

Learning Speaker Embedding with Momentum Contrast

Ke Ding, Xuanji He, Guanglu Wan

Meituan-Dianping Group

{dingke02, hexuanji, wanguanglu}@meituan.com

Abstract

Speaker verification can be formulated as a representation learning task, where speaker-discriminative embeddings are extracted from utterances of variable lengths. Momentum Contrast (MoCo) is a recently proposed unsupervised representation learning framework, and has shown its effectiveness for learning good feature representation for downstream vision tasks. In this work, we apply MoCo to learn speaker embedding from speech segments. We explore MoCo for both unsupervised learning and pretraining settings. In the unsupervised scenario, embedding is learned by MoCo from audio data without using any speaker specific information. On a large scale dataset with 2, 500 speakers, MoCo can achieve EER 4.275% trained unsupervisedly, and the EER can decrease further to 3.58% if extra unlabelled data are used. In the pretraining scenario, encoder trained by MoCo is used to initialize the downstream supervised training. With finetuning on the MoCo trained model, the equal error rate (EER) reduces 13.7% relative (1.44% to 1.242%) compared to a carefully tuned baseline training from scratch. Comparative study confirms the effectiveness of MoCo learning good speaker embedding.

Index Terms: speaker verification, representation learning, unsupervised learning

1. Introduction

Speaker verification (SV) is the task of confirming the claimed identity of a speaker given one’s speech segments. Typically, a fixed-dimensional embedding is extracted for both enrollment and test speech, and compared to give out a same-speaker-or-not decision.

I-vectors [1] are widely used as speaker embeddings. In the standard i-vector system, a GMM is trained with all training data served as universal background model (UBM) to collect sufficient statistics (typically a super vector) from speech segments. A low-rank project matrix, dubbed Total Variability Matrix, is trained with the sufficient statistics. For a given segment, a super vector is extracted by UBM, and projected by the learned project matrix, resulting in a fixed-dimensional vector, i.e. i-vector. The similarity between two i-vectors (usually the log likelihood ratio) can be computed by probabilistic linear discriminant analysis (PLDA) [2].

In recent years, various neural network based methods are proposed to learning more discriminative embeddings.

X-vectors are proposed to replace i-vectors [3]. In the x-vector system, speaker embeddings are extracted by time-delay neural networks (TDNNs). The TDNN is trained as a multi-speaker classifier using cross entropy loss and then the activation from some hidden layer is extracted as the embedding. A temporal statistical pooling layer to applied to tackle segments of variable length. After training, utterances are mapped directly to fixed-dimensional speaker embeddings, just as i-vector systems do. Working with a PLDA backend, the x-vectors can

outperform i-vectors for short speech segments and are competitive on long duration test conditions [3, 4].

The embeddings learned by cross entropy loss is separable by design for the close-set classification task, but not necessarily discriminative enough, which is key to generalize to identify unseen speaker segments. To tackle this issue, various training criterion to enhance the discriminative power of the x-vector embeddings.

Contrastive learning, i.e. contrastive loss [5], triplet loss [6], is introduced for optimizing the distances between embeddings directly. Losses of contrastive style minimize the distance between an anchor and a positive sample while maximizes the distance between the anchor and a negative one, thus encouraging embeddings with compact within-speaker variations and separable between-speaker differences.

However, optimizing contrastive loss can be challenging [6, 7], and selecting training pairs or triplets suffers from combinatorial data expansion and negative samples mining is necessary for effective and stable training. Margin based training criterion is proposed to learn embeddings with compact inter-class variation and bypass the training problems of contrastive learning.

Center loss [8] is used to work with cross entropy loss (to avoid embedding collapsing). Center loss penalises the distance between the embeddings and their corresponding class centers in the Euclidean space to achieve intra-class compactness. By introducing margins between classes into conventional cross entropy loss, angular softmax (A-softmax) is reported being able to learn more discriminative embeddings than cross entropy loss and triplet loss [9]. More recently, Additive Angular Margin (AAM) loss is proposed for extracting highly discriminative features for face recognition [10]. AAM is successfully applied to speaker verification task [11] and achieves state of the art performance in the VoxCeleb challenge [12].

Recently, a novel unsupervised learning framework, Momentum Contrast (MoCo) [13] is proposed. MoCo is an extension of instance discrimination, and can learn representation using contrastive learning criterion. In several vision tasks, MoCo can outperform its supervised pretraining counterpart, thus largely closing the gap between unsupervised and supervised representation.

In this work, we apply MoCo to the speaker verification task. Observe that MoCo trains an encoder directly, which is exactly the conventional neural network based methods (e.g. x-vector system) do. What’s more, MoCo encourages discriminative representation (via contrastive learning), which is key for open set verification. We explore MoCo in the ways: 1) using MoCo trained encoder directly, 2) using MoCo as a pretraining method to relieve downstream training of interest.

The remainder of this paper is organized as follows. Section 2 describes our proposal for using MoCo for learning speaker embedding. Specifically, a SpecAugment [14] based data augmentation is introduced for parallel speech segments generation. Section 3 shows the experiments conducted on a

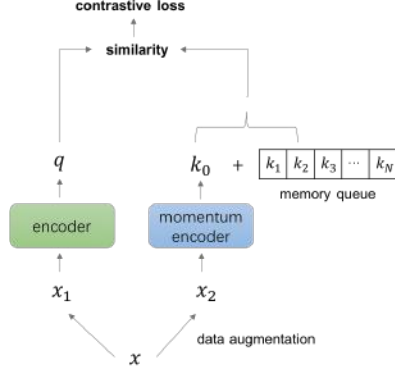


Figure 1: Training framework of MoCo

large scale speaker dataset. Results on the public available Vox-celeb dataset are also presented. We conclude our work in Section 4.

2. Proposed Method

2.1. Momentum Contrast Learning

Momentum Contrast (MoCo) [13] is a general mechanism for unsupervised learning representation using contrastive loss. In MoCo, a dynamic dictionary is built with a queue and a moving-averaged encoder. This enables building a large and consistent dictionary on-the-fly that facilitates contrastive unsupervised learning. The representations learned by MoCo is reported to transfer well to downstream visual tasks.

The training framework of MoCo is depicted by Fig. 1. For each training sample, two corrupted versions are generated by some augmentation strategy. After processing one mini-batch, the encoder’s parameters are updated by some optimizer, say SGD. The encoded representations of the current mini-batch (k_0 in Fig. 1) are enqueued, and the oldest ones are dequeued to keep the queue size consistent. Before processing the next mini-batch, the momentum encoder is update with the encoder with as momentum coefficient (typically close to 1).

As neural network based speaker verification systems learn embedding from utterances, MoCo can be applied to the training process in a natural way. To be specific, the encoder is as same as the network used for conventional xvector training (see Sec. 3). We just initialize the momentum encoder with the encoder and initialize memory queue randomly. For other details and import tricks for training (i.e. contrastive loss, ShuffleBN), we refer the reader to [13].

2.2. Data Augmentation

What’s specific to our task under study (speaker verification) is how to generate parallel corrupted version of the same utterance. The method we propose is depicted by Fig 2. First, we randomly selected two segments from the target utterance, which is the common practice in x-vector training. Second, we apply SpecAugment [14] to the segments, resulting in two different version of the same utterance with various length and spectrum distortion. Via the proposed process, parallel corrupted segments can be generated with both temporal and spectrum variability.

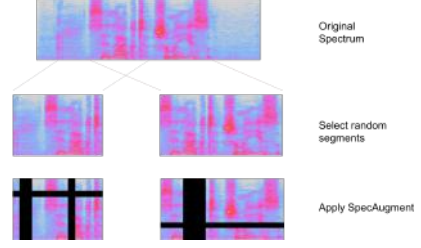


Figure 2: Generating two augmented version of segments from one utterance

2.3. MoCo as embedding extractor

MoCo trains the encoder as an instance discrimination task, treating the parallel corrupted version as positive sample and all in the memory queue as negative samples. So it’s natural to expect the embeddings extracted by a well trained encoder show discrimination between different speakers. Reminiscent of the conventional i-vector extractor, we can use the learned encoder with some backend (e.g. PLDA, cosine) for verification task directly. As MoCo need no speaker information, we also explore if extra unlabelled data can lead to more robust embedding for verification.

2.4. MoCo as Pretraining

Though unsupervised learning may not be optimal for the task of interest, features pretrained by unsupervised learning might be transferred to downstream tasks by fine-tuning. In the previous work, cross entropy training is used for pretraining [12]. In this work, we explore if models unsupervised pretrained by MoCo is helpful for the downstream supervised learning.

3. Experiments

3.1. Datasets

We train and evaluate our models on two datasets, Dataset A with speaker labels and Dataset B without any speaker information.

Dataset A consists of utterances from 2, 943 speakers, each with 1, 000 utterances. All speech are recorded with mobile phones (iOS or Android systems) with a sample rate 16 kHz. The duration of utterances are distributed between 2.5s and 6.5s, with a mode duration 3.5s. We split the dataset into training and test set. The training set consists of 2, 500 speakers. For each of the remaining 443 speakers, we randomly selected three utterances for enrollments, and 50 utterance for evaluation, resulting in about 9.8 million trials. Other remaining utterances of the test speakers are not used in the experiments.

Dataset B consists of mobile phone recorded speech with sample rate 16 kHz, without any speaker specific information. There are 7.6 million utterances in total, or about 9 thousand hours before silences removing.

The information of the two datasets are summarized in Tab. 1.

3.2. Baseline systems

3.2.1. I-vector system

The i-vector system is based on the standard kaldi recipe [15] (egs/voxceleb/v1). The features are 30-dimensional Mel-Frequency Cepstral Coefficients (MFCCs), with a frame shift of

| dataset | training | test |
|---------|--|--|
| A | 2, 500 speakers 1, 000 utterances | 443 speakers (3 for enrollment, 50 for evaluation) |
| B | ~ 7.6 million utterances ~ 9k hours | - |

Table 1: Overview of the training and evaluation data

| Layer | Layer Context | Input \times Output |
|---------------|---------------|-----------------------------|
| frame1 | [t-2, t+2] | $(5 \times d) \times 512$ |
| frame2 | {t-2, t, t+2} | $(3 \times 512) \times 512$ |
| frame3 | {t-3, t, t+3} | $(3 \times 512) \times 512$ |
| frame4 | {t} | 512×512 |
| frame5 | {t} | 512×1500 |
| stats pooling | [0, T) | $1500T \times 3000$ |
| embed_a | {0} | 3000×512 |
| embed_b | {0} | 512×512 |
| softmax | {0} | $512 \times D$ |

Table 2: X-vector network architecture used in this paper, where d , T , and D denote the dimensionality of input feature, the number of utterance frames, and the number of speakers, respectively. Embeddings are extracted at layer embed_b (before non-linearity).

10ms and a window width of 25ms. Delta and acceleration are appended to create 90 dimension feature vectors. Mean normalization is then applied over a sliding window of up to 3 seconds. To filter out non-speech frames, kaldi’s energy-based voice activity detector (VAD) is employed. (We also tried a neural network based VAD, but observe no improvement.) The UBM is a 2048 component full-covariance GMM. The system uses a 400 dimension i-vector extractor. I-vectors are centered, dimensionality reduced to 200 using LDA. Length normalization [16] is applied before PLDA scoring [2].

3.2.2. X-vector system

Our x-vector is based kaldi’s voxceleb recipe. The TDNN structure used is as Tab. 2. We train the x-vector network using pytorch [17], over four V100 GPUs. Synchronous Stochastic Gradient Descent (SGD) with weight decay $= 1e - 5$ and momentum $= 0.9$ is used. The total batch size is 1024 (256 for each GPU). The learning rate decays from $1e - 4$ to $1e - 5$ exponentially. A dropout layer [18] with $p = 0.5$ is applied after before embed_b. A max gradient norm of 2 is applied to stabilize training.

After silence removing, we filter out utterances shorter than 250 frames and make sure each speaker has at least 30 utterances, ending up with 2, 453 speaker, about 1184 hours. We train 30 times of the total training frames, with each epoch processing 72 million frames (i.e. 200 hours). We random select 3 utterances from each speaker as holdout validation set.

Like the ivector-system, a PLDA is used as backend for scoring. The LDA reduction dimension is set to 150.

3.2.3. X-vector with AAM loss

Additive Angular Margin (AAM) loss is reported more effective than the conventional cross entropy loss for both face [10] and

speaker verification tasks [12].

AAM loss is defined as Eq. 1:

$$L = -\frac{1}{N} \sum_{i=1}^N \log \frac{e^{s(\cos(\theta_{y_i} + m))}}{e^{s(\cos(\theta_{y_i} + m))} + \sum_{j=1, j \neq y_i}^D e^{s \cos \theta_j}} \quad (1)$$

where $\cos(\theta)$ is the dot product between the embed_b and the last fully connected layer (softmax layer) (after L2 normalization), additive angular m and scale factor s are hyperparameters.

In our experiments, layers after embed_b are dropped and an AAM loss is applied instead of cross entropy loss. We use $s = 32$, $m = 0.3$ for all experiments. The learning rate decays from $1e - 5$ to $1e - 6$ exponentially. No dropout is used, as we found it degrades the performance. A max gradient norm of 6 is applied to stabilize training. All other hyperparameters stay the same as the cross entropy training.

We try two backend, 1) cosine distance and 2) PLDA as the aforementioned the x-vector system.

The EERs and minDCF of the three baseline systems are listed in Tab 3. It can be found that AAM can achieve better performances than cross entropy based training, and this observation is in line with findings by [11, 12]. Direct cosine distance can achieve decent performance, and a separate PLDA backend contribute no improvement (though no significant performance degradation w.r.t EER, is surpassed by cosine on minDCF criterion by large margins). Thus, we use AAM-loss trained network with cosine backend for future comparison.

3.3. MoCo as extractor

The network used for x-vector system (c.f. Sec. 3.2.2) is used as the MoCo encoder. We use a queue size 10, 000, and $\beta = 0.99$ $\tau = 0.07$. For SpecAugment, time warp window 10, max time mask width 20, max frequency mask width 10. The learning rate decays exponentially from $1e - 4$ to $1e - 5$.

After training, the encoder is used for xvector extraction. As in the xvector-AAM case, we try both cosine and PLDA backends. The results are listed in Tab. 4.

We compare MoCo with i-vector system, as both MoCo and i-vector’s training process is unsupervised. As can be seen, with cosine backend (complete unsupervised), MoCo outperforms i-vector. It is not surprising, as MoCo optimizes dot product (cosine) as the training criterion. Interestingly, helped by a strong backend (PLDA), i-vector can achieve an EER 1.661%, significantly better than MoCo (2.655%).

It is not clear whether some backend other than PLDA used here can improve the performance for MoCo. We leave this topic for future work.

To explore if extra data is helpful for MoCo learning, we train MoCo use extra unlabelled data from Dataset B. We use a larger queue size (20, 000), and other hyperparameters stay the same. The performance of MoCo encoder can improve further, from 4.275% to 2.655% for the cosine backend, and 3.58% to 2.366% for the PLDA backend.

3.4. MoCo as pretraining

In this section, we study if the encoder trained with MoCo is helpful for downstream supervised learning. We initialize x-vector with the encoder pretrained by MoCo in Sec. 3.3. As comparison, we also conduct experiment with conventional pretraining with cross entropy training (as used in [12]). As shown in Tab 5, model pretrained by MoCo can great help the training. The EER improves from 1.44% to 1.242%, a 13.7% relative

| method | ivector (cosine) | ivector (PLDA) | xvector-CE (PLDA) | xvector-AAM (PLDA) | xvector-AAM (cosine) |
|-------------------------|------------------|----------------|-------------------|--------------------|----------------------|
| EER (%) | 5.178 | 1.661 | 1.562 | 1.589 | 1.44 |
| minDCF _{0.01} | 0.397 | 0.191 | 0.217 | 0.24 | 0.165 |
| minDCF _{0.001} | 0.622 | 0.409 | 0.487 | 0.572 | 0.346 |

Table 3: Performance for the baseline systems.

| method | cosine | LDA-PLDA |
|--------------------|-------------|--------------|
| ivector | 5.178 | 1.661 |
| MoCo | 4.275 | 2.655 |
| MoCo (+ Dataset B) | 3.58 | 2.366 |

Table 4: EERs (%) of encoder learned by MoCo with different backends.

| method | EER (%) | minDCF _{0.01} | minDCF _{0.001} |
|--------------------------|--------------|------------------------|-------------------------|
| xvector-AAM | 1.44 | 0.165 | 0.346 |
| xvector-AAM (CE) | 1.481 | 0.169 | 0.357 |
| xvector-AAM (MoCo) | 1.242 | 0.157 | 0.334 |
| xvector-AAM (MoCo-extra) | 1.223 | 0.146 | 0.312 |

Table 5: Results of AAM trained models using different pre-training strategy. CE, Moco, MoCo-extra stands for, respectively, cross entropy pretraining, pretraining with Dataset A, and with both Dataset A and Dataset B. Cosine backend is applied for all models

improvement. At the same time, we find that no improvement is observed with cross entropy pretraining. Cross entropy pretraining does not transfer well to the downstream AAM training. In fact it is harmful for AAM training in our case. One possible explanation is that both MoCo and AAM encourage the representation to hyper-sphere subspace (via softmax over dot product), while cross entropy learns a difference subspace.

The DET curves of the systems under study are illustrated in Fig. 3. Moco pretraining improves the both EER and minDCF criteria; in the meantime, it achieves better performance under different false accept and false reject tradeoffs.

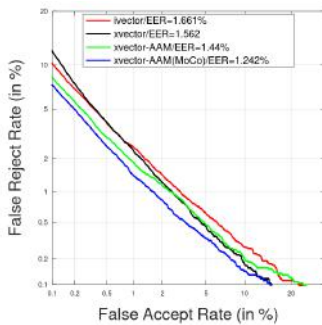


Figure 3: The DET curves of Ivector and Xvector models trained with different losses and strategies trained on Dataset A.

| method | EER (%) | minDCF _{0.01} | minDCF _{0.001} |
|---------------------------------------|--------------|------------------------|-------------------------|
| ivector (PLDA) | 5.467 | 0.4859 | 0.6213 |
| xvector-CE (PLDA) | 3.356 | 0.3591 | 0.5890 |
| xvector-AAM (cosine) | 2.497 | 0.2634 | 0.3888 |
| xvector-AAM (PLDA) | 2.752 | 0.3717 | 0.4971 |
| xvector-AAM (CE pretraing) | 2.572 | 0.2535 | 0.3799 |
| xvector-AAM (MoCo pretraing) (cosine) | 2.402 | 0.2476 | 0.3506 |

Table 6: Results on VoxCeleb1 trained on VoxCeleb1 dev set and all VoxCeleb2 data.

3.5. Results on Voxceleb

To check if the observations on our in-house data generalize, we conduct experiments on the public available dataset Voxceleb. All data from Voxceleb2 [19] and the training part to Voxceleb1 [20] are used for training. After converting the audios from amr to wav format, it ends up with 7,323 speaker and 1,276,888 utterances in total. Voxceleb1 test is used as the test, which has 4,874 utterance from 40 speakers. The official test protocol is used (37,720 trials in total).

The same network architecture and training configuration as previous experiments are used, and no further hyper parameter tuning is conducted for this task. According to Tab 6, similar tendency with the previous experiments is observed. MoCo pretraining does help AAM training, while CE pretraining doesn't. Besides, the results confirm that embeddings learned by AAM works better with cosine backend.

4. Conclusions

In this paper, we explore the effectiveness of MoCo for learning speaker embedding. To apply MoCo with speech data, we propose a parallel distorted data generating strategy based on SpecAugment. The experiments show that MoCo learns good speaker discriminative embedding, and can be used as an effective pretraining method for the downstream supervised training. On a large-scale dataset, we build a strong baseline system with AAM, which can significantly outperform the conventional cross entropy based x-vector system. Compared to the baseline system trained from scratch, MoCo pretraining can achieve a 13.7% relative EER improvement while the conventional cross entropy pretraining gains no improvement. Thanks to MoCo's unsupervised nature, extra unlabelled data could be used to improve the performance further.

5. 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, 2010.
- [2] S. Ioffe, "Probabilistic linear discriminant analysis," in *European Conference on Computer Vision*. Springer, 2006, pp. 531–542.
- [3] D. Snyder, D. Garcia-Romero, D. Povey, and S. Khudanpur, "Deep neural network embeddings for text-independent speaker verification," in *Interspeech*, 2017, pp. 999–1003.
- [4] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-vectors: Robust dnn embeddings for speaker recognition," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 5329–5333.
- [5] K. Chen and A. Salman, "Learning speaker-specific characteristics with a deep neural architecture," *IEEE Transactions on Neural Networks*, vol. 22, no. 11, pp. 1744–1756, 2011.
- [6] C. Zhang, K. Koishida, and J. H. L. Hansen, "Text-independent speaker verification based on triplet convolutional neural network embeddings," *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, vol. 26, no. 9, pp. 1633–1644, Sep. 2018.
- [7] P. Ghosh and L. S. Davis, "Understanding center loss based network for image retrieval with few training data," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018.
- [8] N. Li, D. Tuo, D. Su, Z. Li, and D. Yu, "Deep discriminative embeddings for duration robust speaker verification," in *Interspeech*, 2018, pp. 2262–2266.
- [9] Y. Li, F. Gao, Z. Ou, and J. Sun, "Angular softmax loss for end-to-end speaker verification," in *2018 11th International Symposium on Chinese Spoken Language Processing (ISCSLP)*. IEEE, 2018, pp. 190–194.
- [10] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "Arcface: Additive angular margin loss for deep face recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 4690–4699.
- [11] X. Xiang, S. Wang, H. Huang, Y. Qian, and K. Yu, "Margin matters: Towards more discriminative deep neural network embeddings for speaker recognition," *arXiv preprint arXiv:1906.07317*, 2019.
- [12] H. Zeinali, S. Wang, A. Silnova, P. Matějka, and O. Plchot, "But system description to voxceleb speaker recognition challenge 2019," *arXiv preprint arXiv:1910.12592*, 2019.
- [13] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," *arXiv preprint arXiv:1911.05722*, 2019.
- [14] D. S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le, "SpecAugment: A simple data augmentation method for automatic speech recognition," *arXiv preprint arXiv:1904.08779*, 2019.
- [15] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz *et al.*, "The kaldı speech recognition toolkit," in *IEEE 2011 workshop on automatic speech recognition and understanding*, no. CONF. IEEE Signal Processing Society, 2011.
- [16] D. Garcia-Romero and C. Y. Espy-Wilson, "Analysis of i-vector length normalization in speaker recognition systems," in *Twelfth Annual Conference of the International Speech Communication Association*, 2011.
- [17] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga *et al.*, "Pytorch: An imperative style, high-performance deep learning library," in *Advances in Neural Information Processing Systems*, 2019, pp. 8024–8035.
- [18] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting," *The journal of machine learning research*, vol. 15, no. 1, pp. 1929–1958, 2014.
- [19] J. S. Chung, A. Nagrani, and A. Zisserman, "Voxceleb2: Deep speaker recognition," in *Interspeech*, 2018.
- [20] A. Nagrani, J. S. Chung, and A. Zisserman, "Voxceleb: A large-scale speaker identification dataset," in *Interspeech*, 2017, pp. 2616–2620.