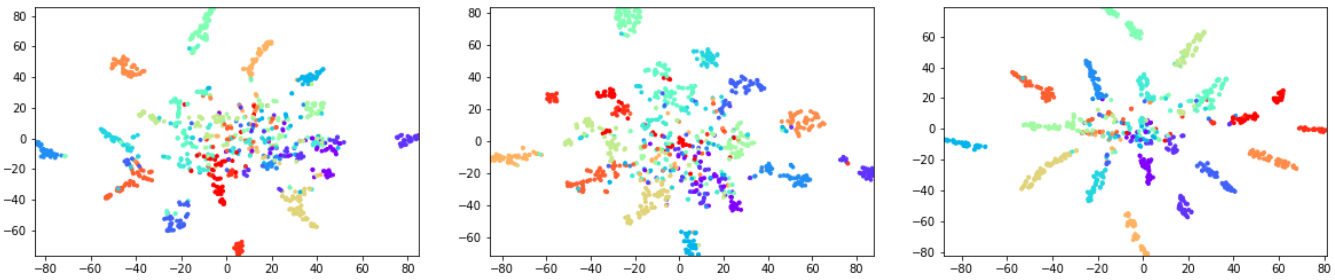
As such, equation (11) establishes a relationship between k-means clustering and the triplet (CT) loss.

C. Overall objective function

The overall objective function of our proposed network model is given by a weighted sum of 1) the dialect identification loss $\mathcal{L}_{\text{d}}$, 2) the adversarial loss $\mathcal{L}_{\text{adv}}$, 3) the reconstruction loss $\mathcal{L}_{\text{rec}}$ from the disentanglement network, and 4) the proposed SCT loss $\mathcal{L}_{\text{SCT}}$:

$$\mathcal{L}_{\text{total}} = \alpha_{\text{d}}\mathcal{L}_{\text{d}} + \alpha_{\text{adv}}\mathcal{L}_{\text{adv}} + \alpha_{\text{rec}}\mathcal{L}_{\text{rec}} + \alpha_{\text{SCT}}\mathcal{L}_{\text{SCT}}, \quad (12)$$

where α_{d}, α_{adv}, α_{rec} and α_{SCT} are adjustable hyper-parameters.



(a) The disentanglement model (b) The disentanglement model with IC (c) The disentanglement model with the proposed SCT

Fig. 3. The t-SNE plots of extracted embeddings from each model such as (a) the disentanglement model [13], (b) the disentanglement model with the IC loss [14] and (c) the disentanglement model with the proposed SCT loss.

TABLE I
PERFORMANCE RESULTS OF THE DISENTANGLEMENT MODEL WITH AND WITHOUT THE PROPOSED SCT LOSS USING ADI-17 DATA SET

SCT Loss	**Dev. Set Accuracy (%)**				**Test Set Accuracy (%)**			
	(< 5s)	($5 \sim 20$s)	(> 20s)	**Overall**	(< 5s)	($5 \sim 20$s)	(> 20s)	**Overall**
without	71.83	79.13	87.85	79.20	73.15	80.89	86.31	78.30
with	84.67	90.51	95.09	88.83	84.86	90.22	94.35	88.45

TABLE II
PERFORMANCE COMPARISON OF ACCURACIES FOR ADI STATE-OF-THE ARTS

Method	**Dev. Set Accuracy (%)**	**Test Set Accuracy (%)**
i-vector [7]	59.7	60.3
x-vector [7]	71.0	72.1
E2E (x-vector) [7]	76.6	77.8
E2E (Softmax) [7]	83.0	82.0
E2E (Tulemax) [7]	78.6	78.6
E2E (AM-Softmax) [7]	62.5	63.7
Transformer [8]	83.2	82.5
Disentanglement [13]	79.2	78.3
Disentanglement + IC [14]	80.1	80.5
Disentanglement + Proposed SCT	88.8	88.5

IV. EXPERIMENTS

A. Data sets

In our experiments, we use the ADI-17 dataset from the MGB-5 challenge [6], [7], which consists of Arabic speech data obtained from YouTube. The training set contains 3,033.4 hours of speech, the development set includes 24.9 hours, and the test set has 33.1 hours. Each utterance also falls under one of three duration ranges: short (< 5s), medium ($5 \sim 20$s) and long duration (> 20s).

B. Experimental settings

According to the disentanglement model in [13], the network architectures of the two encoders E_d and E_r are based on ResNet-34 [23] with a global temporal pooling (TAP) layer to fix the length of input features. The decoder architecture is followed by three fully connected layers and ten transposed convolutional layers [24]. The linear projection network E_{lp} includes three layers of sizes 512×256, 256×128 and 128×64. The training batch size is set to 17, where one sample from each class is randomly selected. The input utterances are randomly segmented into 5 second chunks. A 25ms window is utilized with a 10ms hop size, and mel-spectrograms of size 320×257 are constructed with an FFT of size 512. The Adam optimiser [25] is used for model training, with initial learning rate 10^{-4} and learning rate decay of 10% every 10 epochs. In the SCT loss, the margin m is fixed at 1, and the forgetting factor λ is set at 0.99. Each weight for the overall objective function in equation (12) is empirically set as follows: $\alpha_{\mathrm{d}} = 1$, $\alpha_{\mathrm{adv}} = 0.1$, $\alpha_{\mathrm{rec}} = 0.1$ and $\alpha_{\mathrm{SCT}} = 0.1$. We empirically fixed the top-q number of the negative class means with $q = 5$.

For performance evaluation, we implement and compare the original disentanglement model [13], the disentanglement model with an identity change (IC) loss [14], and the disentanglement model with SCT loss. Following [6], [7], the classification accuracies for the unseen test data set are recorded.

C. Results

In order to observe the effectiveness of our proposed method, Table I shows a performance comparison of the disentanglement model with and without SCT loss using the development and test data sets for the short, medium, and long utterances, as well as for all of the utterances. Using SCT loss improved the performance of the original model for utterances of every duration type.

Fig. 3 shows t-SNE plots of the extracted embeddings from the disentanglement model, the model with the IC loss, and the one with the proposed SCT loss. For the t-SNE plots, we used 60 random samples per class from the test set. It can be qualitatively observed that incorporating the proposed

SCT loss into the model visually provides more discriminative embeddings than the baselines. Additionally, the inter-class distances for the two baselines (Fig. 3(a) and (b)) are visibly closer together compared to those in the model using SCT loss (Fig. 3(c)).

Table II summarizes the state-of-the-art performances obtained from the literature [7], [8]. The disentanglement model with SCT loss outperformed all other state-of-the-art methods, achiving 88.83% and 88.45% accuracy on the development and test sets, respectively.

V. Conclusion

This paper presented a novel supervised clustering with triplet (SCT) loss for a representation disentanglement model. Specifically, we applied the proposed SCT loss for minimizing intra-class variations and maximizing the inter-class variations in an Arabic dialect identification task. We qualitatively showed that the SCT loss allows for the extraction of more discriminative feature representations than baseline methods using t-SNE embedding visualizations. Our proposed disentanglement model outperformed the state-of-the-art methods in terms of classification accuracies on the ADI-17 dataset.

Acknowledgment

This work was supported in part by Artificial Intelligence Graduate School Program under Grant 2020-0-01361.

References

[1] J. Owens, *The Oxford handbook of Arabic linguistics*. Oxford University Press, 2013.

[2] K. Versteegh, *Arabic language*. Edinburgh University Press, 2014.

[3] E. Benmamoun and R. Bassiouney, *The Routledge handbook of Arabic linguistics*. Routledge, 2017.

[4] Google, "Speech to text language support," url update: 2021-02-01. [Online]. Available: https://cloud.google.com/speech-to-text/docs/languages

[5] Amazon, "List of alexa interfaces and supported languages," accessed: 2021-02-18. [Online]. Available: https://developer.amazon.com/en-US/docs/alexa/device-apis/list-of-interfaces.html

[6] A. Ali, S. Shon, Y. Samih, H. Mubarak, A. Abdelali, J. Glass, S. Renals, and K. Choukri, "The mgb-5 challenge: Recognition and dialect identification of dialectal arabic speech," in *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2019, pp. 1026–1033.

[7] S. Shon, A. Ali, Y. Samih, H. Mubarak, and J. Glass, "Adi17: A fine-grained arabic dialect identification dataset," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 8244–8248.

[8] W. Lin, M. Madhavi, R. K. Das, and H. Li, "Transformer-based arabic dialect identification," in *2020 International Conference on Asian Language Processing (IALP)*. IEEE, 2020, pp. 192–196.

[9] S. A. Chowdhury, A. Ali, S. Shon, and J. Glass, "What does an end-to-end dialect identification model learn about non-dialectal information?" *Proc. Interspeech 2020*, pp. 462–466, 2020.

[10] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives," *IEEE transactions on pattern analysis and machine intelligence*, vol. 35, no. 8, pp. 1798–1828, 2013.

[11] X. Chen, Y. Duan, R. Houthooft, J. Schulman, I. Sutskever, and P. Abbeel, "Infogan: Interpretable representation learning by information maximizing generative adversarial nets," *arXiv preprint arXiv:1606.03657*, 2016.

[12] J. Zhou, T. Jiang, L. Li, Q. Hong, Z. Wang, and B. Xia, "Training multi-task adversarial network for extracting noise-robust speaker embedding," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 6196–6200.

[13] J. Tai, X. Jia, Q. Huang, W. Zhang, H. Du, and S. Zhang, "Seef-aldr: A speaker embedding enhancement framework via adversarial learning based disentangled representation," in *Annual Computer Security Applications Conference*, 2020, pp. 939–950.

[14] Y. Kwon, S.-W. Chung, and H.-G. Kang, "Intra-class variation reduction of speaker representation in disentanglement framework," *Proc. Interspeech*, 2020.

[15] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 815–823.

[16] C.-Y. Wu, R. Manmatha, A. J. Smola, and P. Krahenbuhl, "Sampling matters in deep embedding learning," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 2840–2848.

[17] B. Harwood, V. Kumar BG, G. Carneiro, I. Reid, and T. Drummond, "Smart mining for deep metric learning," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 2821–2829.

[18] A. K. Jain, M. N. Murty, and P. J. Flynn, "Data clustering: a review," *ACM computing surveys (CSUR)*, vol. 31, no. 3, pp. 264–323, 1999.

[19] S. Lloyd, "Least squares quantization in pcm," *IEEE transactions on information theory*, vol. 28, no. 2, pp. 129–137, 1982.

[20] H. Bredin, "Tristounet: triplet loss for speaker turn embedding," in *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2017, pp. 5430–5434.

[21] A. M. Mustafa, G. Ayoade, K. Al-Naami, L. Khan, K. W. Hamlen, B. Thuraisingham, and F. Araujo, "Unsupervised deep embedding for novel class detection over data stream," in *2017 IEEE International Conference on Big Data (Big Data)*. IEEE, 2017, pp. 1830–1839.

[22] X. Ji, J. F. Henriques, and A. Vedaldi, "Invariant information clustering for unsupervised image classification and segmentation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 9865–9874.

[23] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[24] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," *arXiv preprint arXiv:1511.06434*, 2015.

[25] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *International Conference on Learning Representations*, 2015.