Improving Similar Face Recognition Using ResNet-50 and FPN with Triplet Loss

Xinbei Miao

School of Computing/Computer Science and Technology, Zhuhai College of Science and Technology, Zhuhai, China

miaoxinbei@stu.zcst.edu.cn

Abstract. Identical twin recognition has been attracting attention as a challenge in face recognition. In this paper, the method of FPN combined with ResNet-50 and triplet loss function is used to achieve better recognition ability on similar face datasets including identical twins without affecting much the performance of ordinary face datasets. In order to illustrate the role of FPN, the author conducted a series of comparative experiments. With or without FPN as a variable, the experiment concluded that ResNet-50 combined with FPN enhanced the recognition ability of similar faces, and the author listed the reasons that may have led to this result. Subsequently, under the condition of different training parameter settings, before overfitting, the learning rate and the number of iterations were positively correlated with the performance of the model on the ordinary faces dataset and negatively correlated with that on the similar faces dataset (including identical twins). The performance of the model on the ordinary face dataset is negatively correlated with the performance on the similar face dataset. Finally, a relatively good set of parameters is decided among the three sets tested.

Keywords: Face recognition, similar face recognition, FPN, neural network.

1. Introduction

Face recognition is now widely used in many areas, such as online payment and smart unlocking, and with the gradual use of face recognition as an important approach to authentication, it is required to be more secure and accurate.

However, although many current facial recognition algorithms have a high recognition rate on ordinary human faces, there are still some challenges in recognizing similar faces and monozygotic twins [1]. That is because most of the recognition technologies recognize faces by their geometric structures, such as the shape of the facial features, which can look very similar among those people.

To solve the problems, methods for recognizing ocular features such as iris recognition were proposed, while those methods always need good illumination and high-resolution camera, which have too many restrictions in normal life [2].

Similarly, the limitation of identifying through ear contours is that the recognition accuracy will be affected by the image angle, while people don't always show their ears head-on [3]. In order to make people's lives safer and more convenient, the technology of relying only on faces, or even half-occluded faces, to identify similar people still needs to be studied.

Gnatyuk V and Moskalenko A proposed a method to test identical twins by facial asymmetry of human faces, which has high accuracy [4]. However, they did not test the algorithm on the ordinary faces datasets, thus it is not certain whether it can be used directly to recognize all faces. In addition, there is no data to suggest that the degree of facial symmetry is unique, and the fact that twins have different facial symmetry does not mean that they will not have the same symmetry as others. Therefore, this is a method that needs to be used with other technology.

2. Method

In this work, Feature Pyramid Network (FPN) is used to obtain the spatial relationship of facial features, which encompass more than just symmetry but are richer and more complex, yet can still be included in a comprehensive model.

In order to prevent the role of the FPN from being obscured by the capabilities of the backbone network, the author did not use the neural networks developed for facial recognition with extremely high accuracy (such as FaceNet, SphereFace and DeepID3) [5]. However, it still needs to achieve recognition of ordinary facial features, with the goal of being able to distinguish similar faces even those of identical twins. Thus the author uses ResNet-50 as the backbone, and combines it with the FPN and triplet loss. This is a multi-scale feature fusion approach that leverages the advantages of FPN and ResNet-50 to allow the model to better extract features at different scales and easier understand the spatial relationships among features.

In the experimental part, the author describes the process of fine-tuning, and conducted comparative experiments to observe the effect of FPN on the recognition of similar faces, and the influence of parameter settings on this effect.

2.1. ResNet backbone

Increasing the depth of a neural network is widely used to extract features from more dimensions. Experiments have shown that deeper networks can learn more complex feature presentations because every layer can build new features depending on the last one. However, the increase in the depth of the neural networks brings about issues of degradation phenomena, which means the train error increases as the number of layers increases, and one of the main reasons for degradation phenomena is vanishing gradients.

Vanishing gradients are that the parameters of the layers which close to the input do not update as significantly as they should. That is because the networks use the chain rule to deliver the gradient, which makes the change of the parameter of the layers become less especially when they are deep [6].

In order to mitigate the degradation phenomena, the advent of residual networks (ResNets) had been created. That is because the ResNets has a special component named residual blocks, which can add skip connections between layers, so that the networks can deliver information effectively even if they are very deep [7]. After several years of development, the ResNet has various versions, including ResNet-18, ResNet-34, ResNet-50, ResNet-101, ResNet-152.etc. The main differences between these networks are the depth and the type of residual blocks they use, such as Basic Blocks and Bottleneck Blocks.

In this experiment, the backbone is ResNet-50, which has 50 layers. The structure is listed as below:

- The input layer: The image size is typically 224x224x3.
- The first convolutional layer: The convolution kernel is 7x7 with stride 2. That is used to extract the initial feature.
- The max pooling layer: The size is 3x3 with stride2. That is used to reduce the feature map dimensions.
- The residual blocks: The residual blocks are distributed across four stages of the network:

1) The first stage: It has 3 residual blocks, each one having 3 convolutional layers (64, 64, 256 channels) with stride 2.

2) The second stage: It has 4 residual blocks, each one having 3 convolutional layers (128, 128, 512 channels) with stride 2.

3) The third stage: It has 6 residual blocks, each one having 3 convolutional layers (256, 256, 1024 channels) with stride 2.

4) The fourth stage: It has 3 residual blocks, each one having 3 convolutional layers (512, 512, 2048channels) with stride 2.

- The global average pooling layer: This is used to compress the feature map into a fixed-length vector. (This will be removed in order to connect with the FPN in this experiment.)
- The fully connected layer: That is the output layer, which typically has 1000 nodes. (This will be removed in order to connect with the FPN in this experiment.)

The residual connection allows the information in the network to be retained and avoids the loss of information in the transmission process, which is beneficial for maintaining the integrity of fine-grained features. Moreover, as the depth of the network increases, the resolution of the feature map of each stage of ResNet-50 decreases gradually, while the number of channels increases. This design enables the network to simultaneously capture different layers of the image, including low-level and higher-level information, which provides a good basis for extracting multi-scale features.

2.2. Feature Pyramid Network

In order to be able to recognize similar faces as well as ordinary faces, the spatial relationships of facial features, which are more accurate than asymmetry, need to be learned. Therefore, FPN is used to fuse multi-scale features.

FPN is a technique, that uses top-down and lateral connections to combine the high-level semantic information and low-level spatial details, which makes it perform better stability, and has made big success on some detection datasets [8, 9]. Meanwhile, when high-level features are fused with low-level features, the network is able to better understand the spatial relationship between different features. In addition, in the top-down path, high-level features are upsampled to restore their spatial resolution. This preserves the semantic information of high-level features while recovering their spatial details. All the above advantages can give FPN a more accurate understanding of the distribution of face grid features in face recognition, which should also include the understanding of symmetry.

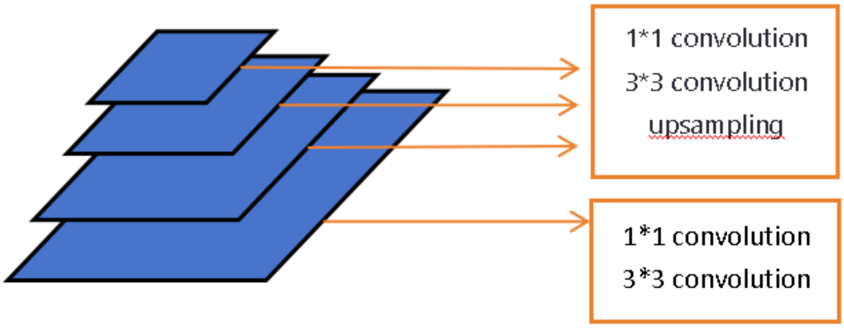

Figure 1. Construction of FPN.

Furthermore, the design of FPN ensures that it will not significantly increase computational cost although it has added some extra routes and connections. Lastly, FPN can be easily integrated with the ResNet-50 backbone, which does not need to make significant changes to the existing architecture. Figure 1 is the construction of the network in this experiment.

2.3. Triplet loss

The triplet loss is a contrastive learning method which is commonly used and fit for scenarios that need to decide the connection between entities by their similarity, such as face recognition. The principle of the function is that the function compares the relative distances among three samples (anchor sample, positive sample, negative sample), and lets the distance between the anchor sample and the positive sample be the least, while those between the anchor sample and the negative sample should be the maximum [10].

Here is the typical definition of the triplet loss:

$$\mathrm{L(A,P,N)} = \max\{\mathrm{d(A,P)} - \mathrm{d(A,N)} + \mathrm{margin}, 0\}$$

$\mathrm{d(A,P)}$ represents the distance between the anchor sample and the positive sample;

$\mathrm{d(A,N)}$ represents the distance between the anchor sample and the negative sample;

α is a positive constant named margin, which controls the minimum difference of the distances between the positive sample and the negative sample.

If the categories show subtle differences among them, the triplet loss can strength these differences by maximize the distance between the anchor sample and the negative sample. Even though the features are hard to recognize, the function can force the model to learn other feature presentations which are more discriminative, which can also enhance the robustness and accuracy of the model. In addition, the function can be adapted to various application contexts, and then enhance the generalizability of the model. Therefore, the triplet loss is chosen to be the loss function of this experiment.

2.4. Measure

Due to the scarcity of datasets specifically focusing on similar-looking faces and the difficulties that individuals can access those data sets, this experiment utilizes the LFW dataset and adds 90 images of similar faces which are generated by the Qwen model powered by Alibaba Cloud to serve as the experimental dataset. Here are some of the generated images in Figure 2 (from left to right, there are anchors, positive samples, and negative samples).

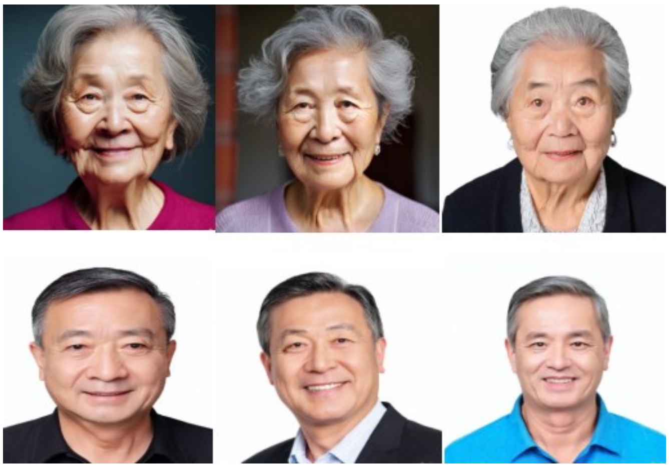

Figure 2. Images generated by the Qwen model.

The data preprocessing included resizing all images to 224x224 pixels and normalization. Because of the usage of the triplet loss, the dataset is divided into 1711 groups of triplets with one anchor image, one positive image and one negative image, and 10 of those groups which are totally composed by generated similar faces are constructed as a special test dataset. Then, 10 percent of those groups are used as the test sets, another 10 percent are the validation sets, while the others are the training sets. These sets are randomly composed to show a more realistic representation of the model's capabilities.

Furthermore, the author searched some images of people who are recognized for their similar appearance by the public and identical twins on the internet as another two test sets (10 groups/5 groups). However, due to copyright issues, it is only for research use, so the pictures will not be put in this paper.

In this experiment, the author initialized the network with weights pre-trained on ImageNet, and a stochastic gradient descent optimizer was used, with a learning rate of 0.001, a momentum term of 0.9, and a total of 10 epochs. To decrease the computational, the output of the last layer of FPN is used to calculate the loss. The margin (parameter of triplet loss) is set to 1.0. The implementation was done in Python using the PyTorch framework, running on a machine equipped with 32 vCPU Intel(R) Xeon(R) Platinum 8352V CPU @ 2.10GHz.

In order to visualize the effect of the model, the author considers groups where the distance between the anchor and the positive sample is greater than the distance from the negative sample as correct, and vice versa as false. Then the accuracy is calculated as the number of the correct groups divided by the total number of groups.

3. Results

3.1. Preliminary results and comparison with other methods

As the preliminary result in Table 1, the ResNet-50 model combined with FPN and triplet loss has a 0.741 accuracy on the test set after ten epochs of training. In addition, it has a 0.4 accuracy on the generated images test set about similar faces, while has 0.7 accuracy on the test set about similar faces of real people. On the identical twins test set, the accuracy is 1.0.

Table 1. ResNet-50 model combined with FPN with 10 epochs and 0.001 learning rate.

	Average loss	Accuracy
Ordinary faces test set (170 groups)	1.935	0.741
Generated similar faces test set (10 groups)	3.489	0.4
Real similar faces recognized by the public (10 groups)	1.315	0.7
Identical twins test set (5 groups)	0.121	1.0

Therefore, it can be seen that the model has a better ability to recognize real faces. While in the dataset of the pictures generated by the Qwen model powered by Alibaba Cloud, which is also AI, it does not have a good ability to distinguish. In addition, due to the scarcity of twins data, the performance of the model on the twins dataset does not mean that it can identify twins with an accuracy of 100%, but only that its ability to identify faces is not weakened by the characteristics of twins.

In order to show the role of the FPN network in this experiment, the author conducted a controlled trial by using ResNet-50 and Triplet loss. The only difference between these experiments is the usage of the FPN. The results are shown in below Figure 3-6.

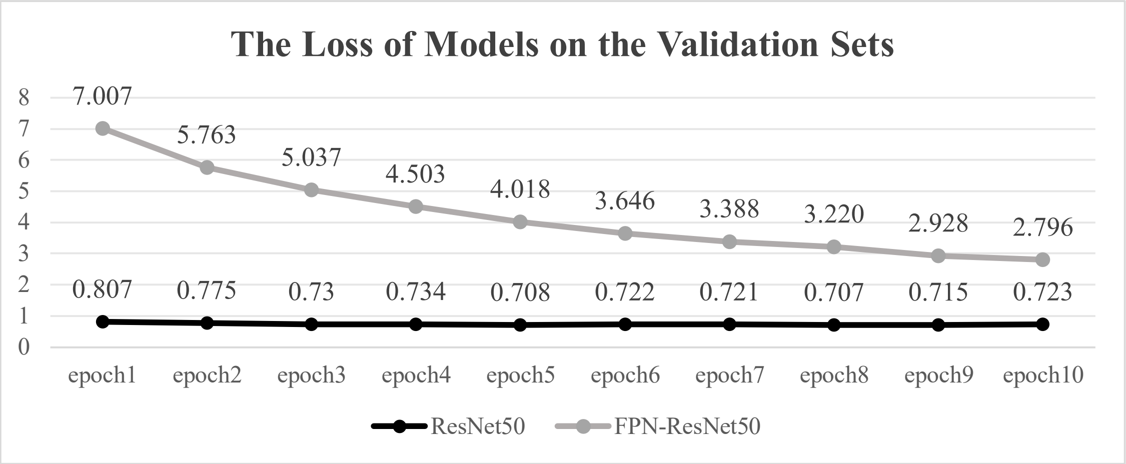

Figure 3. The loss of models on the validation sets.

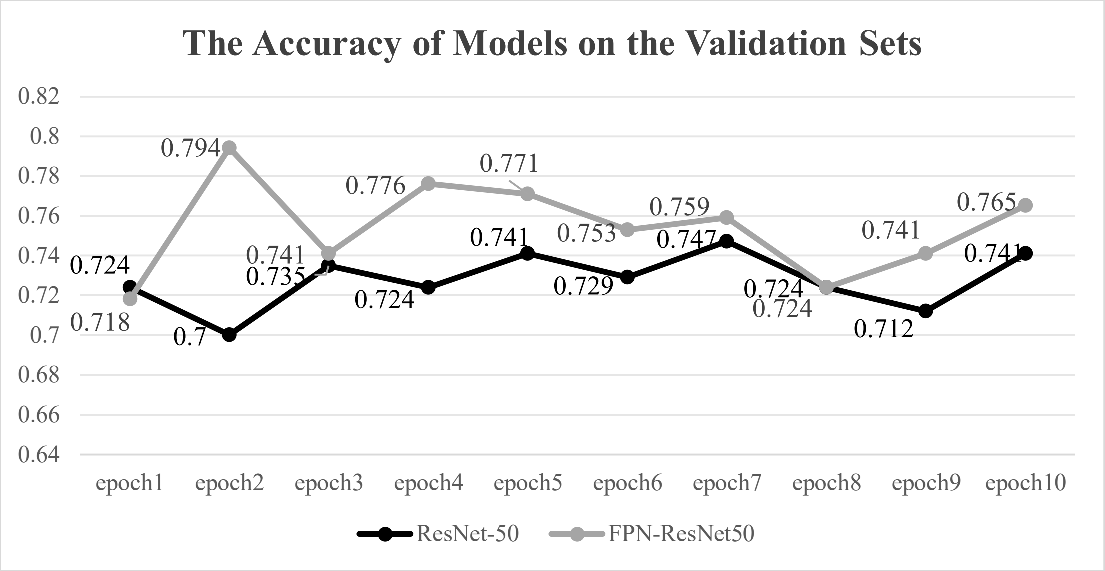

Figure 4. The accuracy of models on the validation sets.

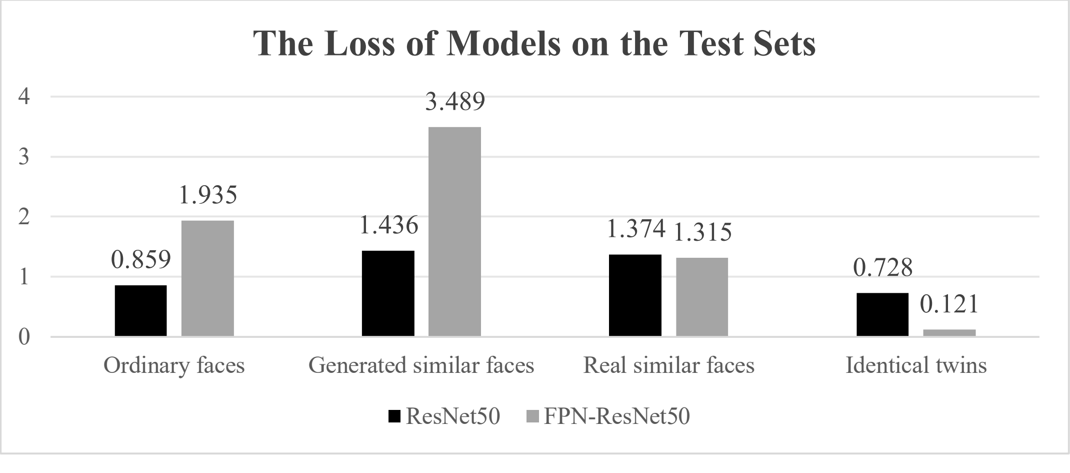

Figure 5. The loss of models on the test sets.

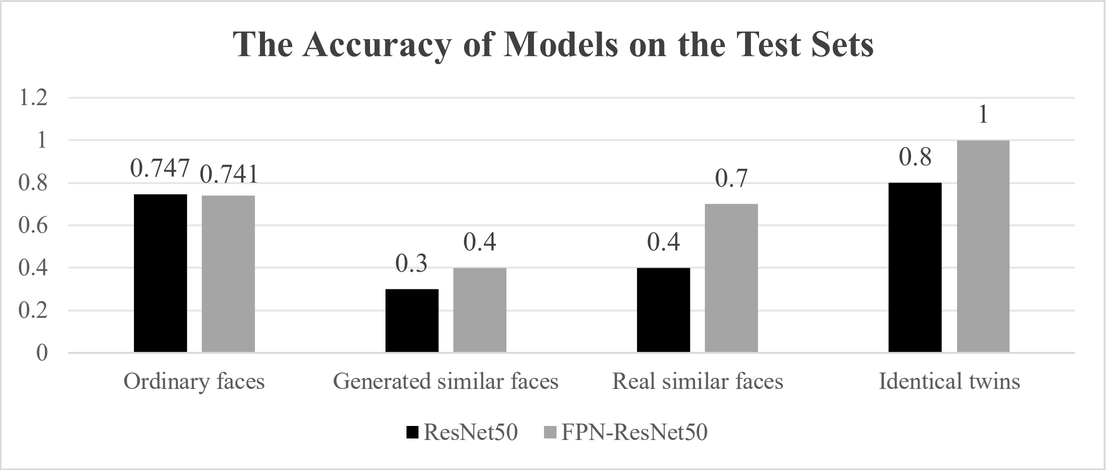

Figure 6. The accuracy of models on the test sets.

For both models, in terms of the rate at which losses decreased, the ResNet-50 combined with FPN is constantly converging, while ResNet-50 converges only at the beginning, that may be because the ResNet-50 is pre-trained, and although ResNet-50 combined with FPN is also pre-trained, the FPN still needs to be trained.

In terms of the accuracy, the accuracy of both models fluctuated and did not improve significantly after 10 epochs of training. Judging by the declining losses of the model, this may be because the learning rate is too low and the model has not been fully converged, thus, although the model is trained to bring positive samples closer to the anchor point, this proximity is not enough to change the model's judgment. To verify this guess and enhance the ability of the model, the author did other experiments below.

On the test sets, the loss of the model combined with FPN is higher on the normal dataset (1.935), the generated dataset about similar faces (3.489), but lower on the dataset about the real people who have similar outlook (1.315), while it is lower on the dataset of the twins (0.121), compared to the ResNet-50 model. However, judging by the accuracy, the model with FPN performed better in most of the datasets, except for the little deficiency in normal datasets.

3.2. Model with different parameters and end result

In order to enhance the ability of the model, two additional experiments were conducted. The first one is the model with a higher learning rate at 0.01. The second is that with more training epochs (20). Figure 7 and 8 are the results compared to the preliminary experiment.

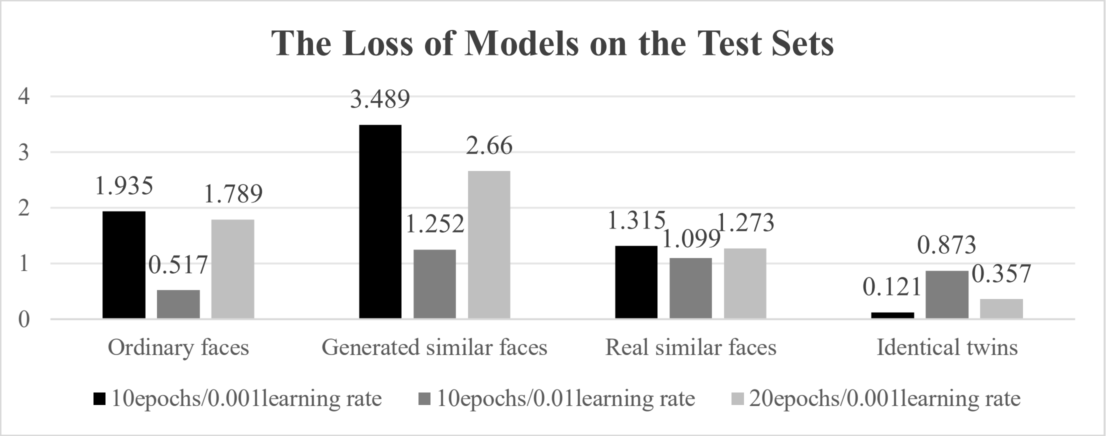

Figure 7. The loss of models on the test sets.

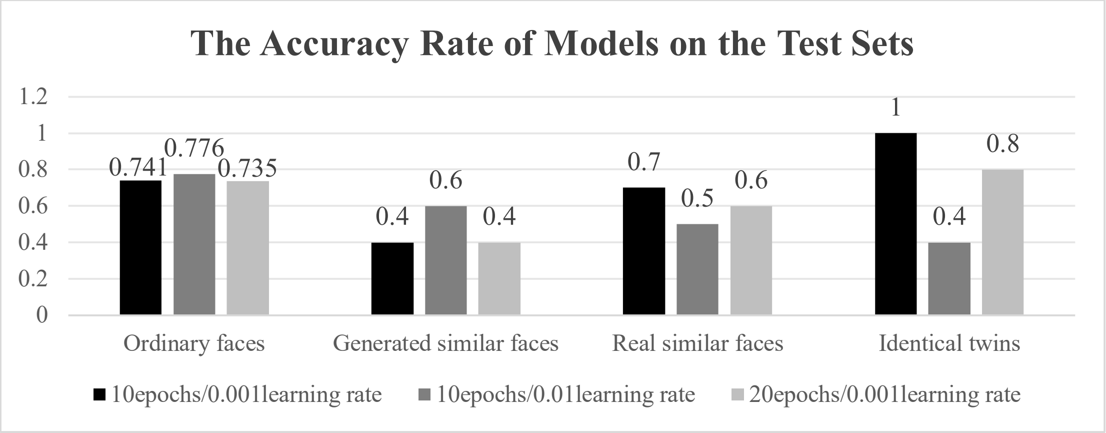

Figure 8. The accuracy of models on the test sets.

It can be easily seen that the increase in learning rate enhances the capability of the model to recognize ordinary faces (accuracy at 0.776) and generate similar faces (accuracy at 0.6). In particular, the recognition accuracy of ordinary faces reached a maximum of 0.824 on the verification set. But as for real similar faces and twins, the model is performing badly. That may be because the high learning rate lets the model ignore some of the details, so it is hard to distinguish identical twins with extremely similar faces.

Although the increase of epoch did not make sense to the accuracy (even reduced some of the accuracy), it makes the loss get lower than those in preliminary experiments, which illustrates that the model has converged further.

Summarizing the above data, it can be seen that with the reduction of losses and the gradual convergence of the model, the performance of the model on the ordinary faces dataset and the AI-generated similar faces dataset will become better, but the performance on the recognized similar face dataset and twin-face dataset will become worse. This is because as the model has a higher speed and extent of converging, the features extracted by the model are more discriminatory, yet some of the details are easier to discard, and it is precisely some details that are needed to distinguish similar faces. From the point of view of distinguishing similar faces, the best model is still the preliminary model.

4. Conclusion

In this paper, the recognition ability of FPN combined with ResNet-50 and triplet loss on ordinary faces and similar faces (including identical twins) is studied, and comparative experiments are carried out from two perspectives: with or without FPN and three different training parameter settings.

The former concludes that FPN can generally improve the recognition accuracy of similar faces of the model, and keep the accuracy of ordinary face recognition not greatly affected. The latter shows that the accuracy of FPN combined with ResNet-50 and triplet loss is positively correlated with the convergence degree of the model before overfitting, and the accuracy of similar face recognition is just the opposite. Thus, the accuracy of ordinary face recognition is negatively correlated with that of similar faces recognition. Furthermore, the higher convergence speed will also affect the model's recognition accuracy of similar faces.

In the second comparative experiment, out of the three sets of parameters, the author finally chose the first set of parameters (learning rate 0.001, epoch 10), under which the performance of the model was: 0.741 (accuracy) on ordinary faces dataset (170 groups), 0.4 (accuracy) on generated similar faces dataset (10 groups), 0.7 (accuracy) on real similar faces dataset (10 groups), 1.0 (accuracy) on identical twins dataset (5 groups), which has better overall performance than ResNet-50.

In the future, the author will further adjust these parameters to achieve a balance between the retention and discarding of detailed features, and further improve the model's capabilities. Also, models that are more targeted to face recognition tasks will be tried as backbone networks to try to get better performance in recognizing both ordinary faces and similar faces (including identical twins).

References

[1] Phillips, P. J., Flynn, P. J., Bowyer, K. W., Bruegge, R. W. V., Grother, P. J., Quinn, G. W., & Pruitt, M. (2011, March). Distinguishing identical twins by face recognition. *In 2011 IEEE International Conference on Automatic Face & Gesture Recognition (FG)* (pp. 185-192). IEEE.

[2] Nguyen, K., Proença, H., & Alonso-Fernandez, F. (2024). *Deep learning for iris recognition: A survey. ACM Computing Surveys,* 56(9), 1-35.

[3] Booysens A, Viriri S. Ear biometrics using deep learning: A survey[J]. *Applied Computational Intelligence and Soft Computing*, 2022, 2022(1): 9692690.

[4] Gnatyuk, V., & Moskalenko, A. (2020, September). Mobile Twin Recognition. *In 2020 IEEE International Joint Conference on Biometrics (IJCB)* (pp. 1-9). IEEE.

[5] Li, L., Mu, X., Li, S., & Peng, H. (2020). A review of face recognition technology. *IEEE access*, 8, 139110-139120.

[6] Roodschild, M., Gotay Sardiñas, J., & Will, A. (2020). A new approach for the vanishing gradient problem on sigmoid activation. *Progress in Artificial Intelligence*, 9(4), 351-360.

[7] Peng, S., Huang, H., Chen, W., Zhang, L., & Fang, W. (2020). More trainable inception-ResNet for face recognition. *Neurocomputing*, 411, 9-19.

[8] Gong, Y., Yu, X., Ding, Y., Peng, X., Zhao, J., & Han, Z. (2021). Effective fusion factor in FPN for tiny object detection. *In Proceedings of the IEEE/CVF winter conference on applications of computer vision* (pp. 1160-1168).

[9] Luo, Y., Cao, X., Zhang, J., Guo, J., Shen, H., Wang, T., & Feng, Q. (2022). CE-FPN: enhancing channel information for object detection. *Multimedia Tools and Applications*, 81(21), 30685-30704.

[10] Sharma, S., & Kumar, V. (2021). 3D landmark - based face restoration for recognition using variational autoencoder and triplet loss. *IET Biometrics*, 10(1), 87-98.