AN EFFICIENT DEEP NEURAL NETWORKS TRAINING FRAMEWORK FOR ROBUST FACE RECOGNITION

Canping Su[1], *Yan Yan*[1,⋆], *Si Chen*[2], *Hanzi Wang*[1]

[1] School of Information Science and Engineering, Xiamen University, Xiamen 361005, China
[2] School of Computer and Information Engineering, Xiamen University of Technology, Xiamen 361024, China

ABSTRACT

In recent years, the triplet loss-based deep neural networks (DNN) are widely used in the task of face recognition and achieve the state-of-the-art performance. However, the complexity of training the triplet loss-based DNN is significantly high due to the difficulty in generating high-quality training samples. In this paper, we propose a novel DNN training framework to accelerate the training process of the triplet loss-based DNN and meanwhile to improve the performance of face recognition. More specifically, the proposed framework contains two stages: 1) The DNN initialization. A deep architecture based on the softmax loss function is designed to initialize the DNN. 2) The adaptive fine-tuning. Based on the trained model, a set of high-quality triplet samples is generated and used to fine-tune the network, where an adaptive triplet loss function is introduced to improve the discriminative ability of DNN. Experimental results show that, the model obtained by the proposed DNN training framework achieves 97.3% accuracy on the LFW benchmark with low training complexity, which verifies the efficiency and effectiveness of the proposed framework.

***Index Terms*—** Face recognition, deep neural networks, triplet loss function

1. INTRODUCTION

During the past few decades, face recognition has received increasing attention in both industry and academia due to its wide range of applications. Face recognition technologies have been successfully used in access control, video surveillance and law enforcement, etc.

Since the proposal of AlexNet by Krizhevsky et al. [1], deep neural networks (DNN) have become one of the most successful techniques in computer vision and pattern recognition. Especially, face recognition has made substantial progress in the recognition accuracy due to the recent development of the DNN-based methods. Now face recognition is at a level that compares favorably with humans for the frontal faces [2].

DeepFace [3] firstly applies convolutional neural networks (CNN) to face recognition. The CNN training is based on the softmax loss function, which can be used for multi-class classification. FaceNet [4] proposes a novel loss function called the triplet loss function, which can significantly improve the discriminative ability of recognizing different persons. However, the triplet loss-based DNN is hard to be trained due to the difficulty in generating high-quality triplet samples. The DeepID based methods [5, 6, 7] use ensemble techniques and joint identification-verification supervisory signals to improve the performance, which makes the training process complicated. Recently, ResNet [8] explores the benefits of the very deep architecture and shows promising results. However, ResNet needs a large amount of memory to store the intermediate results.

Although the performance of the DNN-based face recognition methods is greatly improved, the DNN training process becomes significantly complicated, resulting in high training complexity. Therefore, we focus on the study of efficiently training DNN to reduce the training complexity, while maintaining (even improving) the performance of DNN-based face recognition.

In this paper, we propose an efficient DNN training framework for robust face recognition. The proposed framework takes advantage of softmax loss-based DNN training to initialize the parameters of DNN. Based on it, we further propose an adaptive triplet loss function, which can effectively generate a set of high-quality triplet samples and successfully overcome the problem of slow convergence (caused by adopting a fixed margin in the conventional triplet loss function), to fine-tune the DNN.

The main contributions are summarized as follows. 1) We develop an adaptive triplet loss function, which characterizes the intra-class and inter-class face variations more effectively than the conventional triplet loss function. 2) We propose an efficient DNN training framework, which advantageously combines the softmax loss and triplet loss to obtain the powerful face representation. Experimental results on the tasks of face verification and face identification show that the proposed DNN training framework is highly efficient and the trained DNN model shows superior performance.

[⋆] Corresponding author

978-1-5090-2175-8/17/$31.00 ©2017 IEEE

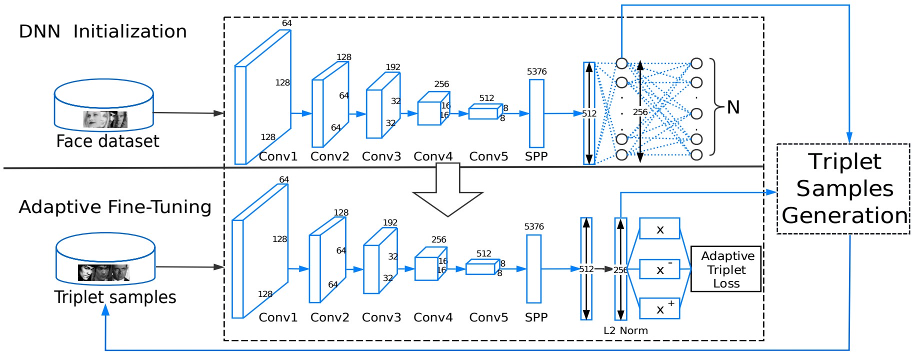

Fig. 1. The schematic diagram of the proposed training framework. Each **Conv** unit is followed by a 2 × 2 max-pooling layer with stride 2. **Conv1** to **Conv5** has $\{1, 2, 2, 3, 3\}$ convolutional layers, respectively. Each convolutional layer is followed by a **Maxout** layer. SPP denotes the spatial pyramid pooling layer. A dropout layer, whose ratio is 0.5, is added after each fully connected layer.

2. THE PROPOSED FRAMEWORK

In this section, we present the proposed framework in detail. The overview of the proposed framework is described in Section 2.1. Section 2.2 introduces the DNN initialization and Section 2.3 describes the process of adaptive fine-tuning.

2.1. Overview

In DNN, the discriminative loss function (such as the triplet loss function) is more appropriate for face recognition. However, such a loss function usually requires high training complexity. Therefore, in this paper, we investigate the DNN training process, which aims to improve the training efficiency while maintaining the recognition accuracy.

To be specific, an efficient and effective DNN training framework is proposed to deal with the training complexity problem mentioned above, which mainly contains two stages. Firstly, a specific DNN architecture with the softmax loss function is designed and the parameters are initialized. Secondly, on the basis of the trained DNN, we generate a set of high-quality triplet samples for the fine-tuning process. Then, we use these triplet samples to fine-tune the network, where an adaptive triplet loss function is developed for updating the parameters. The schematic diagram of the proposed training framework is shown in Fig. 1.

2.2. The DNN Initialization

In this section, we design the DNN architecture for obtaining the initial network parameters (see the top part of Fig. 1). More specifically, the DNN architecture is designed as follows. A fundamental DNN structure is firstly constructed similar to the VGG-16 model [9]. Then, the maxout [10] is used instead of the traditional ReLU [11] as the activation function due to its powerfully discriminative ability. Furthermore, we use a 3-level spatial pyramid pooling (SPP) layer introduced in [12] to combine the multi-scale features before the first fully-connected layer. Finally, the softmax loss is used as the loss function, which can be formulated as,

$$Loss = -\frac{1}{N}\sum_{n=1}^{N}\log(\hat{p}_{n,l_n}) \tag{1}$$

where N is the number of training samples, and $\hat{p}_{n,l_n} = \exp(x_{n,l_n})/\sum_{k=1}^{K}\exp(x_{n,k})$ denotes the probability for the n-th sample to be in the l_n-th class. K is the number of face classes. x_{n,l_n} defines the output of neuron l_n.

The training process stops when the DNN reaches the convergence condition. As we will see in Sec. 3, the DNN initialization based on the softmax loss obtains the effective face representation used for the generation of triplet samples, which can significantly reduce the computational complexity of the subsequent triplet loss-based DNN training process.

2.3. The Adaptive Fine-Tuning

Based on the trained DNN in the previous subsection, we take advantage of the triplet loss function for fine-tuning, which can effectively improve the discriminative capability of DNN, as shown in the bottom part of Fig. 1.

The triplet loss-based DNN [4] maps the input face images into a d-dimensional Euclidean space, which aims to shrink the distances of the faces belonging to the same person while enlarging the distances of the faces belonging to different persons. The performance of the triplet loss-based DNN largely depends on the high-quality training samples. A popular strategy is to directly generate triplet samples in the original image space. Typically, the conventional triplet loss-based DNN is trained to satisfy the following equation,

$$Dap_i + \alpha < Dan_i \quad \forall(x_i^a, x_i^p, x_i^n) \in \mathcal{S} \tag{2}$$

where Dap_i is the distance between the i-th intra-class positive pair (i.e., x_i^a and x_i^p). Dan_i is the distance between the i-th inter-class negative pair (i.e., x_i^a and x_i^n). α is a fixed margin that positive pairs (intra-class samples) and negative pairs (inter-class samples) should be kept. $\mathcal{S}$ is the training set. x_i^a, x_i^p and x_i^n define the anchor, positive and negative image in the i-th triplet sample, respectively.

However, such a strategy cannot effectively generate high-quality triplet samples to characterize the underline distributions of face variations (note that the inter-class and intra-class distances are computed in the image space). Furthermore, the conventional triplet loss enforces all the triplet samples to satisfy a fixed margin, which is not appropriate for DNN training, since the intra-class and inter-class face variations are complex. As a result, the convergence of DNN is slow due to the adopting of a fixed margin.

In this paper, the high-quality triplet samples can be generated according to the trained DNN in the first stage. Compared with the conventional method, the generation of the triplet samples is more easier, and these triplet samples can characterize the face variations more effectively (since the distances are computed in the feature space). To overcome the problem of slow convergence, we propose an adaptive triplet loss function. Therefore, based on the trained DNN in the first stage and the generated high-quality samples, we fine-tune the adaptive triplet loss-based DNN, whose loss function can be formulated as,

$$Loss = \sum_{i=1}^{N} max(err(\mathcal{S}_i), 0) \qquad (3)$$

$$err(\mathcal{S}_i) = Dap_i + \frac{D_{max}}{\sqrt[n]{Dap_i}}\tau - Dan_i \qquad (4)$$

where $\mathcal{S}_i$ is the i-th triplet sample in $\mathcal{S}$. N is the cardinality of training set, and D_{max} is the maximal distance among all the positive pairs. τ is a constant that scales the adaptive margin and $n > 1$ is a scalar to control the enlargement. From Eq. (4), $D_{max}\tau / \sqrt[n]{Dap_i}$ is enlarged when Dap_i is decreased, which effectively ensures the fast convergence of DNN.

In summary, the whole framework is given in Algorithm 1. Note that the initialization and fine-tuning stages used in the proposed framework is significantly different from the conventional ones, which usually refer to random initialization and parameter tuning using a different dataset. The fine-tuning in the proposed framework uses the same dataset as the initialization stage and effectively improves the discrimination by relying on an adaptive triplet loss function.

3. EXPERIMENTS

In this section, we explain the details of experiments and results. Section 3.1 describes the training configuration. Section 3.2 shows the comparison of the training complexity obtained by different methods. We analyze the performance on the tasks of face verification and face identification in Section 3.3 and Section 3.4, respectively.

Algorithm 1: An Efficient DNN Training Framework

```
Input: Training dataset $\mathcal{S}$, sampling interval $K$, and
       maximal epoch $T$
Output: The trained network parameters $\mathbf{W}$
Initialization: Randomize $\mathbf{W}$, $\mathcal{T} = \emptyset$, $t = 1$;
while not converge do // The DNN initialization
    for each training sample $x_i \in \mathcal{S}$ do
        Forward pass to obtain the face representation;
        Backpropagate to update the network parameters $\mathbf{W}$
          via Eq. (1);
    end
end
while $t < T$ do       // The adaptive fine-tuning
    if t mod K then // Generate triplet samples
        Generate positive pair $(x_i^a, x_i^p)$ according the current
          model parameters $\mathbf{W}$;
        Select the negative sample $x_i^n$ via Eq. (4);
        $\mathcal{T} = \mathcal{T} \cup (x_i^a, x_i^p, x_i^n)$;
    end
    for each triplet sample $(x_i^a, x_i^p, x_i^n) \in \mathcal{T}$ do
        Forward pass to obtain the face representation;
        Backpropagate to update the network parameters $\mathbf{W}$
          via Eq. (3);
    end
    $t \leftarrow t + 1$;
end
```

Table 1. The training time obtained by the different methods.

Method	Time (hours)
Softmax	**24**
Triplet	Failed (more than 168)
FaceNet [4]	more than 1000
Proposed method	96

3.1. Training Configuration

We train the proposed DNN using the CASIA-WebFace dataset [13] which contains about 0.5M images of 10,575 persons. We perform face detection on each image and obtain about 0.35M face images (with the size of 128×128 pixels) as the training set. Experiments are implemented on the Caffe platform [14] with a single Titan X GPU. The batch size is set to 128. The base learning rate, momentum and weight decay are set to 10^{-2}, 0.9 and 5×10^{-4}, respectively.

3.2. Training Complexity

In order to show the efficiency of the proposed framework, we compare the training time obtained by several different methods, including the softmax loss-based DNN, the triplet loss-based DNN [4] trained on the CASIA-WebFace, FaceNet [4] and the proposed framework, as shown in Table 1.

The softmax loss-based DNN spends about 24 hours to train the network. Compared with FaceNet, the proposed framework decreases the time complexity by a factor of 10. A triplet loss-based DNN fails to converge after training a week (168 hours) on our computer, while the proposed training framework converges after 96 hours, which benefits from the efficient initialization of the trained DNN and the high-quality triplet samples.

3.3. Face Verification

The LFW dataset [15] is used for evaluation under the unrestricted, labeled outside data protocol, which has 13,233 images of 5,749 individuals. The receiving operating characteristic (ROC) curves obtained by seven different methods are presented in Fig. 2.

Comparison with the traditional methods: We compare the model trained by the proposed DNN framework with the High-dim LBP [16] and TL-FaceVerification [17]. We can see that both the models trained by the proposed framework and the softmax loss-based DNN outperform the traditional methods in a reasonable margin, which shows the powerful discrimination ability of DNN.

Comparison with the existing deep learning methods: We compare the model trained by the proposed DNN with the existing deep learning methods (including DeepFace [3], DeepID [5], ConvNet-RBM [18]). The proposed DNN obtains better results than the softmax loss-based DNN and ConvNet-RBM. Compared with DeepFace [3], which is trained with 4.4M faces, the model obtained by the proposed framework achieves the similar performance with only 0.35M faces, which shows the effectiveness of the proposed training framework. Different from [3] and [18] that respectively use ensemble DNN and hybrid technologies to improve the performance, the proposed DNN achieves similar accuracy while it uses the face representation extracted by the single DNN only. This is because the adaptive triplet loss-based DNN can extract the powerful discriminative face representation by enlarging the intra-class distances and shrinking the inter-class distances effectively.

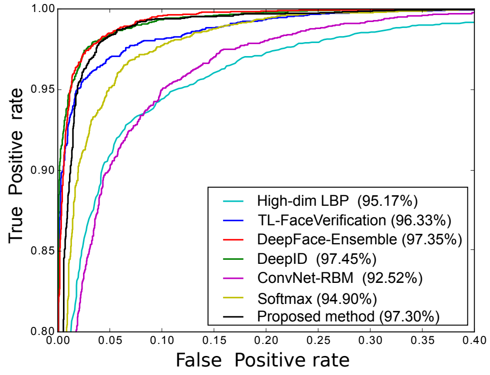

Fig. 2. Performance on the LFW dataset.

3.4. Face Identification

To further evaluate the effectiveness of the proposed training framework, we compare the model trained by the proposed framework with DAE [19], SPAE [20] and the softmax loss-based DNN on different face identification datasets. The correct rates obtained by all the four competing methods are presented in Table 2.

Table 2. The recognition accuracy on the different datasets.

Method	FERET	MultiPIE	FEI	Carmera12
DAE [19]	84.80%	82.50%	N/A	N/A
SPAE [20]	92.50%	91.40%	N/A	N/A
Softmax	99.96%	97.26%	98.72%	98.44%
Proposed method	**99.99%**	**99.31%**	**99.96%**	**99.52%**

Performance on FERET: The FERET dataset [21] contains 200 people with 9 different poses for each person. The dataset is used to evaluate the robust performance on pose variations. We use the same experimental settings as [20]. Due to the high-quality triplet samples, our trained model obtains higher recognition accuracy than DAE [19] and SPAE [20], and achieves slightly better performance than the softmax loss-based DNN on the dataset.

Performance on MultiPIE: The MultiPIE dataset [22] consists of images of 337 identities under different poses, expression and illumination conditions. Experimental results show that the model trained by the proposed framework performs better than the other three competing methods on this more challenging dataset, because it benefits from the powerful discriminative representation of the trained DNN.

Performance on FEI and Camera12: The FEI [23] and Camera12 [24] have 2,800 and 1,600 images of 100 identities, respectively. The former one contains the faces of 10 different poses, while the latter has the faces of 16 different poses. Our trained model achieves better performance than the softmax loss-based DNN on both datasets. This is due to the fact that the adaptive triplet loss function is used to fine-tune the DNN for obtaining the effective network parameters.

4. CONCLUSION

In this paper, a novel DNN training framework, which takes advantage of both the softmax loss and triplet loss functions, has been proposed for efficient face recognition. A specific softmax loss-based DNN architecture is designed to initialize the DNN. Based on it, we improve the discrimination capability of the DNN with a triplet loss function, where an adaptive margin is adopted. We have verified the effectiveness of the proposed DNN training framework on the LFW dataset and four different face datasets.

5. ACKNOWLEDGEMENTS

This work was supported by the National Natural Science Foundation of China under Grants 61571379, 61503315, U1605252, and 61472334, by the Natural Science Foundation of Fujian Province of China under Grant 2017J01127, and by the Fundamental Research Funds for the Central Universities under Grant 20720170045.

6. REFERENCES

[1] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton, "Imagenet classification with deep convolutional neural networks," in *NIPS*, 2012, pp. 1097–1105.

[2] P Jonathon Phillips, Matthew Q Hill, Jake A Swindle, and Alice J O'Toole, "Human and algorithm performance on the pasc face recognition challenge," in *BTAS*, 2015, pp. 1–8.

[3] Yaniv Taigman, Ming Yang, Marc'Aurelio Ranzato, and Lior Wolf, "Deepface: Closing the gap to human-level performance face verification," in *CVPR*, 2014, pp. 1701–1708.

[4] Florian Schroff, Dmitry Kalenichenko, and James Philbin, "Facenet: A unified embedding for face recognition and clustering," in *CVPR*, 2015, pp. 815–823.

[5] Yi Sun, Xiaogang Wang, and Xiaoou Tang, "Deep learning face representation from predicting 10,000 classes," in *CVPR*, 2014, pp. 1891–1898.

[6] Yi Sun, Yuheng Chen, Xiaogang Wang, and Xiaoou Tang, "Deep learning face representation by joint identification-verification," in *NIPS*, 2014, pp. 1988–1996.

[7] Yi Sun, Ding Liang, Xiaogang Wang, and Xiaoou Tang, "Deepid3: Face recognition with very deep neural networks," *arXiv preprint arXiv:1502.00873*, 2015.

[8] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *CVPR*, 2016, pp. 770–778.

[9] Karen Simonyan and Andrew Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[10] Ian J Goodfellow, David Warde-Farley, Mehdi Mirza, Aaron C Courville, and Yoshua Bengio, "Maxout networks.," in *ICML*, 2013, vol. 28, pp. 1319–1327.

[11] George E Dahl, Tara N Sainath, and Geoffrey E Hinton, "Improving deep neural networks for lvcsr using rectified linear units and dropout," in *ICASSP*, 2013, pp. 8609–8613.

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," in *ECCV*, 2014, pp. 346–361.

[13] Dong Yi, Zhen Lei, Shengcai Liao, and Stan Z Li, "Learning face representation from scratch," *arXiv preprint arXiv:1411.7923*, 2014.

[14] Yangqing Jia, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long, Ross Girshick, Sergio Guadarrama, and Trevor Darrell, "Caffe: Convolutional architecture for fast feature embedding," in *ACM MM*, 2014, pp. 675–678.

[15] Gary B. Huang, Manu Ramesh, Tamara Berg, and Erik Learned-Miller, "Labeled faces in the wild: A database for studying face recognition in unconstrained environments," Tech. Rep. 07-49, University of Massachusetts, Amherst, October 2007.

[16] Dong Chen, Xudong Cao, Fang Wen, and Jian Sun, "Blessing of dimensionality: High-dimensional feature and its efficient compression for face verification," in *CVPR*, 2013, pp. 3025–3032.

[17] Xudong Cao, David Wipf, Fang Wen, Genquan Duan, and Jian Sun, "A practical transfer learning algorithm for face verification," in *ICCV*, 2013, pp. 3208–3215.

[18] Yi Sun, Xiaogang Wang, and Xiaoou Tang, "Hybrid deep learning for face verification," in *ICCV*, 2013, pp. 1489–1496.

[19] Yoshua Bengio, "Learning deep architectures for ai," *FTML*, vol. 2, no. 1, pp. 1–127, 2009.

[20] Meina Kan, Shiguang Shan, Hong Chang, and Xilin Chen, "Stacked progressive auto-encoders (spae) for face recognition across poses," in *CVPR*, 2014, pp. 1883–1890.

[21] P Jonathon Phillips, Harry Wechsler, Jeffery Huang, and Patrick J Rauss, "The feret database and evaluation procedure for face-recognition algorithms," *IVC*, vol. 16, no. 5, pp. 295–306, 1998.

[22] Ralph Gross, Iain Matthews, Jeffrey Cohn, Takeo Kanade, and Simon Baker, "Multi-pie," *IVC*, vol. 28, no. 5, pp. 807–813, 2010.

[23] Carlos Eduardo Thomaz and Gilson Antonio Giraldi, "A new ranking method for principal components analysis and its application to face image analysis," *IVC*, vol. 28, no. 6, pp. 902–913, 2010.

[24] Rik Fransens, Christoph Strecha, and Luc Van Gool, "Parametric stereo for multi-pose face recognition and 3d-face modeling," in *AMFG*, 2005, pp. 109–124.