Mask Wearing Specification Detection System Based on Residual Network

YuChen Zhang(✉), ZiQi Shao, YaNan Chen, GuangHan Guo, and HangXu Wu

School of Computer and Software, Nanyang Institute of Technology, Nanyang 473004, Henan, China
zyc@nyist.edu.cn

Abstract. In order to solve the problem of normative detection of mask wearing, this paper proposes an LMSE-ResNet based on improved residual network. Firstly, the cascade classifier is used to strengthen the features of the mask-wearing part, detect whether the face in the image wears a mask, and then load the pre-trained weights and parameters into the convolutional layer of the new model. The number of channels of feature mapping is increased while reducing the residual network depth, and finally the fitting error of the minimum mean square linear model is used as the loss function to improve the detection accuracy. In the experiment on the public dataset Masked Face-Net, the algorithm achieves higher accuracy and lower loss value, and has better effect and robustness in detecting mask wearing irregularities.

Keywords: Residual network · mask detection · fitting error

1 Introduction

Wang Yihao et al. proposed an improved mask detection algorithm based on YOLOv3 to improve the detection accuracy of mask wearing detection by using an improved spatial pyramid pooling structure [1], and Niu Zuodong et al. proposed an improved model based on Retina Face algorithm, which introduces a self-attention mechanism into the feature pyramid to optimize the loss function to realize mask wearing detection in natural scenes [2]. Jiang Yuewu et al. proposed a residual network as the basis, and at the same time introduced the channel and spatial attention mechanism to form an attention residual network to extract features from the image wearing a mask, and the residual network carried out feature fusion at different levels, and the attention mechanism improved the sensitivity of the network to the main features by enhancing the aggregation of useful information and the suppression of useless information [3].

It is found that most mask detection algorithms only identify whether masks are worn, but few people mention how to detect whether to wear masks, and the ResNet residual network uses the Relu function as the activation function to integrate the class-differentiated local information in the convolutional layer or pooling layer. The last fully connected output layer uses the linear classifier softmax for logistic regression classification. Although the degradation problem and gradient disappearance problem

F. Huang et al. (Eds.): ICAIE 2023, AHCS 15, pp. 554–562, 2024.
https://doi.org/10.2991/978-94-6463-242-2_68

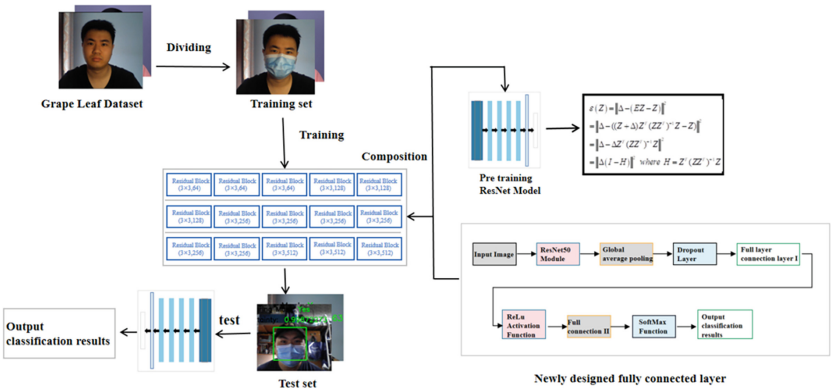


Fig. 1. Implementation process of the algorithm in this paper

of traditional convolutional neural network with the deepening of the number of network layers have been changed, there are many parameters, high complexity and weak expression ability in the convolution stage, and there are large computations in the fully connected layer, the Relu function training process does not adapt to large gradient inputs, after the parameter update, ReLU's neurons will no longer have the function of activation, resulting in a gradient of zero forever, and softmax classification can not well distinguish the extracted features. It will cause the model to be overfitted and even cause the intraclass spacing to be greater than the interclass spacing [4].

To solve the above problems, based on the feature extraction ResNet-50 [5], a new residual network model is designed, and the pre-trained weights and parameters are loaded into the convolutional layer of the new model. The traditional fully connected layer in Res Net is deleted, and a new fully connected module is designed to reduce the residual network depth and increase the number of channels of feature mapping, and at the same time, introduce Dropout between convolutional layers to effectively prevent overfitting. In order to alleviate the gradient disappearance problem in the residuals, the output features of the last set of residual blocks are extracted as feature descriptors, combined with the fitting error of the minimum mean square linear model, so that the features of the mask parts are strengthened and the goal of identifying whether to standardize the wearing of masks is achieved. This is shown in Fig. 1.

2 The Algorithm Model of This Paper

2.1 Image Preprocessing

The algorithm model first needs to denoise the image to prevent the noise of the image from affecting the detection results, convert the dataset into a grayscale map, the purpose of converting to a grayscale map is to simplify the matrix and improve the operation speed, and then use a cascade classifier for face detection, determine whether there is a face in the image, and the position and size of the face in the input image, and then return an array of detection boxes, each box represents a detected face. Set the recognition

number min Neighbors to 7, which means that each picture needs to be found 7 times to determine the true feature target, and the proportion of each image reduction is set to 1.05, which is to convert the image into a binary image, that is, the pixels in the image are divided into two categories according to certain conditions: one is greater than or equal to the threshold, and the other is less than the threshold. After finding the face image, segment or extract the target area in the image, determine the highest confidence level of the image to update the noise reduction average bounding box, if the new bounding box after noise reduction is close to the average, it proves that the noise reduction is effective, update and return the new average bounding box value, and return the last updated bounding box if noise reduction fails. After that, the face image is cropped and adjusted to the correct size, then converted into an array and expanded to facilitate the model to detect.

2.2 Feature Extraction

Although the accuracy of classification can be improved by increasing the depth of the residual network, sometimes it is necessary to double the number of network layers in order to increase a small amount of accuracy, which not only increases the computational cost, but also reduces the reuse of features and reduces the training speed of the network. The use of small convolution kernels can effectively reduce the number of parameters, make training and testing more effective, and increase the width of the network while reducing the depth of the residual network. In this paper, the input image of the model is $224 \times 224 \times 3$, and the convolution kernel set to 3×3 in the input can obtain a feature map with a receptive field of 5. Among them, the output channel size of the first and second 3×3 convolutional layers is 32, the stride size is 2, and the output channel size of the last convolutional layer is 64, which greatly reduces the computational cost of the classification network and reduces the computational amount of the network model under the condition of ensuring that the output backbone information is consistent with the previous output. At the same time, a Dropout layer with a P value of 0.5 is introduced between the convolutional layers, which can effectively prevent overfitting. Resnet is divided into 5 modules, each Resnet module has five sets of convolution, because the fully connected layer requires a lot of calculation, so the fully connected layer is modified, and the output features of the last set of residual blocks are extracted as feature descriptors.

Firstly, let $Z \in M^{3\times n}(R)$ be the homogeneous coordinates of a set of points, and let Δ be the displacement of the same point in homogeneous coordinates, and let $E \in M^{3\times 3}(R)$ be the affine transformation matrix.

$$Z = \begin{pmatrix} x_1 \cdots x_n \\ y_1 \cdots y_n \\ 1 \ \cdots \ 1 \end{pmatrix} \tag{1}$$

The solution to Eq. (1) can be expressed as:

$$E = (Z + \Delta)Z^T \left(ZZ^T\right)^{-1} \tag{2}$$

The error Eq. (3) can therefore be estimated by replacing E with its solution without a minimization step:

$$\begin{aligned}
&\varepsilon(Z) = \|\Delta - (EZ - Z)\|^2 \\
&= \left\|\Delta - ((Z + \Delta)Z^T(ZZ^T)^{-1}Z - Z)\right\|^2 \\
&= \left\|\Delta - \Delta Z^T(ZZ^T)^{-1}Z\right\|^2 \\
&= \|\Delta(I - H)\|^2 where H = Z^T(ZZ^T)^{-1}Z
\end{aligned} \tag{3}$$

The threshold τ is set for each set of calculated fitting errors, and the purpose is to binarize the results calculated by the fitting error $\varepsilon(Z)$, and segment the image to select the detected mask area.

3 Experiment and Result Analysis

3.1 Dataset and Experimental Environment

MaskedFace-Net is a large dataset containing 137,016 high-quality mask face images that can be seen as a benchmark dataset for creating machine learning models related to mask wearing analysis [6]; Especially whether to wear a mask and whether to wear it accurately. This dataset was created using the mask-to-face deformable model. According to the ratio of 8:2, the dataset is divided into a training set and a test set, and the training set is mainly used as the input data for training the mask wearing detection and recognition model, and the function of the test set is to detect the accuracy of the model after the model training is completed. Dataset address: https://hikariming/virus-mask-dataset. This experiment is based on the Python 3.8 PyTorch framework, using NVIDIA RTX 2080Ti GPU to train and test the model, each experimental iteration 200 times, the initial learning rate is 0.01, and the batch_size is set to 10. Test set error rate (Vaildation loss) to evaluate algorithm performance.

3.2 Analysis of Results

The accuracy and error of the training set and test set after VGG16 training are shown in Fig. 2 (a), (b).

From the VGG16 accuracy and error rate line charts, it can be seen that the accuracy of the training set gradually tends to 1 with the increase of the number of iterations epoch, but the accuracy of the test set has a downward trend after the number of iterations epoch exceeds 100 times. The error rate of the training set gradually tends to 0, but the error rate of the test set fluctuates greatly after 100 times. In summary, when VGG16 is selected as the model used by the system, the number of iterations epoch should not exceed 100 times, so as to avoid large errors in the results.

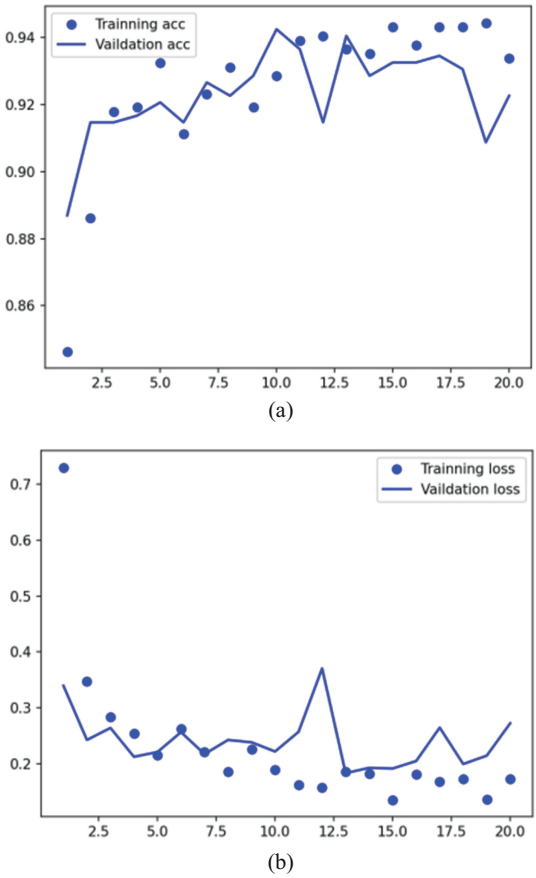
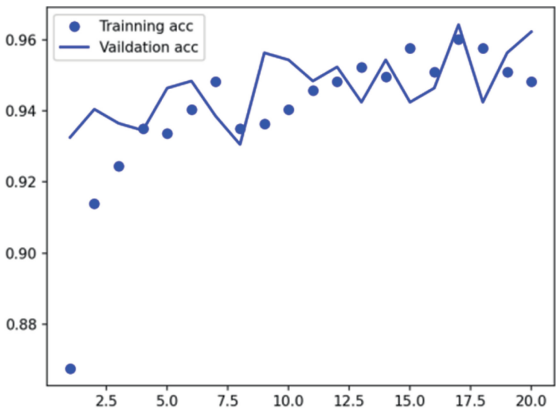


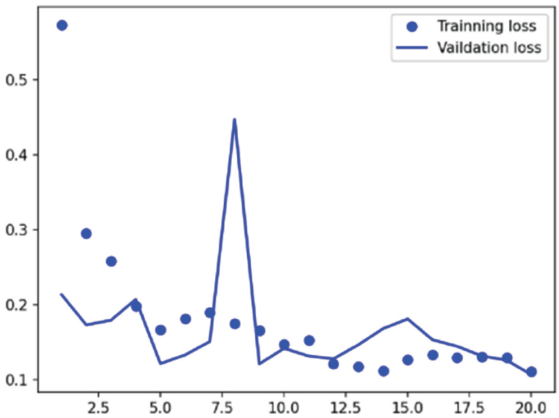
Fig. 2. VGG16 accuracy and error rate line chart

The accuracy and error of the training set and test set after training of ResNet50 are shown in Fig. 3 (a), (b).

It can be seen from the ResNet50 accuracy and error rate line chart that when the number of iterations epoch is between 10–15 times, the accuracy of the training set tends to 1, and after more than 15 times, the accuracy rate has a downward trend. The test set is the same, the number of iterations epoch is stable at 10–15 times, and the rest of the times fluctuate greatly. The error rate of the training set tends to 0 with the increase of the number of iterations and epochs, and the error rate of the test set fluctuates greatly in 6 to 8 times, and the more times, the more times tend to be stable. In summary, when ResNet50 is selected as the network model of this system, the number of iterations epoch should be selected 12 times to prevent large errors in the system.



(a)

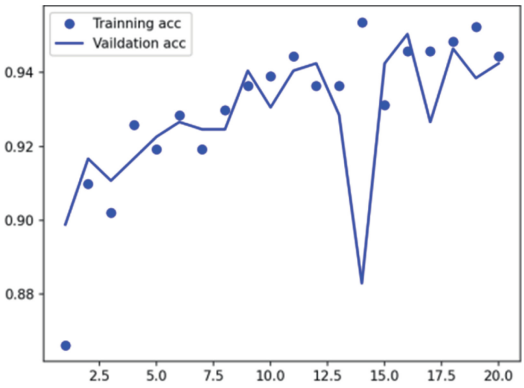


(b)

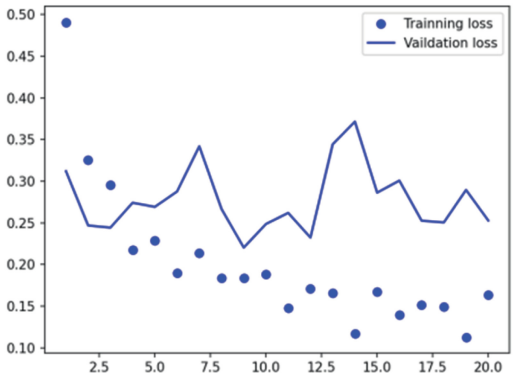
Fig. 3. ResNet50 accuracy and error rate line chart

The accuracy and error of the training set and test set after training of MobileNetV2 are shown in Fig. 4 (a), (b).

It can be seen from the MobileNetV2 accuracy and error rate line chart that the more epochs the number of iterations, the accuracy of the training set tends to 1, and the accuracy of the test set fluctuates greatly. The error rate of the training set gradually tends to 0 with the increase of the number of iterations, but the error rate of the test set fluctuates greatly. In summary, the MobileNetV2 network model is not suitable as the model selected for this system.



(a)



(b)

Fig. 4. MobileNetV2 accuracy and error rate line plots

Analyzed by the above experimental data, combined with the accuracy and error rate line chart and the speed of model training, the system comprehensively selects the ResNet50 neural network, and sets the number of iterations epoch to 12 times to obtain the best accuracy effect, and the accuracy of the training set after the experiment is 0.94 and the error rate is 0.12. The test set accuracy is 0.95 and the error rate is 0.12.

The results of the image detection and recognition function are shown in Fig. 5.

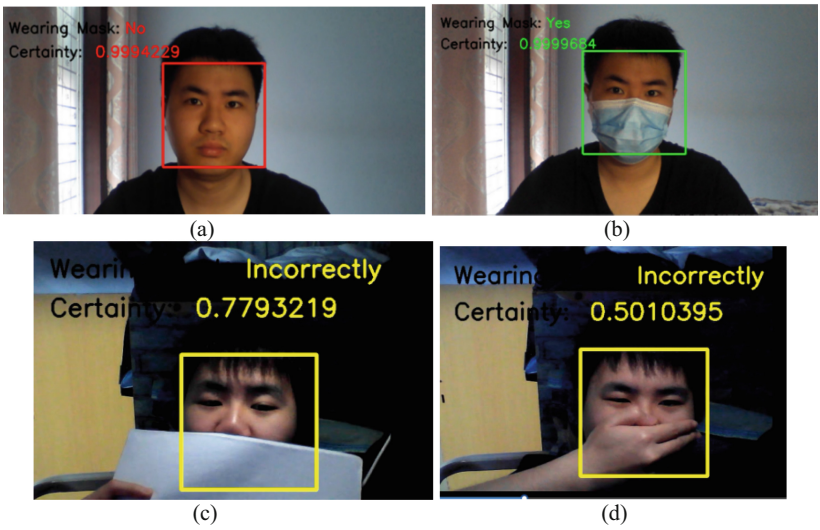


Fig. 5. Camera detection function results

4 Conclusion

Aiming at the normative detection problem of mask wearing, this paper uses a cascade classifier to strengthen the features of the wearing mask, effectively highlights the important features, and then designs a new fully connected module, and then introduces the fitting error of the minimum mean square linear model as the loss function, which makes the positioning more accurate and improves the detection accuracy. In the later stage, it will mainly focus on how to improve the detection speed of the algorithm to make the algorithm performance better.

References

1. Wang Yihao, Jin Yi, Jin Guoqiang, et al. Mask wearing detection algorithm based on improved YOLOv3 in complex scenarios [J]. Computer Engineering. 2020,46(11)
2. Niu Zuodong, Tan Tao, Chen Jinjun. Improved the natural scene mask wearing detection algorithm for RetinaFace [C]// Computer Engineering and Applications. 2020,56(12)
3. Jiang Yuewu, Zhang yujin. Mask wearing norms detection algorithm based on attention residual network [J]. Sensors and Microsystems. 2023,42(02)
4. Howard A, Sandler M, Chu G, et al. Searching for MobileNetV3 [C]//Proc of IEEE/CVF International Conference on Computer Vision. Piscataway, NJ: IEEE Press, 2019: 1314–1324.
5. Long J, Shelhamer E, Darrell T. Fully convolutional networks for semantic segmentation [C]//Proc of IEEE Conference on Computer Vision and Pattern Recognition. Washington DC: IEEE Computer Society, 2015: 3431–344.
6. Adnane Canani, Karim Hammoudi.MaskedFace-Net -- A Dataset of Correctly/Incorrectly Masked Face Images in the Context of COVID-19