TRIPLET DISTILLATION FOR DEEP FACE RECOGNITION

Yushu Feng[1], *Huan Wang*[1], *Haoji (Roland) Hu*[1,*], *Lu Yu*[1], *Wei Wang*[2], *Shiyan Wang*[2]

[1]College of Information Science and Electronic Engineering, Zhejiang University, Hangzhou, China
[2]Chongqing University of Posts and Telecommunications, Chongqing, China.

ABSTRACT

Convolutional neural networks (CNNs) have achieved great successes in face recognition, which unfortunately comes at the cost of massive computation and storage consumption. Many compact face recognition networks are thus proposed to resolve this problem, and triplet loss is effective to further improve the performance of these compact models. However, it normally employs a fixed margin to all the samples, which neglects the informative similarity structures between different identities. In this paper, we borrow the idea of knowledge distillation and define the informative similarity as the transferred knowledge. Then, we propose an enhanced version of triplet loss, named *triplet distillation*, which exploits the capability of a teacher model to transfer the similarity information to a student model by adaptively varying the margin between positive and negative pairs. Experiments on the LFW, AgeDB and CPLFW datasets show the merits of our method compared to the original triplet loss.

***Index Terms*—** Face Recognition, Knowledge Distillation, Triplet Loss, Network Compression

1. INTRODUCTION

Recent years have witnessed the impressive success of CNNs in computer vision tasks, especially in the area of face recognition [1]. However, effective face recognition CNN models typically consume a large amount of storage and computation, making it difficult to deploy on mobile and embedded devices. To resolve this problem, several lightweight face recognition CNN models have been proposed, such as MobileID [2], ShiftFaceNet [3], and MobileFaceNet [4].

Unfortunately, model size reduction usually coincides with performance decline, thus many studies aim to improve the performance of small models. Triplet loss [5], as a metric learning method, is widely used in face recognition to further improve accuracy [6]. Triplet loss explicitly maximizes the inter-class distance and meanwhile minimizes the intra-class distance, where a margin term is used to determine the decision boundaries between positive and negative pairs.

In the original triplet loss, the margin is set to a constant, which tends to push the decision boundaries among different classes to the same value, thus loses the hidden similarity structures of different identities. Generally, the information among different identities help reduce the intra-class variations and enlarge the inter-class differences in the feature space. Therefore, it is necessary to set a *dynamic* margin to take into account the similarity structures. In this vein, Zakharov *et al.* [7] sets the margin term as a function of angular differences between the poses for pose estimation; Wang *et al.* [8] formulates the adaptive margin as a nonlinear mapping of the average distances among different people for person re-identification. However, these researches obtain the dynamic margins by handcrafted rules rather than learned distances, which neglects the instance relationship in feature spaces. In this paper, we propose an enhanced version of triplet loss, named *triplet distillation*, which borrows the idea of knowledge distillation [9] to determine the dynamic margins for face recognition. Specifically, we determine the similarity between two identities according to distances learned by the teacher model. This similarity, as a kind of knowledge, is then applied to guiding the student model to optimize its decision boundaries.

The major contributions of this paper lie in three aspects:

- We propose the triplet distillation method to transfer knowledge from a teacher model to a student model for face recognition.
- We improve the triplet loss with dynamic margins by utilizing the similarity structures among different identities, which is in contrast with the fixed margin of the original triplet loss.
- Experiments on the LFW [10], AgeDB [11] and CPLFW [12] datasets show that the proposed method performs favorably against the original triplet loss.

2. RELATED WORK

Triplet loss. The main purpose of triplet loss [5] is to distinguish identities in the projected space with the guidance of distances among an anchor sample, a positive sample, and a negative sample. There are several revisions for the original triplet loss, which mainly fall into the following three categories: (1) Adding new constraints to the objective function

*Corresponding Author.

to improve the generalization performance [13, 14]; (2) Optimizing the selection of triplet samples to make the triplet samples more informative, which can lead to faster convergence and better performance [15]; (3) Proposing dynamic margins for different triplet combinations, such as [7, 8], which use handcrafted methods to determine the similarities among different identities. Our method belongs to the last category. Different from previous approaches, we exploits a teacher model to obtain the similarity information among identities to set the dynamic margins.

Knowledge distillation. Knowledge distillation, firstly proposed by [16] and then refined by Hinton [9], is a model compression method to transfer the knowledge of a large teacher network to a small student network. Most researches follow [9] to learn the soft-target outputs of the teacher network [17]. The definition of knowledge can also refer to its feature maps [18]. Recent researches further broaden the definition of knowledge to other attributes such as attention maps [19].

Traditional knowledge distillation methods only focus on processing a single instance (or sample) and independently extract instance features as the distilled knowledge, but do not consider the instance relationship, which contains information for a student to reduce the intra-class variations and enlarge the inter-class differences in the feature space. Thus, some methods are proposed to explore the correlation of samples to aid distillation. In [20], the authors use the Radial Basis Function (RBF) as a metric to calculate the correlation between two samples, with the aim to make the correlation extracted by the student similar with the teacher. Park *et al.* [21] not only considers the Euclidean distance between two instances, but also defines the angle-wise distillation losses that penalize structural differences in sample relations. In this paper, we also use the knowledge of feature relationship between instances as a guidance to train the student model. However, the knowledge that we define has physical meanings which reflect the similarities between identities.

3. THE PROPOSED METHOD

3.1. Teacher and student networks

We employ the widely-used ResNet-101 [22] as the teacher model. For the student model, we adopt a slim version of MobileFaceNet [4], which has the same architecture as MobileFaceNet, yet with three quarters of the number of channels in each convolutional layer on average. The detailed statistics of the teacher and student model are summarized in Table 1.

Model	Size/MB	Params/10^6	FLOPs/10^9	Time/s
T	248.8	652.25	24.23	2.45
M	4.0	0.99	0.49	0.31
S	2.3	0.59	0.23	0.18

Table 1. Comparison of the teacher model, MobileFaceNet [4], and the student model. 'T' means the teacher model. 'M' means MobileFaceNet and 'S' represents the student model. The FLOPs are counted by TFProf, a profiling tool in Tensorflow. The inference time is averaged by 5000 runs of forwarding an image of size $112 \times 112 \times 3$ on Intel Xeon(R) CPU E5-2609 v4 @1.70GHz with single thread.

3.2. Triplet distillation

Triplet loss is applied to a triplet of samples, represented as $\{x^a, x^p, x^n\}$. Here x^a is the anchor image; x^p is called the positive image, which belongs to the same identity as x^a, and x^n is called the negative image, which belongs to a different identity of x^a. The triplet loss aims to minimize the distance between the anchor and positive images, and meanwhile maximizes the distance between the anchor and negative images. The objective function of triplet loss can be formulated as

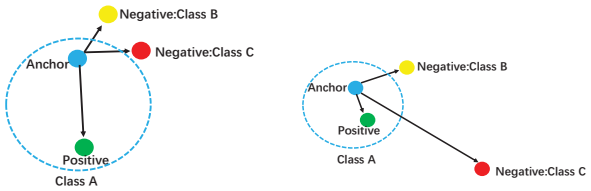
$$\mathcal{L} = \frac{1}{N}\sum_{i}^{N} \max(\mathcal{D}(x_i^a, x_i^p) - \mathcal{D}(x_i^a, x_i^n) + m, 0), \quad (1)$$

where N is the number of triplets in a mini-batch; $\mathcal{D}(*,*)$ denotes the distance between two images, usually Euclidean distance or cosine distance. Notably, the hyper-parameter m represents a margin enforced between the positive and negative pairs, that is, only when the distance difference between the negative pair and the positive pair is smaller than a threshold m, will the loss $\mathcal{L}$ count. Thus, the final distances among different identity clusters will be pushed to the margin m.

In the original triplet loss, m is the *same* for all identities. In other words, all identity clusters are separated with roughly the same distances, which ignores the subtle similarity structures among different identities, since different people are not equally different. Fig. 1 illustrates one example of the above ideas. If Person A looks more similar to Person B than to Person C, then it should be better to set the m for $\{A, B\}$ smaller than the m for $\{A, C\}$ because such setting will push A and B closer than A and C in the hyperspace of the student model.

In a similar spirit to dark knowledge proposed in knowledge distillation [9], this similarity structure is informative and useful, but not considered in the original triplet loss. Our proposed triplet distillation method exploits knowledge distillation to bridge this gap.

First, the teacher model extracts two features from a triplet and obtains the distance between them. Then, we map this distance into the margin and apply it to the training of the student model. Different from previous mathematical angle calculation methods [7, 8], our scheme adopts the well-trained teacher model to calculate the face distance, which has more capability to capture the similarity structures in its learned representations. With the proposed dynamic margin term, the



(a) Before training. (b) After training.

Fig. 1. The idea of setting different m for different identities. If Class A is more similar to Class B than to Class C, then we set the m for $\{A, B\}$ smaller than that for $\{A, C\}$ in order to push A and B closer than A and C in the hyperspace of the student model after training.

objective function can be written as

$$\mathcal{L} = \frac{1}{N}\sum_{i}^{N} \max(\mathcal{D}(x_i^a, x_i^p) - \mathcal{D}(x_i^a, x_i^n) + \mathcal{F}(d), 0), \quad (2)$$

$$d = \max(\mathcal{T}(x_i^a, x_i^n) - \mathcal{T}(x_i^a, x_i^p), 0), \quad (3)$$

where $\mathcal{D}(*,*)$ denotes the distance between two images calculated by the student model, $\mathcal{T}(*,*)$ represents the distance calculated by the teacher model, and d denotes the distance between intra-class and inter-class features extracted by the teacher model. In this paper, we use the cosine distance for all experiments. Here $\mathcal{F}(*)$ represents the function of the margin with regards to the distance. We employ a simple increasing linear function for $\mathcal{F}(*)$,

$$\mathcal{F}(d) = \frac{m_{max} - m_{min}}{d_{max}} d + m_{min} \quad (4)$$

where m_{min} and m_{max} represent the minimum and maximum values of margin; and d_{max} represents the maximum distance in a mini-batch. In this way, the margin is constrained between m_{min} and m_{max}.

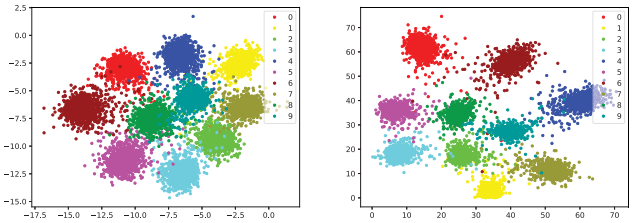
3.3. Feature distribution visualization on MNIST

To better understand the effect of our method, we design a toy experiment to visualize the feature distributions on MNIST. The student network is a 4-layer CNN model with 2-dimensional features and the teacher network is a 7-layer CNN model. Their structures are shown in Table. 2. Because MNIST consists of 10 classes, the two-dimensional vectors output by the last layers of the teacher and student network are fed into another fully connected layer and a softmax layer to obtain probability distributions of 10 classes.

The whole training process lasts for 200 epochs. The learning rate is initialized with 1×10^{-4}, and sequentially divided by 10 at 40, 80, 120, 160 epochs respectively. For triplet loss, we vary the margin m several times and take the best result when $m = 0.75$. For the proposed method, we set the maximum and minimum values of m to 1.0 and 0.5 respectively.

Input	Teacher Layer		Input	Student Layer
$28^2 \times 1$	conv, 5×5		$28^2 \times 1$	conv, 5×5
$28^2 \times 32$	conv, 5×5		$28^2 \times 32$	maxpool
$28^2 \times 32$	maxpool		$14^2 \times 32$	conv, 5×5
$14^2 \times 32$	conv, 5×5		$14^2 \times 64$	maxpool
$14^2 \times 64$	conv, 5×5		$7^2 \times 64$	conv, 5×5
$14^2 \times 64$	maxpool		$7^2 \times 128$	maxpool
$7^2 \times 64$	conv, 5×5		$3^2 \times 128$	2D fc
$7^2 \times 128$	conv ,5×5			
$7^2 \times 128$	maxpool			
$3^2 \times 128$	2D fc			

Table 2. The teacher and student networks for feature distribution visualization. The student network is a 4-layer CNN model and the teacher network is a 7-layer CNN model. Both of them output two-dimensional feature vectors. All convolution layers use stride 1 and padding 2, and all maxpool layers use stride 2.



(a) Triplet loss. (b) Triplet distillation.

Fig. 2. Feature distribution visualization on the MNIST validation set. Different colors represent different classes.

Fig. 2 shows the feature distribution visualization of the outputs of the student network on the MNIST validation set. The visualized two-dimensional feature vectors well demonstrate that triplet distillation by varying margin m could bring the larger margin property to the features compared with original triplet loss with a fixed m.

4. EXPERIMENTS

4.1. Implementation details

Pre-processing. We use MTCNN [23] to detect faces and facial landmarks on the MS-Celeb-1M dataset [24]. To obtain data of higher quality, 3.8 million photos from 85 identities are picked out to make a refined MS-Celeb-1M dataset [6]. All the images are aligned based on the detected landmarks and then resized to 112×112 with normalization (subtracted by mean 127.5 and divided by standard deviation 128).

Model	LFW	AgeDB-30	CPLFW
Teacher	99.73%	98.25%	92.85%
Student	98.75%	93.53%	78.53%
Student+Triplet loss			
margin=0.3	99.21%	94.08%	80.80%
margin=0.4	99.23%	94.00%	81.16%
margin=0.5	99.20%	93.80%	80.38%
Student+Triplet distillation	**99.27%**	**94.25%**	**81.28%**

Table 3. Comparison of the proposed triplet distillation with triplet loss.

Training. The architectures of the teacher and student models are described in Section 3.1. Both of them are first trained from scratch with the ArcFace loss [6]. Stochastic Gradient Descent (SGD) is used with momentum 0.9 and batch size 480. The learning rate begins with 0.1 and is divided by 10 at iteration $30k$ and $70k$, before the training finally ends at iteration $100k$.

Then the proposed triplet distillation is used to fine-tune the student model. During this stage, there are 10 classes, 18 images per class in each mini-batch. The learning rate is 0.001 and the training stops at $34k$ iterations. We randomly sample 1000 triplets from the refined MS-Celeb-1M dataset to obtain different d's (Equation (3)). Then the largest one is chosen as $m_{max} = 0.5$, and the smallest one as $m_{min} = 0.2$. TensorFlow [25] is used in all our experiments.

Evaluation. In the evaluation stage, we extract the features of each image and its horizontally flipped image. Then the two features are concatenated as one for face verification using the cosine distance.

Three popular face verification datasets are considered here: LFW [10], AgeDB [11] and CPLFW [12]. The LFW validation set randomly selects 6000 pairs of face pictures from its total dataset to form picture pairs, of which 3000 pairs are positive samples and 3000 pairs are negative samples; The CPLFW validation set selects 3000 pairs of pictures with obvious pose differences as the positive samples from LFW to detect the influence of face pose, and then selects another 3000 pairs of pictures with the same race and the same gender as negative samples; The AgeDB validation set has significant differences in the characters, poses, expressions, ambient lighting, and character ages, which consists of 5 different year gaps. We choose one of them with 300 positive and 300 negative pairs as our evaluation dataset.

4.2. Experimental results

As shown in Table. 3, the pre-trained teacher model reaches 99.73% on LFW, 98.25% on AgeDB-30, and 92.85% on CPLFW. The student model trained by ArcFace reaches 98.75% on LFW, 93.53% on AgeDB-30, and 78.53% on CPLFW.

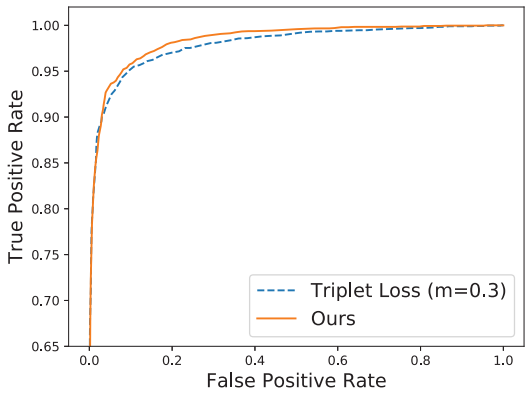


Fig. 3. The ROC curve on the AgeDB-30 dataset.

For comparison with the original triplet loss, we set the fixed margin m as 0.3, 0.4, and 0.5, which are chosen based on experiments for the best performance of the triplet loss. After applying the proposed triplet distillation to the student model, its verification accuracy is boosted to 99.27% on LFW, 94.25% on AgeDB-30, and 81.28% on CPLFW. In other words, the accuracy of triplet distillation is consistently greater than the original triplet loss using the fixed margin.

Fig. 3 shows the ROC curve on the AgeDB-30 dataset in our experiment. It is obviously that at the same False Positive Rate (FPR), the proposed triplet distillation obtains greater True Positive Rate (TPR) than the original triplet loss, which fully illustrates the performance improvement of the proposed triplet distillation in face recognition tasks.

5. CONCLUSION

We propose triplet distillation for deep face recognition, which takes advantage of knowledge distillation to generate dynamic margins to enhance the triplet loss. The distance obtained by the teacher model reflects similarity information between different identities, which can be regarded as a new type of knowledge. Compared with the original triplet loss, experiments have proven that our proposed method delivers encouraging performance improvements.

6. ACKNOWLEDGE

This work is supported by the Natural Key RD Program of China under Grant 2017YFB1002400, the Chongqing Research Program of Basic science and Frontier Technology under Grant CSTC2016JCYJA0542, and the ZJU-SUTD IDEA Innovation Design Project under Grant 188170-11102/017.

7. REFERENCES

[1] Mei Wang and Weihong Deng, "Deep face recognition: A survey," *arXiv preprint arXiv:1804.06655v7*, 2018.

[2] Ping Luo, Zhenyao Zhu, Ziwei Liu, Xiaogang Wang, and Xiaoou Tang, "Face model compression by distilling knowledge from neurons," in *AAAI*, 2016.

[3] Bichen Wu, Alvin Wan, Xiangyu Yue, Peter Jin, Sicheng Zhao, Noah Golmant, Amir Gholaminejad, Joseph Gonzalez, and Kurt Keutzer, "Shift: A zero flop, zero parameter alternative to spatial convolutions," in *CVPR*, 2018.

[4] Sheng Chen, Yang Liu, Xiang Gao, and Zhen Han, "Mobilefacenets: Efficient cnns for accurate real-time face verification on mobile devices," in *Chinese Conference on Biometric Recognition*, 2018.

[5] Florian Schroff, Dmitry Kalenichenko, and James Philbin, "Facenet: A unified embedding for face recognition and clustering," in *CVPR*, 2015.

[6] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou, "Arcface: Additive angular margin loss for deep face recognition," *arXiv preprint arXiv:1801.07698*, 2018.

[7] Sergey Zakharov, Wadim Kehl, Benjamin Planche, Andreas Hutter, and Slobodan Ilic, "3D object instance recognition and pose estimation using triplet loss with dynamic margin," in *IROS*, 2017.

[8] Jiayun Wang, Sanping Zhou, Jinjun Wang, and Qiqi Hou, "Deep ranking model by large adaptive margin learning for person re-identification," *PR*, vol. 74, pp. 241–252, 2018.

[9] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean, "Distilling the knowledge in a neural network," *arXiv preprint arXiv:1503.02531*, 2015.

[10] Gary B Huang, Marwan Mattar, Tamara Berg, and Eric Learned-Miller, "Labeled faces in the wild: A database forstudying face recognition in unconstrained environments," in *International Workshop on Faces in Real-Life Images: Detection, Alignment, and Recognition*, 2008.

[11] Stylianos Moschoglou, Athanasios Papaioannou, Christos Sagonas, Jiankang Deng, Irene Kotsia, and Stefanos Zafeiriou, "AgeDB: the first manually collected, in-the-wild age database," in *CVPR*, 2017.

[12] Tianyue Zheng and Weihong Deng, "Cross-pose LFW: A database for studying cross-pose face recognition in unconstrained environments," Tech. Rep. 18-01, Beijing University of Posts and Telecommunications, February 2018.

[13] De Cheng, Yihong Gong, Sanping Zhou, Jinjun Wang, and Nanning Zheng, "Person re-identification by multi-channel parts-based cnn with improved triplet loss function," in *CVPR*, 2016.

[14] Weihua Chen, Xiaotang Chen, Jianguo Zhang, and Kaiqi Huang, "Beyond triplet loss: a deep quadruplet network for person re-identification," in *CVPR*, 2017.

[15] Xingping Dong and Jianbing Shen, "Triplet loss in siamese network for object tracking," in *ECCV*, 2018.

[16] Cristian Buciluǎ, Rich Caruana, and Alexandru Niculescu-Mizil, "Model compression," in *SIGKDD*, 2006.

[17] Guorui Zhou, Ying Fan, Runpeng Cui, Weijie Bian, Xiaoqiang Zhu, and Kun Gai, "Rocket launching: A universal and efficient framework for training well-performing light net," in *AAAI*, 2018.

[18] Hanting Chen, Yunhe Wang, Chang Xu, Chao Xu, and Dacheng Tao, "Learning student networks via feature embedding," *arXiv preprint arXiv:1812.06597*, 2018.

[19] Sergey Zagoruyko and Nikos Komodakis, "Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer," *arXiv preprint arXiv:1612.03928*, 2016.

[20] Baoyun Peng, Xiao Jin, Jiaheng Liu, Dongsheng Li, Yichao Wu, Yu Liu, Shunfeng Zhou, and Zhaoning Zhang, "Correlation congruence for knowledge distillation," in *ICCV*, 2019.

[21] Wonpyo Park, Dongju Kim, Yan Lu, and Minsu Cho, "Relational knowledge distillation," in *CVPR*, 2019.

[22] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *CVPR*, 2016.

[23] Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao, "Joint face detection and alignment using multitask cascaded convolutional networks," *IEEE Signal Processing Letters*, vol. 23, no. 10, pp. 1499–1503, 2016.

[24] Yandong Guo, Lei Zhang, Yuxiao Hu, Xiaodong He, and Jianfeng Gao, "Ms-celeb-1m: A dataset and benchmark for large-scale face recognition," in *ECCV*, 2016.

[25] Martín Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, et al., "Tensorflow: A system for large-scale machine learning," in *Symposium on Operating Systems Design and Implementation*, 2016.