Face recognition based on improved residual neural network

CHEN Zhenzhou[1], DING Pengcheng[1]

1. School of Electromechanical Engineering, Guangdong University of Technology, Guangzhou 510006

Abstract: For traditional face recognition algorithm based on convolution neural network in the insufficient light, side face of 45 ° environment, for large-scale face recognition accuracy is not high, and when the network layer increases to a certain layer, the performance of the network tends to saturation, continue to increase the network layer instead lead to performance problems. An improved neural network architecture combined with deep residual neural network and ArcFace Loss is proposed, namely, ArcFace Loss replaces the Loss function Softmax of deep residual neural network. Experimental results show that the improved residual neural network has an accuracy of 97.7% in face recognition on LFW data set, which is higher than the traditional and improved deep learning algorithm.

Key Words: face recognition, ResNet, performance, Loss function, accuracy

1 INTRODUCTION

With the continuous research of deep convolutional neural network[1](DCNN), face recognition has made breakthrough progress, and it has a very positive promotion effect in the fields of security, attendance and other fields. Face feature extraction is mainly based on the model structure of CNN. Among them, using deep convolutional neural networks such as Alex-Net[2], VGG[3] to extract image features, and then using classification models(such as SVM) to classify and detect the extracted facial features, from the experience, The depth of the network is crucial to the performance of the model. When the number of network layers is increased, the network can extract more complex feature patterns, so theoretically, better results can be obtained when the model is deeper. However, when the number of network layers increases, the accuracy of the network appears saturated, that is, the phenomenon of gradient dispersion or gradient explosion. Although there are various solutions, such as pre-training with fine tuning[3], gradient clipping[4], weights regularization[5],using ReLu[4] and other activation functions, batchnorm, but only It can alleviate the problem of gradient disappearance and explosion to a certain extent, and does not fundamentally solve the degradation phenomenon. The deep residual network ResNet proposed by He Kaiming[6] and others can deal with the degradation problem well.

The core tasks of face recognition include face verification and face recognition. However, under the supervision of the Softmax cost function of the deep convolutional neural network in the traditional sense, the model studied usually lacks sufficient discriminability. In order to solve this problem, a series of loss functions have been proposed, such as Center Loss[7], L-Softmax[8], A-Softmax[9], ArcFace[10], etc. All these improved algorithms are based on a core idea: Enhance inter-class differences and reduce intra-class differences.

This paper deals with the problem of face recognition from a new perspective, that is, making some adjustments based on ResNet, combined with ResNet deep network, can extract more facial features and ArcFace can enhance without causing network performance degradation. The advantage of inter-class differences achieves high speed and high accuracy face recognition.

2 Residual neural network principle and loss function

2.1 ResNet overall network structure, as shown in Figure 1.

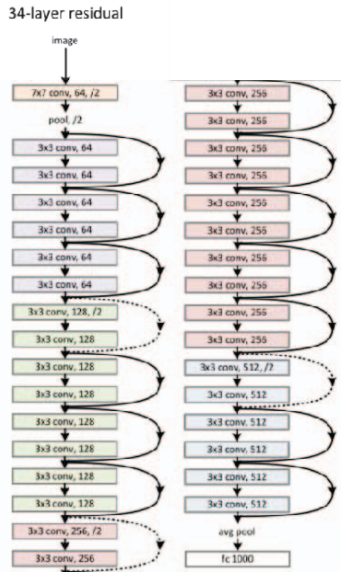

Fig.1 ResNet

Compared to ordinary convolutional neural networks, ResNet has many bypass branches that connect inputs directly to subsequent layers, allowing subsequent layers to learn the residuals directly. This structure is called shortcut or skip connection[11]. The traditional volume base or fully connected layer has more or less information loss and damage when information is transmitted, and ResNet solves this problem to some extent by directly bypassing the input information to the output. The integrity of the information, the entire network only needs to learn the part of the input and output differences, simplifying the learning objectives and difficulty.

In the design of deep residual network, it is usually to pursue a design method of “striving for simplicity”. It is just to deepen the network. All convolutional layers use almost 3×3 convolution kernels, and no design is designed in the

978-1-7281-0106-4/19/$31.00 ©2019 IEEE

hidden layer. The connection layer does not consider any DropOut[12] mechanism during training.

2.2 Residual block structure

As shown in Figure 2, the two-layer residual learning unit has the following expression:

$$F=W_2\sigma(W_1X) \tag{1}$$

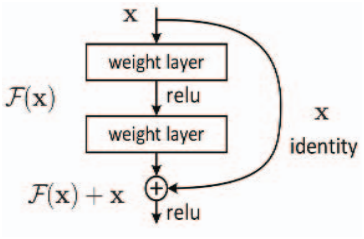

Fig.2 Residual block

Where σ represents the nonlinear activation function ReLU. Then get the output y through a shortcut and the second activation function ReLU:

$$y=F(x,\{W_i\})+X \tag{2}$$

When you need to change the input and output dimensions, you can make a linear transformation Ws for X in the shortcut, as follows:

$$y=F(x,\{W_i\})+W_sX \tag{3}$$

Considering the cost of the calculation, the residual block is calculated and optimized, and the two-layer residual learning unit is replaced with a three-layer residual learning unit. As shown in Fig. 3, two 3×3 convolution layers are replaced by 1×1+3×3+1×1. The middle 3×3 convolutional layer in the new structure is first reduced by a reduced dimensional 1×1 layer and then restored under another 1×1 convolutional layer, which maintains accuracy and reduces computation. The amount, which is equivalent to reducing the amount of parameters for the same number of layers, can be extended to a deeper model.

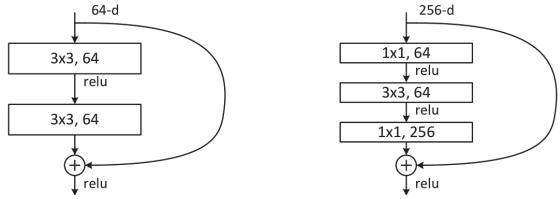

Fig.3 Two and three-layer residual unit

2.3 Loss function ArcFace

The formula for Softmax Loss is as follows:

$$L_1=-\frac{1}{m}\sum_{i=1}^{m}\log\frac{e^{W_{y_i}^Tx_i+b_{y_i}}}{\sum_{j=1}^{n}e^{W_{y_j}^Tx_j+b_j}} \tag{4}$$

Softmax Loss does not explicitly optimize features to positive samples with higher similarity, negative samples can have lower similarity, that is, does not expand decision boundaries, so the offset b_j is then set to 0, then the weight and the inner product of the input can be expressed by the following formula:

$$W_{y_i}^Tx_j=\|W_j\|\bullet\|x_j\|\cos\theta_j \tag{5}$$

The L2 regularization process is such that ‖Wj‖=1, that is, each value in the W_j vector is divided by the modulus of W_j, respectively, to obtain a new W_j, and the modulus of the new W_j is 1. At the same time, the angle interval m is introduced, that is, the angle is expanded by m times, and the following formula of A-Softmax Loss is obtained:

$$L_{ang}=\frac{1}{N}\sum_i-\log\left(\frac{e^{\|x_i\|\cos(m\theta_{yi},i)}}{e^{\|x_i\|\cos(m\theta_{yi},i)}+\sum_{j\neq y_i}e^{\|x_i\|\cos(m\theta_{yi},i)}}\right) \tag{6}$$

Improve the way angle margins are added in the Loss function:

$$L_{lsF}=\frac{1}{N}\sum_i-\log\left(\frac{e^{s(\cos(m\theta_{yi},i)-m)}}{e^{s(\cos(m\theta_{yi},i)-m)}+\sum_{j\neq y_i}e^{s(\cos(m\theta_{yi},i)-m)}}\right) \tag{7}$$

Compared with CosineFace, the margin of arc space of ArcFace corresponds to the arc distance on the hypersphere. There is a clearer geometric interpretation, which not only compresses the feature area, but also corresponds to the geodesic distance on the hypersphere. Figure 4 shows the geometric interpretation of ArcFace, and Figure 5 shows the comparison of the four Loss methods.

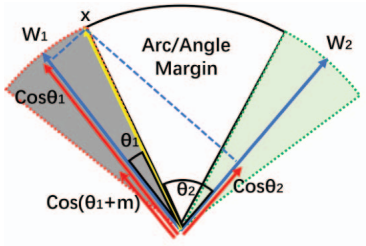

Fig.4 The geometry of ArcFace

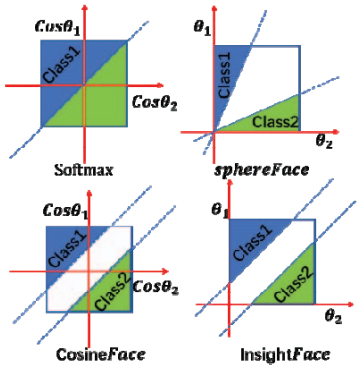

Fig.5 Comparison of four Loss function

3 Experimental preparation and results

3.1 Training methods

First, use ResNet training to extract facial features, then replace the final Softmax Loss of ResNet with ArcFace Loss, separate the facial features of different people, and

finally calculate the Euclidean distance between the newly acquired facial features and the learned model.

In order to accelerate the training of ArcFace Loss, the traditional face recognition model is trained by the traditional Softmax. Because of the strong supervised nature of the classification signal, the model will be quickly fitted, then the top layer classification layer will be removed, and the model will be fine-tuned with ArcFace Loss.

After experimental comparison, the squared Euclidean Distance is not as good as the non-squared Euclidean Distance, so as Equation 8:

$$\| x_i^a - x_i^p \|_2^2 + \alpha < \| x_i^a - x_i^n \|_2^2 \quad (8.1)$$

$$\| x_i^a - x_i^p \|_2 + \alpha < \| x_i^a - x_i^n \|_2 \quad (8.2)$$

3.2 Data set: training data set and test data set

CAS-PEAL[13]: A database of 99450 face images containing 1040 volunteers. The database covers changes in features such as gestures, expressions, decorations, lighting, background, distance, and time.

LFW[14]: The LFW data set was created to investigate the face recognition problem in an unrestricted environment. This dataset contains more than 13000 face images (all from the internet, not the lab environment), each face is labeled with a person's name, of which about 1680 people have more than two faces.

3.3 Data cleaning[15]

Get the center of mass of each person's identity and sort each person's picture by distance from the centroid. Images that are too far from the center are automatically removed. The image at the edge of the threshold is then manually confirmed, and the final data set contains approximately 23000 images of 9500 people.

3.4 Network settings

Using the MxNet architecture, CAS-PEAL was first used as training data, and various network settings were verified using the Softmax loss function. A GTX1080Ti, the learning rate starts at 0.1, is reduced by 10 times at 100k, 140000 and 160000, the total iteration is 200000, the momentum is 0.9, and the weight attenuation rate is 0.0005.

3.5 experiment and result analysis

First, the cleaned CAS-PEAL data set was used for training, and then the LFW data set was used for testing. The experimental comparison results are shown in Table 1.

Table1. Performance comparison of various network structures

Network@Loss	LFW(%)
AleNet@Softmax	95.81
VGGNet@Softmax	94.4
ResNet@Softmax	96.38
ResNet@ArcFace	97.7

3.6 experiment was conducted in ORL Face Database.

ORL Face Database is composed of 10 images of 40 people with 112 x 92 pixel (figure 6), some of which were taken at different times. The details of their facial expressions vary to varying degrees, for example, whether they laugh or not, the eyes are open or closed, and they wear or don't wear glasses; There is also considerable variation in facial posture, with depth and plane rotation up to 20%. There are also up to 10 percent variations in face size. In order to reduce measurement, each face image is processed with third order wavelet transform, and the low-frequency part is taken to reduce each image to a 14x12 pixel image.

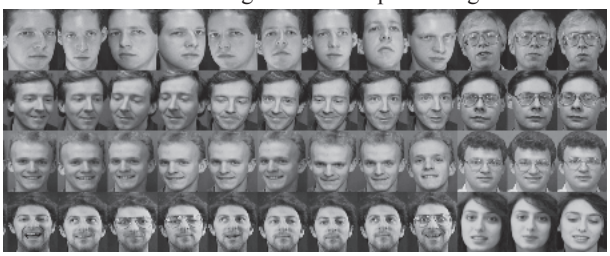

Fig.6 the of ORL Face Datebase

In the experiment of this paper, the number of different training samples was tested, and the number of samples selected for each type was the same. After determining the number of training samples, the training samples were randomly selected, and all the remaining images were the test samples. Four hundred images of 40 people were randomly divided into two groups, a training sample and a test sample. The number of training samples in each test was 1~9. Different methods in the same experiment all select the same training samples and test samples.

In order to eliminate the randomness of single selected samples, each test in this paper was carried out for 30 times, and then the average recognition rate was taken. Face recognition experiments were carried out based on four different combinations, and then the results were compared. The experimental results are shown in figure 7.

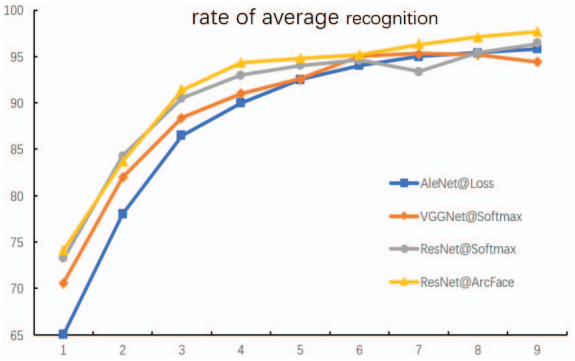

Fig.7 the rate of average recognition

4 Conclusion

In this paper, the loss function is studied and explored. The problem of network performance degradation occurs when the traditional deep convolutional neural network is too deep in the network layer, and the face recognition rate is not enough under the light and the side face 45°[17]. In the high case, an improved deep residual neural network is

proposed. After training, ArcFace Loss is used to replace Softmax in ResNet to fine-tune, expand the categories of different facial features, and finally calculate the newly acquired people. The Euclidean distance between the face feature and the learned model,

By combining their respective advantages, the experimental results show that:

(1) First, use a small number of trained Softmax, and then use a larger number of data sets to fine-tune the accuracy of the ArcFace Loss, and also demonstrate the effectiveness of the two-step training mechanism, the training speed is faster than direct training ArcFace Loss a lot of;

(2) Accuracy can also be improved by performing ArcFace fine-tuning on the same dataset, indicating that local elevations contribute to the improvement of the global model;

(3) The method of local metric learning can complement the global hypersphere metric learning method.

REFERENCES

[1] Goodfellow I, Bengio Y, Courville A. Deep Learning[M]. Oxford: Academic Press, 2019: 8-15.

[2] Krizhevsky A, Sutskever I, Hinton G E. ImageNet Classification with Deep Convolutional Neural Networks[C]. International Conference on Neural Information Processing Systems. 2012: 1097-1105.

[3] Simonyan K, Zisserman A. Very Deep Convolutional Networks for Large-Scale Image Recognition[J]. Computer Science. 2014.

[4] He K M, Zhang X Y, Ren S Q, et al. Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification[J]. IEEE International Conference on Computer Vision. 2015: 1026-1034.

[5] Zhang C Y, Bengio S, Hardt M, et al. Understanding deep learning requires rethinking generalization[J]. International Conference on Learning Representations. 2017.

[6] He K M, Zhang X Y, Sun S Q, et al. Deep Residual Learning for Image Recognition[C]. IEEE Conference on Computer Vision and Pattern Recognition. 2016

[7] Wen Y D, Zhang K P, Li Z F, et al. A Discriminative Feature Learning Approach for Deep Face Recognition[J]. European Conference on Computer Vision. 2016.

[8] Liu W Y, Wen Y D, Yu Z D, et al. Large-Margin Softmax Loss for Convolutional Neural Networks[J]. International Conference on Machine Learning. 2016.

[9] Liu W Y, Wen Y D, Yu Z D, et al. SphereFace: Deep Hypersphere Embedding for Face Recognition[J]. IEEE Conference on Computer Vision and Pattern Recognition. 2017.

[10] Deng J K, Guo J, Zafeiriou S. ArcFace: Additive Angular Margin Loss for Deep Face Recognition[J]. arXiv. 2018: 1801-7698.

[11] Szegedy C, Ioffe S, Vanhoucke V, et al. Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning[J]. arXiv. 2017: 1602-7261.

[12] Srivastava N, Hinton G, Krizhevsky A, et al. Dropout: A Simple Way to Prevent Neural Networks from Overfitting[J]. Journal of Machine Learning Research. 2014: 11-15.

[13] Yi D, Lei Z, Liao S C, et al. Learning Face Representation from Scratch[J]. Computer Science. 2014.

[14] Huang G B, Ramesh M, Berg T, et al. Labeled Faces in the Wild: A Database for Studying Face Recognition in Unconstrained Environments[J]. University Of Massachusetts, Amherst, Technical Report. 2007: 7-49.

[15] Ng H W, Winkler S. Advanced Digital Sciences Center (ADSC), University of Illinois at Urbana-Champaign, Singapore[C]. IEEE International Conference on Image Processing. 2015: 343-347

[16] Schroff F, Kalenichenko D, Philbin J. FaceNet: A Unified Embedding for Face Recognition and Clustering[C]. Computer Vision and Pattern Recognition. 2015

[17] (Zhou Guo-Feng, Fu Gui-Lei, Li Hai-Tao, et al. Summary of multi-pose face recognition[J]. Pattern Recogniton and Artificial Intellgence. 2015(07):613-625)