2020 International Conference on Machine Vision and Image Processing (MVIP)
Faculty of Engineering, College of Farabi, University of Tehran, Iran, 19 & 20 Feb. 2020

Using Siamese Networks with Transfer Learning for Face Recognition on Small-Samples Datasets

Mohsen Heidari
M.Sc. in Information Technology Engineering
Deep Learning Research Lab
Department of Computer Engineering
Faculty of Engineering, College of Farabi
University of Tehran, Iran
heidari.mohsen@ut.ac.ir

Kazim Fouladi-Ghaleh
Assistant Professor
Deep Learning Research Lab
Department of Computer Engineering
Faculty of Engineering, College of Farabi
University of Tehran, Iran
kfouladi@ut.ac.ir

***Abstract*—Nowadays, computer-based face recognition is a mature and reliable mechanism that is significantly used in many access control scenarios along with other biometric methods. Face recognition consists of two subtasks including Face Verification and Face Identification. By comparing a pair of images, Face Verification determines whether those images are related to one person or not; and Face Identification has to identify a specific face within a set of available faces in the database. There are many challenges in face recognition such as angle, illumination, pose, facial expression, noise, resolution, occlusion and the few number of one-class samples with several classes. In this paper, we are carrying out face recognition by utilizing transfer learning in a siamese network which consists of two similar CNNs. In the siamese network, a pair of two face images is given to the network as input, then the network extracts the features of this pair of images and finally, it determines whether the pair of images belongs to one person or not by using a similarity criterion. The results show that the proposed model is comparable with advanced models that are trained on datasets containing large numbers of samples. furthermore, it improves the accuracy of face recognition in comparison with methods which are trained using datasets with a few number of samples, and the mentioned accuracy is claimed to be 95.62% on LFW dataset.**

***Index Terms*—Face Recognition, Convolutional Neural Networks, Siamese Network, Transfer Learning, Small-Sample Dataset**

I. Introduction

Face Recognition is a perfect biometric method to identity confirmation and it is widely used in several domains such as military affairs, financial issues, public security and daily common life. Face recognition task can be divided into two subtasks including “Face Verification” and “Face Identification”. In each scenario, first, a set of known persons images are recorded in a gallery and the probe image is shown to the system when the experiment is to be performed. Face verification computes a one-by-one similarity index of gallery and probe images in order to determine whether those two images belong to one person or not; while face identification calculates a one-to-many similarity index to determine the specific identity of a probe face image. One of the face recognition challenges is the intra-class (intra-persona) varieties. This means that an identity may have changes in appearance that are made as the result of changes in lighting, facial expression, pose, make up, hair style, aging, and etc. The other challenge is inter-class similarities (between people or identities). For example, different identities may have similar appearances such as similarities between twins, relatives and even strangers. In recent years, with the emergence of deep learning, avalability of big training data, and improvements in hardwares and computational capabilities, “convolutional neural networks” (CNN) are widely used as one of the most important models by many researchers in various computer vision problems such as image classification [1], object detection [2], image retrieval [3], and etc. In order to simplify CNN training and improve image classification efficiency in public datasets, it is required to have sufficient and rich samples based on the number of available classes or categories. However, there is a small number of samples in some situations to make a real face recognition while there are large number of classes available; and this issue significantly decreases the face recognition efficiency.

In this paper, we are implementing face recognition using a “siamese network” [4] architecture which consists of two similar CNN networks- and transfer learning [5]. In the proposed model, a pair of face images is given to the network as input and the network determines whether the pair of images belong to one person or not by extracting the features of this pair of images and computing a similarity index.

This paper is organized in five sections. In section 2, the most recent and new related works are reviewed. In section 3, the proposed method is presented. The empirical results and evaluation is discussed in section 4 and finally, section 5 is dedicated to conclusion and recommendations for future works.

II. Related Works

In recent years, CNN-based algorithms have had significant achievements in face recognition and face verification applications [6]. Authors of [7] have used Weighted PCA-EFMNet deep learning feature extraction method to solve problems related to changes in expression, position, illumination and occlusion. In [8], a new approach (Auto-Encoder) is presented as a class sparsity supervised encoding (CSSE) for face

978-1-7281-6832-6/20/$31.00 ©2020 IEEE

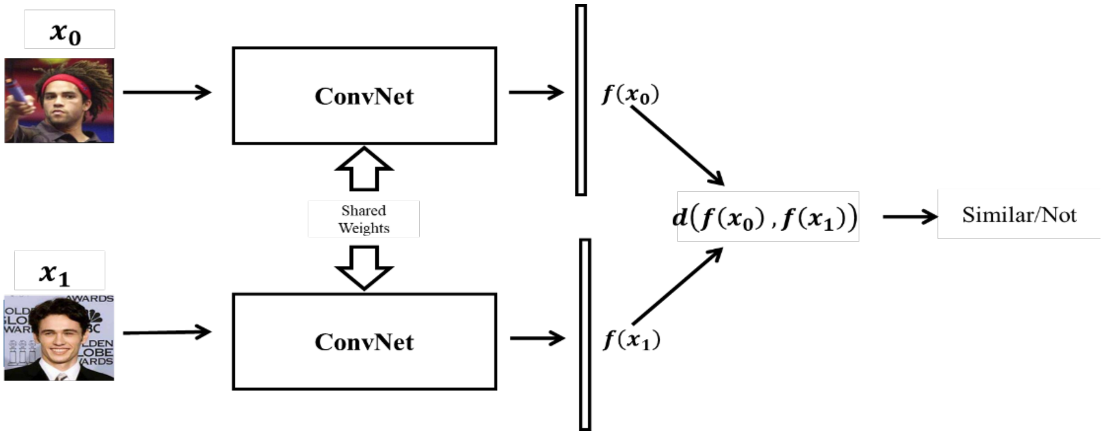

Fig. 1. Siamese network architecture for face recognition. $(X0, X1)$ are a pair of input images, $(f(X), f(X1))$ is the extracted feature vector for the pair of input images by using a convolutional neural network, $d(f(X0), f(X1))$ is the a similarity function which calculates the distance between output vectors of the two networks.

verification, in which feature representation is learned using supervised training data. Authors of [9] have proposed a part-based learning method for face verification, in which feature representation is extracted by convolutional fusion network (CFN). In [10], a double layer block (DLB)-based metric learning method is presented for better resolution of pair of face images and faster general procedure in face verification.

Jianming Zhang et al. [11] introduced a method by designing a new CNN and using it in siamese architecture, so that they could reach 94.8% accuracy in face recognition by training their model with small-samples dataset LFW.

The proposed method in this paper is similar to Jianming Zhang et al. [11], with siamese network architecture. We improve the accuracy up to 95.2% by using transfer learning and VGG-16 Model which is pre-trained on ImageNet dataset.

III. Face Recognition Using Siamese Architecture and Transfer Learning

A siamese network is an architecture with two similar parallel neural networks. The networks have identical configurations with identical weights and parameters, and the weights are shared between these two networks. Each network has a different input (image) and their outputs are combined to present some predictions. The fundamental idea of siamese networks is that they could learn the useful data descriptors that are used to compare subnetwork inputs. Figure 1 shows a siamese network architecture.

Transfer learning is a popular approach in machine learning specially in deep learning, in which pre-trained models are used as the beginning point of solving several computer vision problems. This method is mostly used when small number of data is available for modeling a new problem. Therefore, we can utilize deep learning models that are previously trained on big datasets and have common fundamentals with the new problem at hand, with the aim of building the transfer learning model on the gained knowledge from previous model.

In this paper, we use transfer learning to solve the face recognition problem. For this purpose, we utilize the pre-trained VGG-16 [12] as the convolutional neural network available in siamese architecture for feature extraction.

A. Pre-Trained VGG-16

VGG-16 is a convolutional neural network model proposed by A.Zisserman and K.Simonyan [12]. This model has reached to top-5 test 92.7% accuracy by training 1000-class ImageNet dataset with over 1.4 million labeled images. The architecture of this network consists of 13 convolutional layers with 3×3 kernel-size, 4 pooling layers with 2×2 kernel-size, and 3 fully connected (FC) layers that first two layers have 4096 neurons and the third fully connected layer has 1000 neurons.

B. Face Recognition Using Siamese-VGG

We use pre-trained-VGG-16 [12] in order to extract features in siamese architecture and then fine-tune it. For this purpose, as shown in figure 2, we first eliminate all VGG-16 fully connected layers and add our three fully connected layers each of which has 512 neurons along with ReLU activation function to the network. Then, we freeze all VGG-16 convolutional layers except block-5 layers that include three convolutional layers and one pooling layer, so that their weights will not change during training and the only changeable weights would be for block-5 and subsequently added fully connected layers. Since fully connected layers are randomly initialized, updating very big weights may be propagated throughout the network and destroy previously learned representations. We also set the network input size to $128 \times 128 \times 3$. To train the proposed siamese model, we use “contrastive loss Function” [13] to make an adaptive estimation of the model between Defaults (D) and Ground-Truth (y). Contrastive loss function tries to minimize the square Euclidean distance for similar pairs of images and maximize this distance for dissimilar pair of images, so that similar samples will get closer and dissimilar

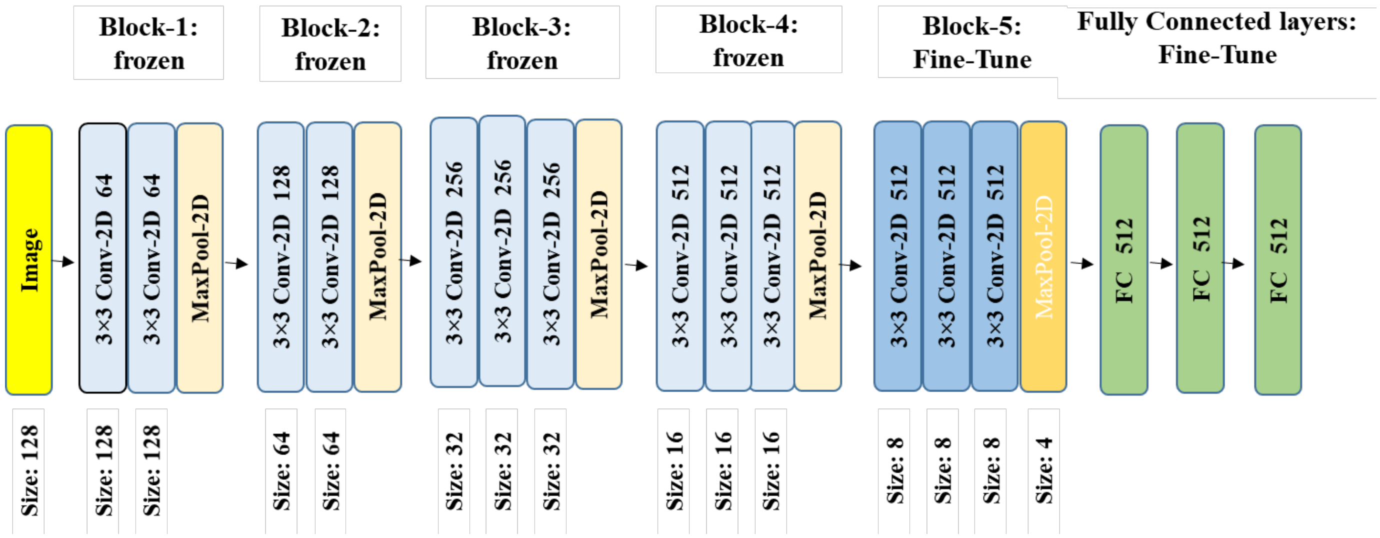

Fig. 2. VGG-16 network architecture with fully connected layers on top for fine-tuning

samples will get farer. Calculation of contrastive loss is as follows:

$$L = (1-y)(D)^2 + y[\max((m-D),0)]^2 \qquad (1)$$

where, y is the binary label that represents the similar ($y = 0$) or dissimilar ($y = 1$) pair of input images. m is a margin value that has a value more than zero. The existence of the margin shows that different pairs that are beyond this margin, do not affect the loss function. In this research, we assume the value of m to be equal to 1. D is the Euclidean distance of output vectors for both convolutional neural network available in siamese architecture (the value that the model predicts) and is calculated as follows:

$$d = ||f(x_0) - f(x_1)||_2 \qquad (2)$$

where, D is the Euclidean distance between f(x0) and f(x1) which are the outputs of the model for the pair of images (inputs) x_0 and x_1, and each of them includes a feature vector with 512 parameters. If the output vectors are sufficiently close ($d < 0.5$) then the model decides whether the input pair of images is similar ($D = 0$) or ($d \geq 0.5$) different ($D = 1$). Since the labels of each pair of images is initialized as 0 or 1, the predicted values of the model should also be 0 or 1, so that the comparison between the predicted value and the actual value would be more accurate and correct. The final predicted value of the model is calculated as follows:

$$D = \begin{cases} 0 & d < 0.5 \\ 1 & d \geq 0.5 \end{cases} \qquad (3)$$

IV. Experimental Results and Evaluation

Achieving high accuracy in recognition by using a small number of training samples for single-class samples is a difficult task for face recognition algorithms. LFW dataset has small-sized samples of face images for each person. Deep learning models need large numbers of samples to train network model for reaching high accuracies. It is difficult for them to reach an acceptable accuracy by using a small number of training samples. Nonetheless, we are conducting face recognition task using the small dataset of LFW.

Fig. 3. Some samples of LFW dataset images

LFW dataset is a database of facial images to study face recognition problem. This dataset includes over 13000 facial images that are collected from web pages. Each image is labeled with the person name. About 1680 individuals have two or more separate images in this dataset. Each facial image is a color image with size of 250 × 250 [14]. Figure 3 shows a part of this dataset.

A. The Results of Implementation of the Proposed Model on LFW Dataset

The inputs of our siamese network are the pairs of images and the labels. Therefore, it is necessary to prepare and generate training data that are compatible with expected siamese network structure. We generate the pair of images by using the suggested algorithm in [11]; so that if the pair of images belongs to one person, we mark (label) it as 0 and if it belongs to two different people, we mark (label) it as 1. The dataset includes about 15000 pairs of images that are generated

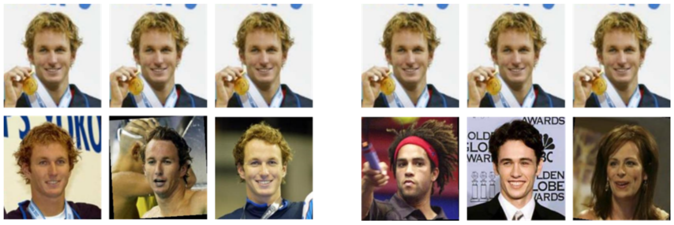

Fig. 4. Similar and Different Pairs Related to One Individual

TABLE I
Face Recognition Rate in Different Methods with LFW Dataset

Method	Face Recognition Rate (%)
DLB [10]	88.50
CFN+APEM [9]	87.50 ± 1.57
L-CSSE+KSRC [8]	92.02
SiameseFace1 [11]	94.80
Weighted Pca-Efmnet [7]	95.00 ± 0.71
Siamese-VGG (Ours)	**95.62 ± 0.42**
CosFace [15]	99.73

at the rate of 1:1. In this dataset, 60% of image pairs are used for training and the remained 40% are used for testing. Segmentation of dataset into two train and test subsets is done randomly. Figure 4 shows the generated similar and different sample pairs for an individual.

This experiment is carried out on a system with GeForce GTX 780 GPU and a 16 GB main memory. We used Keras as deep learning framework and carried out our experiment on LFW dataset. We assumed that batch sizes are equal to 32. We used ADAM Optimizer Function with $\alpha = 0.000005$ as learning rate. In the table I, the results of our experimet is represented in comparison with other face recognition algorithms. As clarified below, our model obtained the best accuracy compared to other models that use small training datasets alike us.

In Table I, CosFace is a deep learning method proposed by Hao Wang et al. [15]. This algorithm is trained by data that are different from LFW dataset. This algorithm is trained by 5 million facial images and has gained the best accuracy. As it is clear, our algorithm has less accuracy than the foresaid one, but it has the best accuracy compared to other methods using small dataset to train their model.

In Table I, CosFace is a deep learning method proposed by Hao Wang et al. [15]. This algorithm is trained by data that are different from LFW dataset. This algorithm is trained by 5 million facial images and has gained the best accuracy. As it is clear, our algorithm has less accuracy than the foresaid one, but it has the best accuracy compared to other methods using small dataset to train their model.

V. Conclusion and Future Works

In this paper, we carried out a face recognition task by using a siamese network architecture which consists of two similar CNNs and also applying transfer learning from VGG-16 model. For this purpose, we exploited a VGG network model pre-trained on ImageNet dataset to extract features from images along with Euclidean distance to calculate the similarity level. We also conducted network training with contrastive loss function to minimize the similarity between pairs of images related to one person and maximize the similarity between pairs of images for different individuals. The results show that the proposed model has been capable of improving the accuracy rate compared to other similar methods that are trained on datasets with small number of samples. As a recommendation to continue this work in the future, we suggest using other CNN networks in siamese network architecture; particularly those convolutional networks that are capable of extracting high level features in addition to low level ones, perfectly. Moreover, using "triplet loss" can be an appropriate alternative for Siamese architecture which compares simultaneously positive and negative pairs. Data augmentation methods can also be used as effective methods for small-samples datasets.

References

[1] W. Rawat and Z. Wang, Deep convolutional neural networks for image classification: A comprehensive review, Neural Comput., vol. 29, no. 9, pp. 2352-2449, 2017.

[2] J. Han, D. Zhang, G. Cheng, N. Liu, and D. Xu, Advanced deep-learning techniques for salient and category-specific object detection: a survey, IEEE Signal Process. Mag., vol. 35, no. 1, pp. 84-100, 2018.

[3] L. Zheng, Y. Yang, and Q. Tian, SIFT meets CNN: A decade survey of instance retrieval, IEEE Trans. Pattern Anal. Mach. Intell., vol. 40, no. 5, pp. 1224-1244, 2017.

[4] S. Chopra, R. Hadsell, and Y. LeCun, Learning a similarity metric discriminatively, with application to face verification, in CVPR (1), 2005, pp. 539-546.

[5] J. Qiu, Q. Wu, G. Ding, Y. Xu, and S. Feng, A survey of machine learning for big data processing, EURASIP J. Adv. Signal Process., vol. 2016, no. 1, p. 67, 2016.

[6] M. Wang and W. Deng, Deep face recognition: A survey, arXiv Prepr. arXiv1804.06655, 2018.

[7] B. Ameur, M. Belahcene, S. Masmoudi, and A. Ben Hamida, Weighted PCA-EFMNet: A deep learning network for Face Verification in the Wild, in 2018 4th International Conference on Advanced Technologies for Signal and Image Processing (ATSIP), 2018, pp. 1-6.

[8] A. Majumdar, R. Singh, and M. Vatsa, Face verification via class sparsity based supervised encoding, IEEE Trans. Pattern Anal. Mach. Intell., vol. 39, no. 6, pp. 1273-1280, 2016.

[9] C. Xiong, L. Liu, X. Zhao, S. Yan, and T.-K. Kim, Convolutional fusion network for face verification in the wild, IEEE Trans. Circuits Syst. Video Technol., vol. 26, no. 3, pp. 517-528, 2015.

[10] S.-C. Chong, A. B. J. Teoh, and T.-S. Ong, Unconstrained face verification with a dual-layer block-based metric learning, Multimed. Tools Appl., vol. 76, no. 2, pp. 1703-1719, 2017.

[11] J. Zhang, X. Jin, Y. Liu, A. K. Sangaiah, and J. Wang, Small Sample Face Recognition Algorithm Based on Novel Siamese Network., J. Inf. Process. Syst., vol. 14, no. 6, 2018.

[12] K. Simonyan and A. Zisserman, Very deep convolutional networks for large-scale image recognition, arXiv Prepr. arXiv1409.1556, 2014.

[13] R. Hadsell, S. Chopra, and Y. LeCun, Dimensionality reduction by learning an invariant mapping, in 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR06), 2006, vol. 2, pp. 1735-1742.

[14] G. B. Huang, M. Mattar, T. Berg, and E. Learned-Miller, Labeled faces in the wild: A database forstudying face recognition in unconstrained environments, 2008.

[15] H. Wang et al., Cosface: Large margin cosine loss for deep face recognition, in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 5265-5274.