Dual Minkowski Loss for Face Verification of Convolutional Network

Wang Dandan[1, 2, a] and Chen Yan[1, b]

[1]School of Maritime Economics and Management of Dalian Maritime University, Dalian Maritime University, Dalian 116026, China
[2]City Institute, Dalian University of Technology, Dalian 116600, China

Abstract: Despite face recognition and verification have achieved great success in recent years, these achievements are experimental results on fixed data sets. Implementing these outstanding technologies in the field of undeveloped data sets presents serious challenges. We adopt three state-of-the-art pre-trained models on an entire new dataset University Test System Database (UTSD), however the results are far from satisfactory. Therefore, two methods are adopted to solve this problem. The first way is data augmentation including horizontal flipping, cropping and RGB channels transform, which can solve the imbalance of label pairs. The second way is the combination of Manhattan Distance and Euclidean Distance, we call it Dual Minkowski Loss (DML). Through the implementation of photo augmentation and innovative method on UTSD, the accuracy of face verification has been significantly improved, achieving the best 99.3%.

1 Introduction

Nowadays, a large variety of photos, videos and text scripts were applied to deep learning. The large scale of datasets contributes a lot to the improvement of recognition accuracy, such as Labeled Faces in the Wild (LFW), YouTube Faces (YTF), CASIA-WebFace, and CAS-PEAL et al [6]. Thanks to the valuable datasets that models and algorithms have been greatly developed and promoted. The recent state-of-the-art face recognition models such as DeepFace have achieved an accuracy of 97.35% on LFW dataset, the later published technique FaceNet refreshed the latest record and pushed the precision to the highest 99.63%. So far, FaceNet is considered as a baseline for face verification and recognition [1, 2].

Due to the existence of an intermediate bottleneck layer, the operation speed and accuracy of the convolution neural network have been greatly affected. By abandoning the bottleneck layer and choosing optimized embedding, FaceNet has an obvious advantage in image processing. Therefore, FaceNet model is transferred to our new database (UTSD) for learning, three pre-trained models are adopted as initial input. Comparing the performances on the novel database, several improved measurements are adopted for the effect of the model and accuracy of prediction.

2 Related Work Details

2.1. Face Alignment and Label Generation

Photos in UTSD are 250*250 pixels, in order to put all the images into Inception Network for training, we change the whole database of pictures cutting into 160*160 shape through programs [3, 4]. Then photos of the same person are placed in a fixed folder and each of them is renamed.

Once the photos have been processed, the following task become straight-forward. For the same identities in a folder, pictures are randomly formed into a set as positive sample tag pairs. As for the negative sample tag pairs, we do not assemble all of the distinct photos. Because if all the possible negative tag pairs are generated, the result in triplets is more inclined to get satisfied. Taking this into consideration, only the first and second pictures from different folders are selected as negative ones.

As shown in Fig. 1, the output figures are Euclidean space distance between two pairs of faces，which from either the same person or two different ones. The compared photos are sampled respectively from identity card and certificate of identification. The figure of 0.0 represents the two pictures are of the same person, and a figure of 5.0 are completely distinct. You can see that a threshold of 0.98 would be fine to distinguish the pairs of person.

* Corresponding author:[a]wangdandan0707@dlut.edu.cn，[b]chenyan_dlmu@163.com

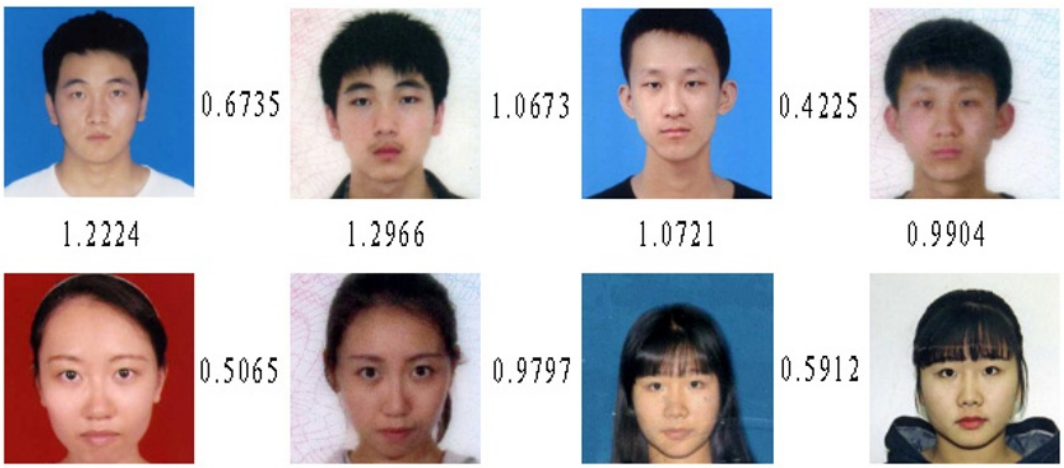

Fig. 1 Distances of Euclidean Distance between pairs of faces

2.2. Pre-trained Network Architecture

Although AlexNet has showed that sufficient large and deep Convolution Networks trained by standard back-propagation can achieve an excellent recognition accuracy, millions of parameters have to be trained and plenty of hours need to be consumed under this condition. The more important thing is the hard solving problem of bottleneck layer, which reducing the performance of the whole network [10].

Avoiding extra time consumption and aiming to have a good performance start, we adopt 3 pre-trained dataset models for training. The network architectures are selected including CASIA-WebFace dataset which contains more than 10,000 persons and 500,000 photos, VGGFace2 which has 9131 identities and 3.31 million number of images, and MS-Celeb-1M consists of 1 million famous people and 100 photos of each. Each of the models is trained for a great deal of time and all of them have achieved extraordinary accuracy. The whole datasets are well trained in neural network. Therefore, choosing 3 pre-trained dataset models as the initial input of CNN is a pretty good choice for efficient training.

2.3. Database and data augmentation

1280 different identities of boy students and girl students among the age 17 to 21are stored in University Test System Database(UTSD), and nearly 3980 sample photos are contained including identity card, certificate of identification and instant photos. For the sake of comparing effects of face verification, data augmentation methods are adopted to enlarge the amount of database. The first form of data augmentation consists of generating image translations and horizontal flipping. Four corners are cropped on the original and flipping images. The second way of data augmentation consists of altering the intensities of the RGB channels in training photos. We perform PCA on the set of RGB pixel values throughout the whole training set. In order not to affect the accuracy of face verification, we slightly changed the RGB channels. The two steps of procedure making photos expand to 12-15 times but size of pixels are maintained.

3 Model Design

The models we chosen are treated as a black box, the main works lie in the end-to-end learning of the whole system. To achieve face verification we employ a novel method. Under this circumstance, the Triplet Loss Function is modified according to the performance of real data. We calculate the distance between two pictures by using the combination of Manhattan Distance and Euclidean Distance rather than Euclidean Distance only. The main idea of face verification is to minimize the distance of the same category (Anchor and positive) and maximize the distinct (Anchor and negative). In order to have fast convergence, hard positive should be seriously selected.

4 The Dual Minkowski Loss

Triplet Loss is found that there is still a certain distance from the higher accuracy in practice. As a result we improve the method through the combination of Manhattan Distance and Euclidean Distance, which is called Dual Minkowski Loss (DML).

In DML, we want to meet the following three conditions.

$$\| f(x_i^a) - f(x_i^p) \|_2^2 + \alpha_1 < \| f(x_i^a) - f(x_i^n) \|_2^2 \quad (1)$$

$$\| f(x_i^a) - f(x_i^p) \| + \alpha_2 < \| f(x_i^a) - f(x_i^n) \| \quad (2)$$

$$\forall f(x_i^a),\ f(x_i^p),\ f(x_i^n) \in \Phi \quad (3)$$

α_1 is a margin enforced by the positive and negative pairs by using the Euclidean Distance. Similarly, α_2 is a margin by using Manhattan Distance. Φ is the generated possible triplets assemble, including positive and anchor tag pairs, negative and anchor tag pairs, the cardinality of Φ is N.

For the first condition, which is a classic expression of Euclidean Distance in [1], the second is the expression of Manhattan Distance. Combining of the first and second

conditions is the method we found to be more effective in improving experimental results.

The loss function is represented as follows, one condition will be needed.

$$L=\sum_{i}^{N}\beta[\| f(x_i^a)-f(x_i^p)\|_2^2-\| f(x_i^a)-f(x_i^n)\|_2^2+\alpha_1]_+ +\chi[\| f(x_i^a)-f(x_i^p)\| -\| f(x_i^a)-f(x_i^n)\| +\alpha_2]_+ \quad (1)$$

$$\beta+\chi=1 \quad (2)$$

β and χ are the weights of adjustment, they can get distinct values according to the effect of experiment in iteration process, and they follow the equation that the sum of β and χ equals 1 . The loss function is a weight control combination of Manhattan Distance and Euclidean Distance.

5 Training

In the process of training convolutional networks, we use the basic Stochastic Gradient Descent with standard back propagation [5, 7, 8]. The learning rate is set to 0.06 at the beginning and gradually decreases as the number of iterations increasing. The model is trained on a CPU running for nearly a week, the loss dramatically drop down when running at 120 hours of training. The margin α_1 is set to 0.2 and α_2 is 0.08. We fine tune the parameters and get good performance when value of β is around 0.375 and χ around 0.624.

6 Experiments

We evaluate our method on three Pre-trained datasets with Dual Minkowski Loss Function comparing to the FaceNet with Triplet Loss Function. As shown in Table 1, the latter method trained on CASIA-WebFace, VGGFace2 and MS-Celeb-1M datasets is fine, the accuracy is around 0.875, 0.893 and 0.885. Surprisingly, validation rates are quite dissatisfied.

We analyze the distribution and characteristics of the data in UTSD, trying to find out the reasons affecting validation rate. The out-off-balance distribution of data in the labeled pairs' dataset may be one of the reasons accounting for this phenomenon. Two effective methods are applied to practice in order to solve this problem. The first solution is photo augmentation, because of the limitation of data acquisition, there is no more than 5 photos of the same identity including identity card, certificate of identification and instant photos. The generated negative labels are obvious more than the positive ones, giving rise to imbalance of data distribution. Therefore, we employ horizontal flipping, cropping and RGB channels altering. Another solution contributed to this is Dual Minkowski Loss Function.

As shown in Table 1, the accuracy of the three models shows excellent results. Not only the accuracy promotes, but the Validation Rates are also outstanding.

Table 1 Two training methods on three pre-trained architectures comparing

Numble	Pre-trained dataset	DataSet	Model	Accuracy	Validation Rate
1	CASIA-WebFace	UTSD	FaceNet using Triplet Loss	0.875+-0.006	0.03455+-0.18263
2	VGGFace2	UTSD	FaceNet using Triplet Loss	0.893+-0.003	0.07577+-0.3649
3	MS-Celeb-1M	UTSD	FaceNet using Triplet Loss	0.885+-0.005	0.05546+-0.3210
4	CASIA-WebFace	UTSD	Dual Minkowski Loss	0.974+-0.005	0.9150+-0.02036
5	VGGFace2	UTSD	Dual Minkowski Loss	0.982+-0.005	0.9203+-0.01382
6	MS-Celeb-1M	UTSD	Dual Minkowski Loss	0.993+-0.003	0.9256+-0.02178

7 Conclusions

Our work demonstrates that using a combination of Manhattan Distance and Euclidean Distance loss function which we called Dual Minkowski Loss(DML) can effectively learn many examples to overcome the drawbacks and limitations of previous method. Experiments show that DML can significantly improve the predicted accuracy on University Test System Database (UTSD) compared to Triplet Loss. As a result, our proposed method achieves the state-of-the-art results on the fixed data sets.

References

1. Florian Schroff, Dmitry Kalenichenko, James Philbin. FaceNet: A Unified Embedding for Face Recognition and Clustering, IEEE Computer Society Conference

on Computer Vision and Pattern Recognition, in: Proc. IEEE Conf. Comput. Vis.Pattern Recognit. (CVPR), Jun. 2015, pp. 815-823.

2. Y. Taigman, M. Yang, M. Ranzato, and L. Wolf, Deepface: Closing the gap to human-level performance in face verification, in Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on. IEEE, 2014, pp. 1701-1708.
3. M.Lin, Q. Chen, S.Yan. Network in network. CoRR, abs/1312. 4400, 2013, 2-6.
4. [4]C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions.In CVPR, 2015.
5. A. Kumar, A. C. Berg, P.N. Belhumeur, G. Hinton. ImageNet classification with deep convolutional neural networks. In ANIPS, 202, 1-4
6. K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In ICLR, 2015
7. M. D. Zeiler and R. Fergus. Visualizing and understanding convolutional neural networks. In ECCV, 2014.
8. T. Vatanen, T. Raiko, H. Valpola, and Y. LeCun. Pushing stochastic gradient towards second-order methods–back propagation learning with transformations in nonlinearities. In Neural Information Processing, 2013
9. S. Ren, K. He, R. Girshick, and J. Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. In NIPS,2015.
10. G. Montufar, R. Pascanu, K. Cho, and Y. Bengio. On the number of linear regions of deep neural networks. In NIPS, 2014