ISSN: 2664-9640

Masked face recognition algorithm based on convolutional neural network

Xing Hu

College of Mechanical and Electronic Engineering, Northwest A&F University, Shaanxi, China, 712100, China

Abstract

COVID-19 has emerged and spread all over the world, and wearing masks is one of the most effective ways to prevent infection. In order to ensure the wearing of masks in crowded scenes, the automatic recognition of masks has quickly become a research hotspot. In the actual face mask detection scene, the model is required to have high accuracy and operation speed. Considering the accuracy and operation speed, this paper constructs a mask classification model based on resnet-18 network, evaluates the model with confusion matrix, and visualizes the model with class activation thermodynamic diagram. The results show that the accuracy of the mask classification model constructed in this study is 98.64%, the precision is 98.64%, the recall is 99.62%, and the comprehensive measure F1 score is 99.13%. The model has achieved good results. In the class activation diagram, its attention is focused on the mask, which proves that the model has good generalization performance. It has certain application value in the development of mask recognition equipment.

Keywords

Mask face detection; Convolutional neural network; Class activation diagram.

1. Introduction

Since December 2019, new coronaviruses have emerged and spread rapidly around the world. While the new coronavirus is mainly transmitted through the respiratory tract, one of the most effective ways to prevent infection is by wearing a mask. Therefore, wearing a mask has become an important measure in people's daily life. To ensure that everyone wears masks in special scenarios such as crowded areas, the automated recognition of masks is rapidly becoming a research hotspot in related fields.

And in recent years, deep learning has been widely applied and combined in the field of computer vision with good results. Convolutional neural network (CNN) [1] is a kind of Artificial Neural Network(ANN), which has become a current research hotspot in the field of image recognition. Its weight-sharing[2] network structure makes it more similar to biological neural networks, reducing the complexity of the network model and the number of weights. 1994 LeNet-5[3] handwriting recognition network was born, and its use of convolution and pooling operations to extract image features, avoiding a large amount of computational costs, and then finally using a fully connected network for classification recognition, became the architecture of a large number of convolutional neural networks It became the starting point of a large number of convolutional neural network architectures. Since then, convolutional neural networks began to gradually enter the vision of researchers. 2012, AlexNet[4] emerged in the ImageNet competition with more layers than LeNet, including ReLU activation layer and introducing Dropout mechanism in the fully connected layer to prevent overfitting. Subsequently, convolutional neural networks have been developed considerably.

ISSN: 2664-9640

Under such conditions, for the mask recognition problem, researchers have constructed an algorithmic framework for mask recognition based mostly on deep learning with image classification. Xinyu Tang[5] et al. designed a whole set of face recognition system based on paddlehub, using PyramidBox face detector and FaceBoxes backbone network for face detection and mask recognition in the target region, respectively; Youxi Ke et al[6] designed a contactless temperature temperature measurement and mask recognition system based on openmv, which used open mv to collect The system uses open mv to collect mask and maskless data, train to form a model, and develop related devices; Baihan Zou et al[7] compared the performance gap between PyramidBox-Lite model, Keras model, and CenterFace-based mask detection model in terms of processing speed, correct detection classification rate, and outlined the optimization trend of detection methods; Zixin Liu[8] designed an embedded video streaming-based video streaming based mask wearing detector and equipped with convolutional neural network on the basis; Ran Pengfei[9] et al. proposed proposed an improved mask wearing detection algorithm based on YOLOv4 and proposed an image screening algorithm used to select images that satisfy the condition of complex lighting to produce the dataset.

In addition, in the image classification application scenario of mask classification, where the target identifier and background are complex, an important issue is how to ensure that the trained model can correctly learn different features of different types of objects. It often happens that the generalization performance of the model is difficult to meet the application requirements because the chosen model cannot learn the feature parts in the image. As for the visualization of neural networks, there are three commonly used methods, including the visualization of convolutional kernel output, the visualization of convolutional kernels, and the visualization of activation-like heat maps. Visualization of convolution kernel output[10], i.e., visualization of the result after convolution operation, visualizes the result after activation of the convolution kernel to be able to see the result of the image after convolution and helps to understand the role of the convolution kernel. Visualization of the convolution kernel[10] i.e. visualization of the convolution kernel itself, which explains the behavior learned by the convolution kernel and helps to understand how the convolution kernel perceives the image. Class activation heat map visualization[11] is used to understand which parts of an image play a key role in an image classification problem by using class activation heat maps, and also to locate the position of objects in an image. It helps to understand which part of an image makes the convolutional neural network make the final classification decision. This helps to debug the decision process of the convolutional neural network, especially effective in case of classification errors. In this study, a class activation heat map visualization method is used, which should be used for classification recognition of images in mask recognition scenarios.

In response to the existing technical approaches, this study proposes the following research objectives.

1. To design a masked face recognition network model with high recognition accuracy and fast operation speed.
2. To complete the visualization of the results of the masked face convolutional neural network.

2. Model Construction

In this study, a masked face recognition algorithm is designed based on resnet-18 network, which has high accuracy and adaptability.

In the researcher's vision, the more complex the deep learning is and the more parameters it has, the more it has stronger feature extraction and classification capabilities. With this basic criterion, CNN classification networks have evolved from 7 layers of AlexNet to 16 and even 19 layers of VGG[12], and later to 22 layers of GoogLeNet[13]. However, as the number of layers

ISSN: 2664-9640

increased, it was found that when the network reached a certain depth, increasing the number of layers could not bring further improvement in classification performance, but would bring problems such as gradient disappearance and gradient explosion, which made the model converge more slowly and the classification accuracy on the test set became worse. To solve this problem, ResNet was born.

ResNet is a residual neural network proposed by Kaiming He et al[14] in 2015, and its residual learning structure diagram is shown in Figure 1.

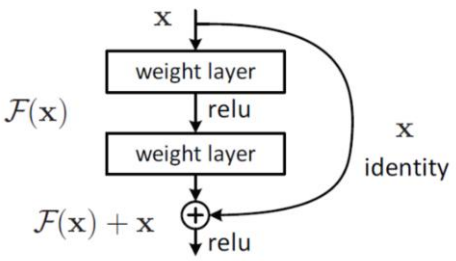


Figure 1 Residual learning structure diagram

The residual unit can be expressed as:

$$y_1 = h(x_l) + F(x_l, w_l) \tag{1}$$

$$x_{l+1} = f(y_1) \tag{2}$$

where x_l and x_{l+1} denote the input and output of the Lth residual unit, respectively, noting that each residual unit generally contains a multilayer structure. F is the residual function, which denotes the learned residual, while $h(x_l) = x_l$ denotes the constant mapping and f is the ReLu activation function.

Based on the above equation, this research derive the learned features from the shallow l to the deep L layer.

$$x_L = x_l + \sum_{i=l}^{l-i} F(x_i, W_i) \tag{3}$$

Using the chain rule again, the gradient of the reverse process can be found as follows.

$$\frac{\partial loss}{\partial x_1} = \frac{\partial loss}{\partial x_L} \cdot \frac{\partial x_L}{\partial x_l} = \frac{\partial loss}{\partial x_L} \cdot \left(1 + \frac{\partial}{\partial x_L} \sum_{i=l}^{L-1} F(x_i, W_{i)}\right) \tag{4}$$

The first factor $\frac{\partial loss}{\partial x_L}$ of the equation represents the gradient of the loss function arriving at L. The 1 in parentheses indicates that the short-circuiting mechanism propagates the gradient losslessly, while the other residual gradient needs to pass through the layer with weights and the gradient is not passed directly. The residual gradient will not be all -1, and even if it is smaller, the presence of 1 will not cause the gradient to vanish.

In this way, the gradient decay is further suppressed, and the computation of addition makes training more stable and easier. So the number of layers of trainable networks is also greatly increased.

Based on the basic structure described above, Resnet constructed several models with the architectural parameters shown in Table 1.

ISSN: 2664-9640

Table 1 ResNet architecture parameter diagram

Layer name	Output name	18-layer	34-layer	50-layer	101-layer	152-layer
conv1	112×112	7×7, 64, stride 2				
conv2_x	56×56	7×7, 64, stride 2				
		$\begin{bmatrix}3\times3,64\\3\times3,64\end{bmatrix}\times2$	$\begin{bmatrix}3\times3,64\\3\times3,64\end{bmatrix}\times2$	$\begin{bmatrix}1\times1,64\\3\times3,64\\1\times1,256\end{bmatrix}\times3$	$\begin{bmatrix}1\times1,64\\3\times3,64\\1\times1,256\end{bmatrix}\times3$	$\begin{bmatrix}1\times1,64\\3\times3,64\\1\times1,256\end{bmatrix}\times3$
conv3_x	28×28	$\begin{bmatrix}3\times3,128\\3\times3,128\end{bmatrix}\times2$	$\begin{bmatrix}3\times3,128\\3\times3,128\end{bmatrix}\times4$	$\begin{bmatrix}1\times1,128\\3\times3,128\\1\times1,512\end{bmatrix}\times4$	$\begin{bmatrix}1\times1,128\\3\times3,128\\1\times1,512\end{bmatrix}\times4$	$\begin{bmatrix}1\times1,128\\3\times3,128\\1\times1,512\end{bmatrix}\times8$
conv4_x	14×14	$\begin{bmatrix}3\times3,256\\3\times3,256\end{bmatrix}\times2$	$\begin{bmatrix}3\times3,256\\3\times3,256\end{bmatrix}\times6$	$\begin{bmatrix}1\times1,256\\3\times3,256\\1\times1,1024\end{bmatrix}\times6$	$\begin{bmatrix}1\times1,256\\3\times3,256\\1\times1,1024\end{bmatrix}\times23$	$\begin{bmatrix}1\times1,256\\3\times3,256\\1\times1,1024\end{bmatrix}\times36$
conv5_x	7×7	$\begin{bmatrix}3\times3,512\\3\times3,512\end{bmatrix}\times2$	$\begin{bmatrix}3\times3,512\\3\times3,512\end{bmatrix}\times3$	$\begin{bmatrix}1\times1,512\\3\times3,512\\1\times1,2048\end{bmatrix}\times3$	$\begin{bmatrix}1\times1,512\\3\times3,512\\1\times1,2048\end{bmatrix}\times3$	$\begin{bmatrix}1\times1,512\\3\times3,512\\1\times1,2048\end{bmatrix}\times3$
	1×1	Average pooling, 1000 Fully-connected, softmax				
FLOPs		1.8 ×109	3.6 ×109	3.8 ×109	7.6 ×109	11.3 ×109

3. Experiment

3.1. Experimental computer environment

The experiments were conducted in a server virtual environment with an Intel(R) Core(TM) i3-10100F CPU @ 3.60GHz, with 8 cores allocated in the virtual environment; a RAM size of 31GB; a GPU of GeForce RTX 2080ti with 11GB of video memory; and an operating system of ubuntu 18.04 64-bit, based on Pytorch framework and accelerated with CUDA.

3.2. Experimental data

The experimental data come from the Real-World Masked Face Dataset (RMFD), which is a dataset initiated by the National Multimedia Software Engineering Technology Research Center of Wuhan University and has tens of thousands of image data including the real masked face recognition dataset[15] in the annotated dataset, which specifically includes, real Mask face recognition dataset: the samples are crawled from the web, and after sorting, cleaning and labeling, it contains 5,000 mask faces and 8,000 normal faces of 525 people.

The samples of RMFD dataset are shown in Table 2, and this study uses this dataset for training. Assuming that the non-masked sample is labeled Negative and the masked sample is labeled Positive, the number of picture samples in each category is counted and the distribution of sample categories is plotted in Figure 2.

ISSN: 2664-9640

Table 2 Selected samples of RMFD dataset

Image	Lab Status
	Negative
	Positive

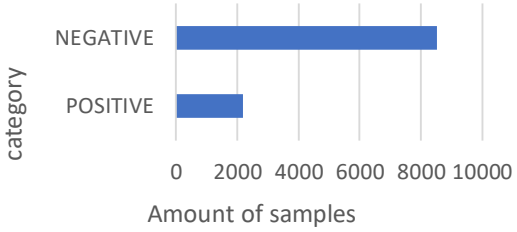


Figure 2 ResNet18 model structure

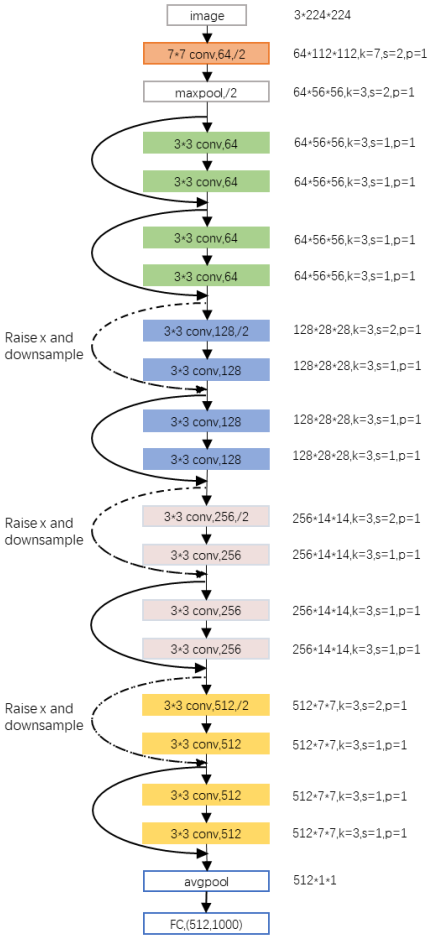


Figure 3 ResNet18 model structure

ISSN: 2664-9640

In this study, the RMFD dataset is split into a training set and a validation set in the ratio of 9:1.

3.3. Model solving

Since ResNet uses a residual structure, it achieves good accuracy on datasets such as ImageNet and COCO, etc.; and in terms of operation speed, considering that the deeper the network structure is, the slower the operation speed is relatively. Based on the above analysis, ResNet-18 was selected as the network model in this study. Its network structure is shown in Figure 3. On this basis, the model training parameters are set as shown in Table 3.

Table 3 Training parameter table

Parameter	Num
epochs	200
batch_size	128
Learing_rate	1e-3
momentum	0.9
Loss_function	Cross Entropy Loss

The loss_function is selected as Cross Entropy Loss, the learing_rate is set to 1e-3, the batch_size is set to 128, and the optimizer is selected as stochastic gradient descent, SGD. after the parameters are set and the model is determined, the hornet image report classification model is trained in PaddlePaddle deep learning framework starts training. During the training process, its overall loss curve is shown in Figure 4, and the variation of the validation set accuracy is shown in Figure 5.

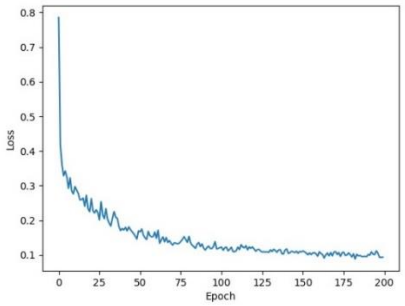


Figure 4 Loss graph

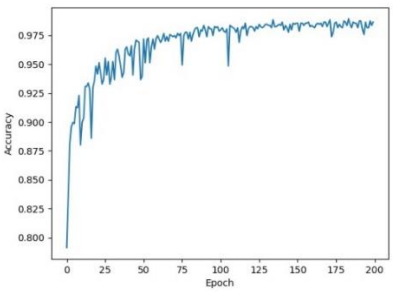


Figure 5 The variation of the accuracy of the validation set

After 200 iterations, it can be seen by observing the loss curve that the model has converged.

ISSN: 2664-9640

3.4. Model evaluation

3.4.1. Confusion matrix evaluation

In order to evaluate the model effect, save the model, evaluate the model performance through the test set, and establish the confusion matrix, as shown in Table 4:

Table 4 Confusion matrix table of classification model of wasp image report

Class	Negative	Positive
Negative	796	3
Positive	11	224

From the precession and the recall, respectively.

$$P = \frac{TP}{TP + FP} \tag{5}$$

$$R = \frac{TP}{TP + FN} \tag{6}$$

Where,F_1 is a comprehensive measure based on the harmonic average definition of precision and recall:

$$F_1 = \frac{2RP}{R + P} \tag{7}$$

The total accuracy of the model is expressed as:

$$Acc = \frac{TP + TN}{TP + TN + FP + FN} \tag{8}$$

The model evaluation results can be obtained, as shown in Table 5:

Table 5 overall average index

Class	Precision	Recall	F1-score
Over_all	0.9864	0.9962	0.9913

It can be seen that the mask classification model has achieved an accuracy of 0.9864 on the test set, its precision P = 0.9864, recall r = 0.9962, and its comprehensive measure F1 score = 0.9913. It can be seen that the model has a good classification effect.

3.4.2. Class activation map evaluation

Class activation heat map visualization is used to generate a class activation heat map for an input image by weighing the activation maps according to their gradients or their contribution to the output, thus representing the importance of each location for that class. The class activation heat map helps to understand which part of an image motivates the convolutional neural network to make the final decision and also locates a specific target degree weighted classification activation map in the image[16].

In an image classification problem, suppose that the training network recognizes an image as an object with a probability of 0.9 and wants to know what the last layer of convolutional layers contributes to the 0.9 probability. For example, suppose the last convolutional layer has 512 convolutional kernels, and you want to know how many votes each of these 512 convolutional kernels cast on the image being this object. The more convolutional kernels that vote, the more confident they are that the picture is this object, because the features they extract tend to the actual features of the object.

As shown in Figure 6, when doing class activation heat map visualization, the feature mapping of the class corresponding to the last convolution layer is first extracted to obtain the output element map of the final convolution layer, and then the convolution layer parameters

ISSN: 2664-9640

(including information on convolution kernels, gradients, etc.) are extracted, and then the global average pooling is applied to the gradients to convolve the input image, after which the element maps are multiplied with the corresponding pooling gradients, i.e., the output of the feature mapping is used relative to the The gradients are weighted (multiplied) for each filter, and finally the obtained results are normalized to output the class-activated heat map, which completes the whole process of visualizing the class-activated heat map.

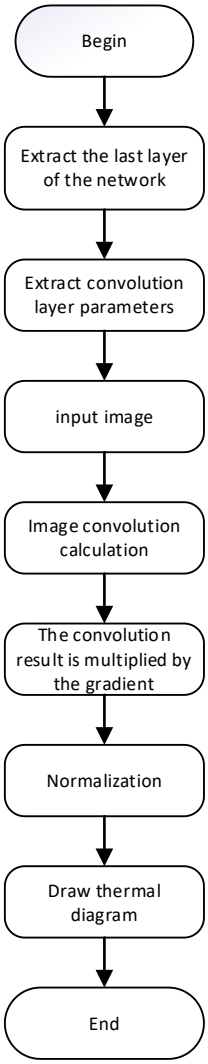


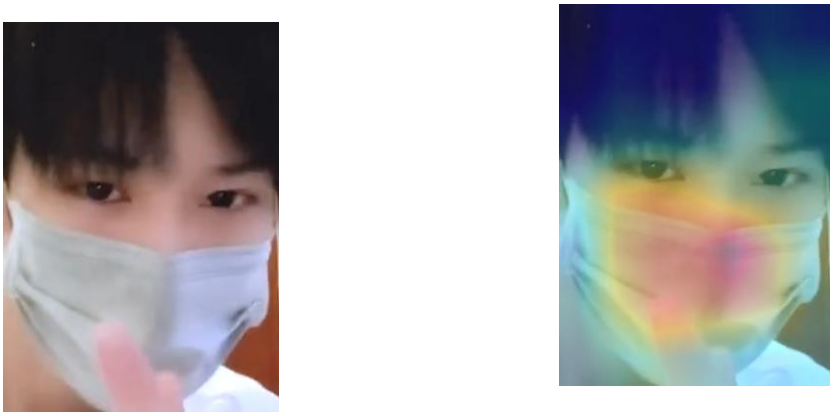
Figure 5 Class activation heat map visualization steps

The images in the dataset are visualized as class activation heat map using Pytorch-based ResNet-18 network, and the original image processing results and the class activation heat map visualization results are overlaid to output the result images. The original images are shown in Figure 6, 7, where Figure 6, 7(a) are both sample images from the original dataset, and Figure 6, 7(b) are the results of the class activation heat map visualization on the original image overlay, and it can be seen that the region with the highest heat (red) in the figure which is the focus of the model, and such region is concentrated at the mask of the face.

Compared with Figure 6, the human face in Figure 7 has eyes as distractors and there is some occlusion, but the region that the model focuses on is still concentrated at the mouthpiece, which further illustrates that the model has good results.

By comparing and analyzing the visualization results, we can judge whether the model's extraction judgment of different state image features is relatively reasonable, thus improving the reliability of the recognition model applied to real-life engineering scenarios.

ISSN: 2664-9640



(a) Original drawing (b) Class activation diagram

Figure 6 Original drawing and visualization results of glasses free samples

(a) Original drawing (b) Class activation diagram

Figure 7 Original drawing and visualization results of glasses wearing samples

4. Results

This research constructed a masked face classification model based on ResNet-18 network for the masked face recognition problem, and completed the model evaluation using the confusion matrix and the visualization of the model using CAM class activation heat map. The results show below .

1.The accuracy of the constructed model is 0.9864, its checking accuracy P is 0.9864, its checking completeness R is 0.9962, and its comprehensive measure F_1-score is 0.9913.

2.The class activation diagram shows that, the model achieves good results and its attention is focused on the masks on the class activation graph, which proves that the model has good generalization performance and has certain application value.

References

[1] Hinton, G. E. , and R. R. Salakhutdinov . "Reducing the Dimensionality of Data with Neural Networks." Science 313.

[2] Razavian, A. S. , et al. "CNN Features off-the-shelf: an Astounding Baseline for Recognition." 2014 IEEE conference on computer vision and pattern recognition workshops IEEE, 2014.

[3] Lecun, Y. , et al. "Comparison of learning algorithms for handwritten digit recognition." International Conference on Artificial Neural Networks 1995.

ISSN: 2664-9640

[4] Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E. Hinton. "Imagenet classification with deep convolutional neural networks." Advances in neural information processing systems 25 (2012): 1097-1105.

[5] Tang Xinyu, et al. "Face mask recognition system based on paddlehub." Journal of Hainan Normal University (NATURAL SCIENCE EDITION) 34.02 (2021): 177-184. Doi: CNKI: Sun: hnxz. 0.2021-02-011

[6] Ke Youxi, Ke Zhengtao, and Wu Yueping. "Contactless temperature measurement and mask recognition system based on openmv." automation and instrumentation. 05 (2021): 104-108. Doi: 10.14016/j.cnki.1001-9227.2021.05.104

[7] Zou Baihan, et al. "Research on the current situation of mask face detection methods based on lightweight CNN." software 41.08 (2020): 186-188. Doi: CNKI: Sun: rjzz.0.2020-08-051

[8] Liu Zixin. "Mask wearing detector based on embedded video stream." digital communication world. 02 (2021): 25-26 + 48. Doi: CNKI: Sun: sztj. 0.2021-02-009

[9] Ran Pengfei, and Liu Yinhua. "Wearing mask detection based on deep learning under complex light." automation and instrumentation 36.4:7

[10] Yosinski, J. , et al. "Understanding Neural Networks Through Deep Visualization." Computer Science (2015).

[11] Zeiler, Matthew D., and Rob Fergus. "Visualizing and understanding convolutional networks." European conference on computer vision. Springer, Cham, 2014.

[12] Simonyan, Karen, and Andrew Zisserman. "Very deep convolutional networks for large-scale image recognition." arXiv preprint arXiv:1409.1556 (2014).

[13] Szegedy C, Liu W, Jia Y, et al. “Going deeper with convolutions.” arXiv preprint arXiv:1409.4842 (2014)

[14] He, Kaiming, et al. "Deep residual learning for image recognition." Proceedings of the IEEE conference on computer vision and pattern recognition. 2016.

[15] “Real-World Masked Face Dataset” GitHub, Inc. January 2021. National multimedia software engineering technology research center of Wuhan University .<https://github.com/X-zhangyang/Real-World-Masked-Face-Dataset>

[16] Liu Zhiyong, et al. "Application of convolutional neural network visualization in process industry image recognition." Chemical automation and instrumentation 48.1:5