Mask Wearing Detection Based on InceptionV4 and Multi-Scale Retinex Image Enhancement Algorithm

Hongfei Zhang\textsuperscript{1,a}, Chun Liu\textsuperscript{2,b}, James Ho\textsuperscript{1,c} and Zhiwei Zhang\textsuperscript{1,*d}

\textsuperscript{1} School of Information Science and Engineering, Jinling College, Nanjing University, Nanjing, Jiangsu, China
\textsuperscript{2} School of Physics, Nanjing University, Nanjing, Jiangsu, China

Email: \textsuperscript{a}saraanberlin@gmail.com;\textsuperscript{b}liuchun@nju.edu.cn;\textsuperscript{c}james.ho@nju.edu.cn;\textsuperscript{*d}zhangzhiwei@nju.edu.cn

Abstract. The novel coronavirus is a contagious virus with a high mortality rate, and the international health emergency due to COVID-19 has never stopped since the outbreak, due to this situation, wearing masks has become a basic recommended public epidemic prevention approach for many countries. In the current shortage of medical human resources, we urgently need a non-artificial mask-wearing detection method. In this paper, the multi-scale Retinex algorithm is involved as the preprocessing step of the input image. The mask-wearing detection is based on the InceptionV4 convolutional neural network model. During the experiment, we compared and verified the superiority of the Inception part of the InceptionV4 model compared with the Stem structure of the GoogLenet model, and from LFW (Labeled Faces in the Wild), RMFD (Real-World-Masked-Face-Dataset) more than 10,000 samples were selected from the public datasets for model training. Finally, the precise rate reaches 97.3%.

1. Introduction
The COVID-19 is quite contagious and can spread from an infected person to others through the air by coughing and sneezing before handwashing \cite{1}. With the spread of the virus, wearing a mask has become a significantly important protective measure in crowded areas. Wearing masks can not only effectively protect the medical staff and community epidemic prevention personnel who are fighting to prevent the infection of the new crown virus, but also effectively reduce the infection rate of the virus among ordinary people. In public places such as hospitals, mask-wearing recognition has become a necessary work to fight the epidemic.

In the process of image sampling and quantization, the final quality of the image will be greatly affected by factors such as lighting and fog. Therefore, before entering the prediction of the neural network model, it is quite necessary to preprocess the image. This paper will briefly describe and realize how the multi-scale Retinex algorithm \cite{2} process digital images.

GoogLenet \cite{3} is a convolutional neural network model based on Inception structure and auxiliary classifiers launched by Google, which was winner of the classification competition, ILSVRC in 2014. Two years later, based on InceptionV3 of SOTA \cite{4} and residual structure, the team proposed the InceptionV4 structure \cite{5}. Compared with the GoogLenet model, the InceptionV4 model improves the Inception structure and proposes the Stem structure, which not only ensures a certain amount of calculation but also reduces the convolution complexity. This paper will make a certain comparison.
between the two neural networks and prove the superiority of the InceptionV4 model, and finally, use the InceptionV4 as the training model for mask wearing detection.

With respect to the theory mentioned above, a mask detection method is proposed, which is based on the InceptionV4 neural network and a preprocessing step involving the multi-scale Retinex algorithm. Experiments are carried out based on such theory and expected results are obtained.

2. Related work
Among the past works in the field of prediction, instead of using neural network models most of the researchers used the tricks of key point labelling, that is, labelling the feature points of the face and then identifying them [6]. However, during the prediction process, if the key points corresponding to the face are occluded, the algorithm will be failure because it will misjudge that covering is a mask. Since proposed by Alexnet [7] in 2012, the convolutional neural network has been rapidly developed, and the method of feature learning through the convolutional neural network has shown the better transferability than the traditional method of extracting and identifying fixed features. Instead of focusing on a single feature, researchers can allow the machine to learn multiple features to make fast and accurate judgments. In 2020, Huang, G.B. et al. involved VGG16 as a model to recognize and classify facial emotions. Their VGG16 model is trained on the KDEF dataset and achieved a precise rate of 88% [8].

Within a certain threshold, the recognition accuracy of the convolutional neural network is proportional to the depth and width of the network. However, when the width and depth are too large, the following two problems emerges:

I. The parameters of the neural network model will become too many. And the worse, if the training dataset is limited, overfitting is likely to occur, whereas if the size of dataset is too large, it will take very long training time.

II. In the process of forwarding propagation or backpropagation, the problems of gradient disappearance and explosion are likely to occur, which might lead to poor model generalization ability.

This article will discuss the significance of the Inception module for solving these two problems and point out the superiority of the InceptionV4 structure.

In engineering applications for detection, most of the sample image data for training purposes are obtained under normal conditions, which means the images distorted by exposure or real noise in the scene are quite normal. For example, images generated under bad weather conditions like snow or fog suffer from low contrast, faint light, and shifted luminance [9].

To solve the problem mentioned above, a pre-training model of a convolutional neural network based on the InceptionV4 structure and iteration process based on multi-scale Retinex are implemented for the final prediction purpose.

3. Dataset
Two particular datasets, Labeled Faces in the Wild Home and Real-World Masked Face Dataset, are involved as training samples in the experiment.

3.1. Labeled Faces in the Wild Home [10]
The LFW database contains images of 5749 different individuals, among which, 1680 people have two or more images while the remained 4069 have only one facial image. The images are available as the form of JPEG images of 250*250 pixels. Most images are in three color channels, i.e. the RGB format, although a few are grayscale only.

3.2. Real-World Masked Face Dataset [11]
Three categories of masked face datasets were proposed by RWMF. The official provider divides the database into two specific categories, of which first is gained from real world, whereas the second part is obtained through putting stimulated masks on face of people in the images. In this paper, about more than 10,000 images were randomly selected for training purposes.

Some training samples are shown as below (figure 1):
Figure 1. Four particular samples of the dataset

4. Method

4.1. The convolutional neural network InceptionV4
GoogLenet[3] and InceptionV4[5] are both CNN based on Inception and auxiliary classifiers proposed by Google. To highlight the advantages of InceptionV4, here comes the comparison of them in the aspect of structure [12].

In GoogLenet, the researchers paralleled multiple convolutional layers and one pooling layer in an Inception structure, and the sizes of these convolutional kernels were not the same. Such difference is meaningful because in the actual classification process, equalizing all the pictures is impossible due to the target scales of the pictures are different. Therefore, different convolution kernels for convolution operations should be used to generate different receptive fields and allow the machine to learn features of different scales.

The InceptionV4 model has made certain improvements to the Inception structure. Comparing the Inception structure of InceptionV4 (figure 2) with the Inception structure of GoogLenet (figure 3), the $5 \times 5$ convolution layer transforms into two $3 \times 3$ convolution layers, and the average pooling downsampling is used. In other Inception modules, several $n \times n$ convolution layers of GoogLenet have been converted into $1 \times n$ convolution layers and $n \times 1$ convolution layers of the same quantity.

Figure 2. Inception A module of InceptionV4
4.2 Multi-scale Retinex

After the image is quantized and sampled, the multi-scale Retinex algorithm is applied as image preprocess before it enters the neural network model.

4.2.1 Incident reflection model and representation of original image function. In the real world, the reflection surface-object-light source can be abstracted as the incidence-reflection model. In the case discussed in this article, the light source is generally sunlight while the face wearing the mask is considered to be the reflective surface, and the recognition device is treated as the observer.

For a given image function $G$, let the reflected image function be $R$ and the original image function be $O$, we transform the formula to the logarithmic domain [13]:

$$G(x, y) = R(x, y) \cdot O(x, y)$$

$$O(x, y) = G(x, y) / R(x, y)$$

(1)
\[ g(x, y) = \ln G(x, y) \]
\[ r(x, y) = \ln R(x, y) \]
\[ o(x, y) = \ln O(x, y) \]
\[ o(x, y) = g(x, y) - r(x, y) \]
\[ O(x, y) = e^{g(x, y) - r(x, y)} \] (3)

4.2.2. The iterative process of multi-scale Retinex algorithm. The processing function can be depicted as below:

\[ F_i(x, y) = \sum_{n=1}^{N} W_n \cdot \log [S_i(x, y)] - \log [S_i(x, y) \cdot M_n(x, y)] \] (4)

where \( i \in R, G, B \), which stands for Red, Green, Blue. \( N \) is the number of scales being used, for which \( W_n \) is the weighting factors.

The \( M_n(x, y) \) is the spiral and iterative functions given by

\[ M_n(x, y) = K_n \cdot e^{-\left(x^2 + y^2\right)/\sigma_n^2} \] (5)

where \( \sigma_n \) determines the scale follows Gaussian distribution. The Retinex offers different types of information with regard to the various scale, that is, features like color constancy are more likely to be given by those of large scale. In eq.5, the \( K_n \) are calculated and selected so that \( \iint F(x, y) \, dx \, dy = 1 \).

Each of the expressions within the summation in eq.4 and eq.5 represents a single SSR process [14].

5. Experiment

5.1. Comparison test of GoogLenet and InceptionV4

To verify the superiority of the InceptionV4 model compared to GoogLenet, we took the CIFAR100 dataset as the training sample and trained 200 epochs by the batch size of 128 with a single NVIDIA 3090RTX GPU. As shown in the charts, the Softmax loss of InceptionV4 (figure 6) is lower than the loss of GoogLenet (figure 7), but the accuracy rises to 76.36% of InceptionV4 (figure 8) from 72.15% of GoogLenet (figure 9) to, which is 4.21% higher.

![Figure 6. Loss of InceptionV4](image1)

![Figure 7. Loss of GoogLenet](image2)

![Figure 8. Accuracy of InceptionV4](image3)

![Figure 9. Accuracy of GoogLenet](image4)
5.2. InceptionV4 model training and Multi-scale Retinex preprocessing

To generalize better, we collected samples from the LFW (Labelled Faces in the Wild) face database from the computer vision laboratory of the University of Massachusetts, Amherst, USA, and the RMFD organized by Wuhan University. About 12,000 photos were selected in the data set for model training.

We built the InceptionV4 model (figure 4) on the training device and divided the dataset into training and validation sets in a 7:1 ratio. After building the convolutional neural network model, we carried out the training process in the environment of NVIDIA 3090 single GPU. The training was iteratively trained 200 times with a batch size of 64. It achieved an accuracy of about 96.7% (figure 10), and after adding Retinex preprocessing, the accuracy of the total sample came to 97.3% (figure 11).

6. Conclusion

This paper proposes a mask-wearing detection method that is preprocessed by the multi-scale Retinex algorithm and predicted based on the InceptionV4 convolutional neural network model training after the improved GoogLenet. It can be seen from the experimental results that, compared with GoogLenet, InceptionV4 reduces the complexity of convolution calculation while ensuring a certain computational complexity. What’s more, higher accuracy can be obtained in a shorter time. After adding the multi-scale Retinex algorithm for preprocessing, we can see from the confusion matrix that our method can better deal with how the reflection function negatively influences the images generated through quantization and sampling in the real world. At last, the accuracy of the model with multi-scale goes 0.6% higher, achieving the expected result on the data set.

References

[1] Humphreys, J. (2020). The importance of wearing masks in curtailing the COVID-19 pandemic. Journal of Family Medicine and Primary Care, 9(6), 2606.
[2] Rahman, Z.U., Jobson, D.J. and Woodell, G.A., 1996, September. Multi-scale retinex for color image enhancement. In Proceedings of 3rd IEEE international conference on image processing (Vol. 3, pp. 1003-1006). IEEE.

[3] Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V. and Rabinovich, A., 2015. Going deeper with convolutions. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 1-9).

[4] Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J. and Wojna, Z., 2016. Rethinking the inception architecture for computer vision. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 2818-2826).

[5] Szegedy, C., Ioffe, S., Vanhoucke, V. and Alemi, A.A., 2017, February. Inception-v4, inception-resnet and the impact of residual connections on learning. In Thirty-first AAAI conference on artificial intelligence.

[6] Mian, A.S., Bennamoun, M. and Owens, R., 2008. Keypoint detection and local feature matching for textured 3D face recognition. International Journal of Computer Vision, 79(1), pp. 1-12.

[7] Krizhevsky, A., Sutskever, I. and Hinton, G.E., 2012. Imagenet classification with deep convolutional neural networks. Advances in neural information processing systems, 25, pp. 1097-1105.

[8] Huang, G.B., Mattar, M., Berg, T. and Learned-Miller, E., 2008, October. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. In Workshop on faces in 'Real-Life' Images: detection, alignment, and recognition.

[9] Choi, L.K., You, J. and Bovik, A.C., 2015. Reference less prediction of perceptual fog density and perceptual image defogging. IEEE Transactions on Image Processing, 24(11), pp. 3888-3901.

[10] Huang, G.B., Mattar, M., Berg, T. and Learned-Miller, E., 2008, October. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. In Workshop on faces in 'Real-Life' Images: detection, alignment, and recognition.

[11] Wang, Z., Wang, G., Huang, B., Xiong, Z., Hong, Q., Wu, H., Yi, P., Jiang, K., Wang, N., Pei, Y. and Chen, H., 2020. Masked face recognition dataset and application. arXiv preprint arXiv: 2003.09093.

[12] Krizhevsky, A. and Hinton, G., 2009. Learning multiple layers of features from tiny images.

[13] Jobson, Daniel J., Zia-ur Rahman, and Glenn A. Woodell. "Properties and performance of a center/surround retinex." IEEE transactions on image processing 6.3 (1997): 451-462.

[14] Rahman, Zia-ur, Daniel J. Jobson, and Glenn A. Woodell. "Multi-scale retinex for color image enhancement." Proceedings of 3rd IEEE International Conference on Image Processing. Vol. 3. IEEE, 1996.