Journal of Image Processing Theory and Applications (2020) Vol. 3: 11-19
Clausius Scientific Press, Canada
DOI: 10.23977/jipta.2020.31002
ISSN 2560-6239

Mask detection algorithm based on pyramid box

Yueyuan Liu, Jun Liu and Kai Chen

College of information science and technology, Chengdu University of Technology (College of network security, Oxford Brooks College), Chengdu, 610059

Keywords: PyramidBox algorithm, Deep learning, Convolutional neural network, Face mask image detection

Abstract: The novel coronavirus outbreak there was no parallel in history. In early 2020, the outbreak of an unprecedented coronavirus outbreak could effectively block the spread of the epidemic in the crowd in the public areas of the society if it could use artificial intelligence technology to detect people without masks in the crowded area. This paper introduces the face mask image detection algorithm, mainly based on the face recognition algorithm pyramid box to complete, describes how to use Baidu's paddlehub mask detection project, through the use of Python language to complete the mask detection of multiple images and get accurate detection data. The accuracy rate of face detection based on the pyramid box mask detection model is not satisfactory. It can also be deployed to the server or even the mobile terminal to achieve rapid real-time detection. With the resumption of work of enterprises, I believe that the face mask detection scheme can solve many pain points that need to be solved for many enterprises, communities and manufacturers.

1. Introduction

At the beginning of this year, a new type of coronavirus (named 2019 ncov) broke out in Wuhan. After infected with the new coronavirus, respiratory symptoms, fever, cough, shortness of breath and dyspnea will appear obviously. If the virus develops further, it will lead to pneumonia, severe acute respiratory syndrome, renal failure, and even death of the infected person. Respiratory droplets transmission, aerosol and contact transmission are the two main transmission routes of new coronavirus. In order to deal with this virus infection, everyone wearing a mask is an effective means not only to protect themselves, but also to block the transmission path. Therefore, the government needs to closely monitor the wearing of masks in crowded places in public places, and prohibit people without masks from entering In public places, if manual inspection is adopted here, it will be time-consuming, laborious and dangerous, and it is easy to have many problems such as dense, slow crowd flow and supervision omission. The use of in-depth learning using camera detection is not only time-saving and efficient, but also safe supervision, which is an unusual economic and practical means that can be used on a large scale.

At the same time, Baidu company open source the first mask face detection model in the industry for free. The main function of the model is to accurately detect each face in the dense flow of people, and judge whether the person is wearing a mask. This face detection model is based on the pyromidbox algorithm developed in the paper of ECCV of international computer vision conference

in 2018. It can detect massive face data in public field, and quickly recognize faces with and without masks don’t label. I believe that this deep learning model can play a great role in this epidemic. It can not only timely detect and avoid risks in various public places, but also promote the government and enterprises to ensure safety in the resumption of work and production.

2. Deep learning

Convolution neural network is a kind of feedforward multilayer neural network, which has the ability of autonomous learning and can extract accurate features from a large number of labeled data. If we want to detect the visual pattern from the image pixels, we only need to carry on the simple preprocessing to the transmitted image, and can better identify the image object with more changes, and after the image is distorted and geometric transformation, its recognition ability is not easy to be affected.

In 1962, biologists studied the cat's brain. After a lot of experiments and analysis, the characteristic cells with the function of directional selection in the local area were called simple cells. The cells with such characteristics were similar to the filter of convolution network [1]. The concept of complex cell was also proposed in the research process, its function and aggregation layer in convolution neural network it’s very similar. Then Fukushima proposed and simulated a network model based on biological vision for the first time. The network model is composed of complex cell layer and serial simple cell layer, and many simple cell planes form a simple cell layer. Image features are extracted and convoluted with neurons in the plane and the previous cell layer to obtain [2]; many complex cell planes constitute a complex cell layer, It is mainly used to improve the deformation tolerance of simple cell layer and reduce the number of complex cell layer; convolution network is the output layer, and the types identified are represented by special neurons. After the network is simplified, the back-propagation algorithm is used to train the network in the supervised situation [3].

In the 1960s, convolutional neural networks using hardware began to appear. At that time, many companies and research institutions were committed to the research and development of hardware implementation neurons. Among them, Adaline developed the perceptron and neural network model, while Rosenblatt proposed the perceptron, and Widrow proposed the Adaline neural network model. From the model point of view, Adaline neural network and perceptron can be regarded as the same neuron model, which can input multiple inputs and adaptively change the strength of synapses (interconnection weights), but they have some differences in implementation methods. The synaptic strength is regulated by the sensory machine through the rotation of an electric motor, while the Adaline of Widrow uses the resistance to describe the strength of the synapse.

3. Convolution neural network

Convolution neural network consists of input layer, convolution layer, pooling layer (also called sampling layer), full connection layer and output layer. Convolution layer and pooling layer usually connect one or more fully connected layers and one output layer after multiple series connection (song Yue, 2017). Figure 1 shows the structure of convolution neural network

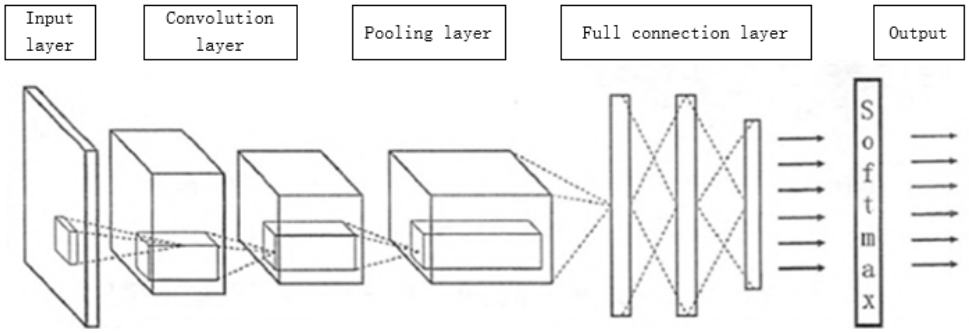

Figure. 1 Structure of convolution neural network

3.1 Convolution layer

Convolution layer is composed of some feature maps [11], which contains a large number of neurons, and neurons are connected with the local area of the upper layer of feature surface through convolution nucleus. The convolution kernel is a matrix (such as 3x3 or 5x5). The convolution layer of convolution neural network obtains the difference characteristics of input layer through convolution calculation, so the key module is convolution kernel. The design of convolution kernel needs to consider its step size and number. Generally speaking, the number of convolution kernels means the number of feature surfaces obtained from convolution filtering in the upper layer. If the number of feature maps is more, the learning ability of representative features is stronger and the space is larger, and the recognition effect is more accurate. However, if there are too many convolution kernels, the complexity of the network model will be greatly increased, and the calculation difficulty will be greater. Therefore, it is easier to over fit. Therefore, the number of convolution kernels should be determined according to the specific application data. For example, the number of convolution kernels extracted from MNIST database is generally between 20 and 50. However, if the same number of convolution kernels is used in cifar dataset, the effect is not good and needs to be redesigned.

After convolution calculation, the new feature surface needs to be mapped to a pixel value on the feature surface by using the activation function. Therefore, activation function plays an important role in convolution neural network algorithm. Activation functions are generally nonlinear and mainly include the following three functions. The mathematical expressions of the functions are as follows:

Sigmoid function:

$$f(x) = \frac{1}{1+e^{-x}} \tag{1}$$

Tanh function:

$$f(x) = \frac{e^{x}-e^{-x}}{e^{x}+e^{-x}} \tag{2}$$

Relu function:

$$f(x) = \{x, x > 0; 0, x \leq 0\} \tag{3}$$

3.2 Pooling layer

The pool layer follows the convolution layer and is composed of several characteristic surfaces [12]. The characteristic surface is the only feature surface on the upper layer, and the number will not change. The feature map obtained by convolution calculation is used to pool the image data in a

small range through pooling operation to extract new feature attributes. Through the upper pooling operation, the result will reduce the parameters (reduce the dimension of feature surface), and by enhancing the attributes of features, some feature attributes can be saved. Daily use of the maximum sampling, mean sampling, maximum sampling and random sampling, the characteristics of the three methods are shown in table 1.

Table 1 Characteristics of three common pooling methods

Pooling method	characteristic
Maximum sampling	Taking the largest feature point in the domain can preserve the image texture information
Mean sampling	The average value of feature points in the domain can keep the background information of the image
Random sampling	According to the probability of random selection of feature points, the larger the feature points, the greater the probability of acquisition

Figure 2 shows three common pool operation diagrams:

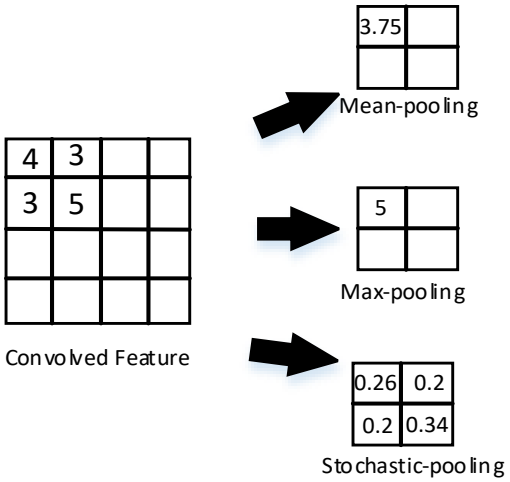

Figure. 2 schematic diagram of convolution and pooling operation

3.3 Full connection layer

In the convolution network model, after completing a series of convolution pooling operations, the whole layer is usually connected to the back. Each single neuron in the total connective layer connects all the neurons in the anterior layer, and the total connective layer also plays a role in distinguishing more specific local feature information in the convolution layer or pool layer. In order to improve the performance of convolution network, the excitation function of each single neuron in the full connection layer usually uses relu function. The output eigenvalues of the last full connection layer are transmitted to the output layer of network model, which can be classified by softmax regression, which can also be called softmax layer. For a specific classification task, it is very important to select the appropriate loss function.

When a large network model is used to train a few small image data sets, because of its high capacity, it is often not very good in saving the test feature set [13]. In order to prevent the over fitting phenomenon in the training process, the regularization method dropout technology is often used in the full connection layer. Even if the characteristic value of the neurons in the hidden layer

becomes 0 with a probability of 0.5, the node in the hidden layer will fail. Since a single neuron is not allowed to rely on other specified neurons to survive, this technique reduces the complexity of interdependence among neurons and makes neuron learning more robust. At present, the research methods of convolutional neural network usually adopt the relu + dropout method, and have achieved good classification performance. The full connection model is shown in Figure 3.

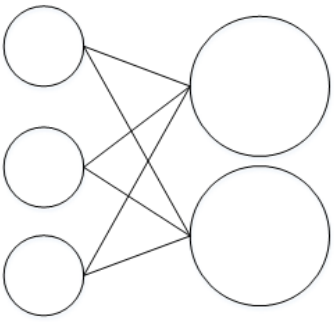

Figure. 3 Full connection model

4. Pyramid box algorithm

Pyramid box algorithm is a deep learning algorithm for face detection developed by Baidu. Pyramid box designs an effective data enhancement strategy (data anchor sampling) and a context based face detection module. In this article, we improve each part to further improve performance, including balanced data anchor sampling, dual pyramid anchor, and dense context module. This paper proposes an anchor based context assist method, pyramid anchors, which introduces supervised information to learn the context features of small, fuzzy and partially occluded faces; designs a low-level feature pyramid network (lfpn) to better integrate contextual features and facial features. At the same time, this method can deal with different scales of faces in a single shot; a context sensitive prediction model is proposed, which is composed of a hybrid network structure and a maximum input-output layer, which can learn accurate location and classification from the fusion features; a scale aware data anchor sampling strategy is proposed to change the distribution of training samples with less attention In the general face detection benchmark fddb and wider face, it achieves the current best level. The network structure is shown in Figure 4.

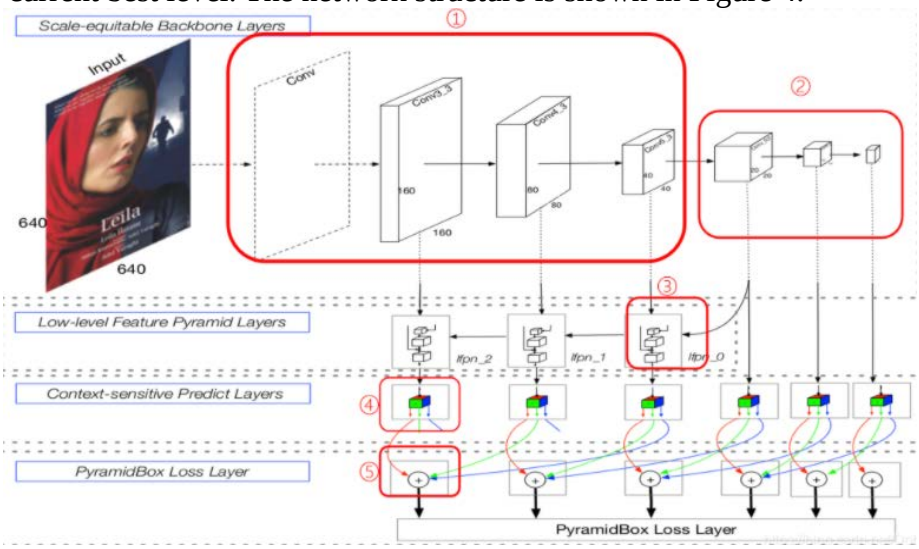

Figure. 4 Network structure diagram

5. Mask recognition model based on pyramid box algorithm

5.1 Pyramidbox_ Introduction and deployment of mask pre training mode

Pyramidbox_ lite_ mobile_ Mask model and pyramid box_ lite_ server_ Mask models have been trained by a large number of effective data sets, which can detect a large number of faces in the public scene with dense flow of people, and quickly recognize and label the faces with and without masks. Based on this pre training model, developers can quickly complete the model development of their own scenarios with only a small amount of their own data. According to the introduction of Baidu R & D Engineer, the mask face detection and classification model is composed of two functional units, which can complete the mask face detection and mask face classification respectively. After testing, the face detection algorithm based on the backbone network of faceboxes can significantly improve the recall rate by 30% with 98% accuracy. The face mask judgment model can be used to determine whether the face is wearing a mask, and the accuracy rate of mask identification is 96.5%, which meets the requirements of conventional mask detection. We and developers can optimize the quadratic model based on our own scene data, which can further improve the accuracy and recall rate of the model. The two models can be deployed on the server side and the mobile terminal respectively. With the help of paddlehub, the server-side deployment is very simple. A command line is used to start the mask face detection and classification model on the server_ lite_ server_ Mask-p88662. Deployed to the mobile terminal; paddlelite is the end-to-side reasoning engine of the propeller, which is specially for the mobile terminal model reasoning deployment. If it is necessary to embed the mask face detection and classification model into mobile devices such as mobile phones. There are only three steps to deploy mask face detection and classification model in mobile terminal: ① download prediction library, paddlelite will provide compiled prediction library; ② optimize model and use model_ optimize_ Tool tool is used to optimize the model; ③ the prediction API is used to realize the call.

5.2 Test and result of mask recognition model

In this paper, pyramidbox model is used to train and detect personal images. Firstly, the definition of prediction data is treated. In this step, images are read in and drawn as the standard form. If there are many images to be predicted and stored in a folder, then the pre training model is loaded. Paddlehub provides two groups of pre training models, namely, the pyramidbox_ lite_ mobile_ Mask and pyramid box_ lite_ server_ mask。 Here we usually use the pyramidbox_ lite_ mobile_ Mask model, pyramid box_ lite_ server_ the mask is used as a backup. The results are shown in Figure 5, 6, table 2 and 3.

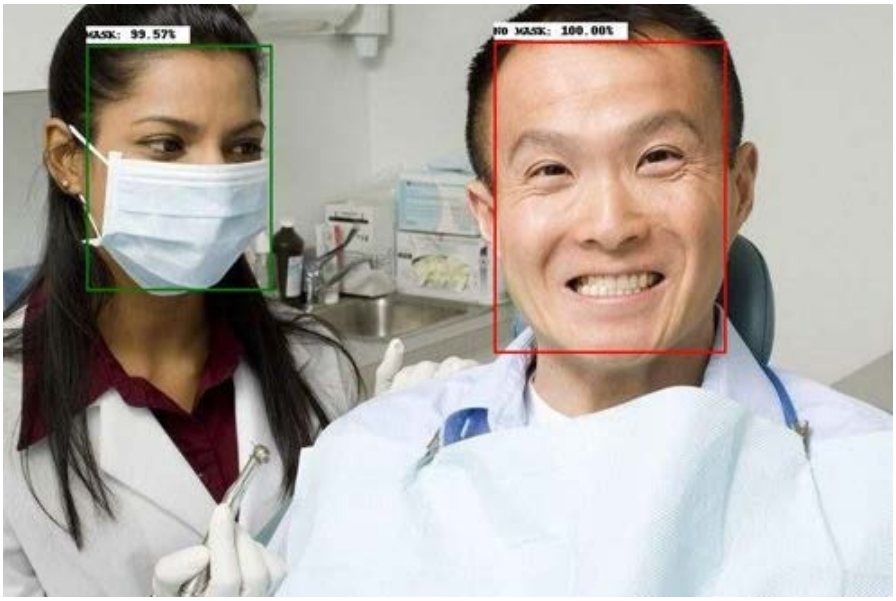

Figure. 5 Figure 1 prediction results

Table 2 Prediction results of test figure 1

0	1	2	3	4	5	6	7
NO	403.9669	35.65117	588.4654	282.7294	0.999981	1	./test_mask_detection.jpg
MASK	78.1618	38.74145	226.2013	232.7501	0.996196	1	./test_mask_detection.jpg

Figure. 6 Figure 2 prediction results

Table 3 Prediction results of test figure 2

	0	1	2	3	4	5	6	7
0	NO	495.371	227.75	506.19	241.73	0.99884	1	./test_mask_detection.jpg
1	NO	94.5027	217.34	105.99	231.86	0.99999	1	./test_mask_detection.jpg
2	NO	444.706	232.82	454.11	244.65	0.99168	1	./test_mask_detection.jpg
3	NO	66.8394	217.22	78.564	231.42	0.99904	1	./test_mask_detection.jpg
4	NO	208.512	225.94	218.97	239.72	0.99642	1	./test_mask_detection.jpg
5	NO	124.037	231.01	135.87	245.13	0.99997	1	./test_mask_detection.jpg
6	NO	401.184	235.09	410.97	247.36	0.99899	1	./test_mask_detection.jpg
7	NO	347.481	232.39	357.12	245.06	0.99805	1	./test_mask_detection.jpg
8	NO	146.841	233.09	158.1	247.28	0.99974	1	./test_mask_detection.jpg

In the model test, it is not difficult to find that in the prediction result of the first picture, the positive recognition rate is very high, close to 100%. However, although the recognition of the side face is successful, the recognition rate is not as high as that of the front. I also found this problem in many other scenes, that is, the front recognition is accurate, the side face is in the middle, there are 18 faces in the scene, but the final recognition loss of the model There are only 8 masks detection data, so we can clearly see that in complex scenes, such as when the face is small and more, the recognition accuracy of the model is not high, and there may be missing recognition and lack of recognition, which will cause relatively large interference to the actual use.

6. Conclusion

Artificial intelligence technology is a new revolution in the development of science and technology. Deep learning can solve many complex pattern recognition problems. My graduation project topic based on pyromidbox mask detection algorithm has achieved the expected goal and can use pyramid box the mask detection model can detect whether the face in the picture is wearing a mask quickly, and can give the accuracy of judgment. The mask detection model is very suitable for the current epidemic environment. After being deployed to the server and mobile terminal, it can make real-time judgment on the current image detection, and the effect is relatively significant. In the close-up scene, the basic recognition success rate is more than 90%, which shows the pyramid box face detection algorithm it is powerful, which proves that deep learning is of great help to human beings. The highlight of this model is that it can carry data sets to carry out more personalized training, and can also generate data results to facilitate developers' judgment. The disadvantage of this model is that the accuracy of recognition needs to be improved, especially in the case of more faces in public. More training of such scenes is needed to make the recognition more stable and efficient.

References

[1] Zhai Junhai, Zhang Sufang, Hao Pu. Convolutional neural network and its research progress [J]. Journal of Hebei University (NATURAL SCIENCE EDITION) (06): 85-96

[2] Chang Liang, Deng Xiaoming, Zhou Mingquan, et al. Convolutional neural network in image understanding [J]. Acta automatica Sinica, 2016, 42 (9)

[3] Wan shining. Research and implementation of face recognition based on convolutional neural network [D]. University of Electronic Science and technology, 2016

[4] Lu Hongtao, Zhang Qinchuan. A review of the application of deep convolution neural network in computer vision [J]. Data acquisition and processing (1): 1-17, total 17 pages

[5] Ke Xiaolong. Application of convolution neural network in image classification [D]. Shenzhen University

[6] Su Yue. Research and analysis of image recognition technology based on deep learning convolutional neural network [J]. Information and communication, 2019 (7)

[7] Gao Fei, Jiang Jianguo, an Hongxin, et al. A fast moving target detection algorithm [C] // abstracts of the 22nd National Conference on computer technology and application (cacis · 2011) and the 3rd National Conference on key security technologies and Applications (SCA · 2011). 2011
[8] Li Xudong, Ye Mao, Li Tao. A review of target detection based on convolutional neural network [J]. Computer Application Research (10): 7-12 + 17
[9] Huang Z. research on target detection model based on convolutional neural network [D]. Shanghai Jiaotong University
[10] Zhang zemiao, Huo Huan, Zhao Fengyu. A survey of target detection algorithms based on deep convolution neural network [J]. Minicomputer system, 2019, 40 (9)
11] Li Yandong, Hao Zongbo, Lei hang. A review of convolutional neural networks [J]. Computer applications (9): 2508-2515
[12] Zhou Feiyan, Jin Linpeng, Dong Jun. a review of convolutional neural networks [J]. Acta Sinica Sinica (6)
[13] Li Feiteng. Convolutional neural network and its application [D]. Dalian University of technology, 2014
[14] Ren S, He K, Girshick R, et al. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39 (6): 1137-1149.
[15] Liu W, Anguelov D, Erhan D, et al. SSD: Single Shot MultiBox Detector [J]. 2016.
[16] Zhang S, Zhu X, Lei Z, et al. S^3FD: Single Shot Scale-invariant Face Detector [J]. 2017.
[17] Lin T Y, Dollár, Piotr, Girshick R, et al. Feature Pyramid Networks for Object Detection [J]. 2016.
[18] Tang X, Du D K, He Z, et al. PyramidBox: A Context-assisted Single Shot Face Detector [J]. 2018.