YOLOv3_Slim for Face Mask Recognition

Xiaoming Jiang, Fugui Xiang*, Minghong Lv, Wei Wang, Zhonghua Zhang and Yi Yu
School of Bioinformatics, Chongqing University of Posts and Telecommunications, Chongqing, China

*Corresponding author’s e-mail: 1294978807@qq.com

Abstract. The use of face mask is advised by World Health Organization (WHO) for preventing transmission of Coronavirus disease 2019 (COVID-19). It is of great value to solve the multi-task object detection problem of non-wearing mask, wrong way wearing mask and standard wearing mask. In this paper, a network YOLOv3_Slim based on YOLOv3 is implemented. It’s faster than YOLOv3. Detection speed increased from 15.67 fps to 16.89 fps. In the mean time, we found the effect of the difference of inner class on the classification ability of the model. The large error of inner class will reduce the accuracy of the model and make the attention mechanism ineffective. So after changing the labels of the third data set, We add ECA module to our network. YOLOv3_Slim is more accurate than YOLOv4 in face mask recognition based on our data set. The mAP increased from 89.45% to 92.50%.

1. Introduction
Due to the impact of coVID-19 since 2019, experts advise people to wear masks during the epidemic in order to achieve joint prevention, control and reduce the risk of transmission and recurrence[1]. Wearing masks has become the focus of attention. Wearing a mask can effectively prevent harmful gases, smells, droplets, viruses and other airborne objects from invading the human body. Because of the wide range of people's activities and the complex distribution of workplaces, we cannot remind them to wear masks in real time just by manpower. It is a good choice to realize automatic recognition of wearing condition of mask by new technology.

In recent years, deep learning has made great progress in object detection, image classification, semantic segmentation and other fields. Combined with all kinds of algorithms of convolutional neural network, great progress has been made in both accuracy and speed. Among them, object detection is a multi-task learning problem combining object classification and object positioning.

1.1. Related work
Object detection can be realized by one-stage detector or two-stage detector[2]. The two-stage method includes the regional proposal stage and the positive sample classification and location stage. The one-stage method takes the background as one of the categories and regards the target detection as a regression problem. The detection speed of the one-stage detector is faster than that of the two-stage detector, which can often achieve real-time. Although one-step detector is usually less accurate than two-step detector, its detection accuracy is constantly optimized to meet the requirements of the industry. In the case of real time detection field, the focus is often on YOLO[3-6]. Yolov1 is the first one-stage detector based on deep learning, which is very fast. After continuous improvement, YOLO is almost the best choice to realize object detection.
The attention mechanism can further improve the accuracy of the one-stage detector while ensuring the speed. The attention mechanism enables the network to allocate attention to the area that is really useful for the tasks, thus improving the efficiency of the network. The common attention mechanisms are divided into three types: spatial attention model, channel attention model, and spatial and channel mixed attention model. Adding the attention mechanism to image classification tasks achieved great results[7,8]. SENet (Sequeeze and Net)[9] is the simplest channel attention mechanism. Many Attention modules improve SENet Block to introduce more complex channel Attention or mining spatial Attention, such as CBAM (Convolutional Block Attention Module)[10]. Empirical studies show that dimensionality reduction will bring side effects to channel attention prediction, and it is inefficient and unnecessary to capture the dependencies between all channels[11]. It also proposes the latest lightweight Attention mechanism, ECANet (Efficient Channel Attention Net). ECANet based on SENet, greatly reducing the number of parameters while maintaining high performance.

At present, there are many face mask recognition models. Tencent, Baidu, and Jingdong all reach a recognition accuracy of more than 99%. PaddleHub based implementation of mask wear detection application, the recognition accuracy can reach 80%-90% and 99% in certain cases at: https://github.com/ PaddlePaddle/ PaddleHub/ tree/ release/ v1.7/ demo/ mask_detection. A facial-eyebrow multi-particle mask face recognition model was proposed and the recognition accuracy on the data set reached 95%[12]. The recognition speed of these models is limited and the high detection accuracy can only be achieved in specific occasions.

1.2. Contributions
This paper proposes a single network that can accurately detect faces and at the same time classify the wearing conditions of masks. The net work is based on YOLOv3, but our network is more simple, faster and more accurate.

Our contributions in this paper are as follows:

- Inspired by Yolov4-Tiny, subsequent networks have reasonably deleted the first dimension feature extracted by Darknet-53.
- It is found that the improvement of the network detection accuracy by the attention mechanism module is affected by the difference of data within the same class, and the network detection accuracy is improved by modifying the category label and adding ECANet attention mechanism.
- A new data set of four categories was made according to the actual requirements of mask wear detection.

2. Method
A single network is implemented, which uses a convolutional neural network for face detection and classification to estimate the bounding boxes of faces and classify the wearing conditions of face masks at the same time. This is a multi-task learning progress. For the sake of overall simplicity and efficiency, we adopted YOLOv3 for multi-object detection.

2.1. YOLOv3
As is shown in Figure 1, YOLOv3 uses the backbone network called Darknet-53 for multi-scale training on full images. The backbone network which introduced the idea of feature pyramid was used to extract multi-scale characteristic. Three layers of features with different scales being extracted to predict boxes and detect objects of different sizes. The smaller scale feature layer was upscaled and transformed to the same size as the previous feature layer through deconvolution, and then concatenated. In this way, the information between the three feature layers with different scales can be better connected. By obtaining more fine-grained information from the early feature map, and then conducting convolution processing, tensors with different scales of the same feature can be obtained.
2.2. ECANet
In order to ensure the speed and improve the accuracy of the model, the lightweight attention mechanism ECANet is introduced. The structure of ECANet is shown in Figure 2. After the ECA module uses the global average pooling aggregated convolution feature without dimension reduction, it firstly determines the kernel size $K$ adaptively, then carries out one-dimensional convolution, and then learns channel attention through sigmoid function.

![Figure 1](image1.png)

**Figure 1.** The network structure of YOLOv3.

![Figure 2](image2.png)

**Figure 2.** Efficient channel attention (ECA) module.

Experiments show that the network performance is better without dimension reduction. Visualizing channel characteristics finds that capturing dependencies between all channels is inefficient. Therefore, the author only considers the information exchange between the current channel and its $k$ neighborhood channels. The number of parameters for each channel is $C$, then the number of parameters is $k \times C$. And the parameters of all channels are Shared by the following formula, as in Equation (1)
\[ \omega_i = \sigma(\sum_{j=1}^{k} w_i^j y_j^i), \ y_j^i \in \Omega_i^j, \]  

(1)

Where \( \Omega_i^j \) indicates the set of \( k \) adjacent channels of \( y_j^i \). \( \sigma \) is the activation function. This strategy can be easily and quickly implemented by a one-dimensional convolution with kernel size of \( k \), as in Equation (2)

\[ \omega = \sigma(C1D_\omega(y)), \]  

(2)

where \( C1D \) indicates 1D convolution. ECA module makes the final number of parameters is \( k \) by invoking Equation (2).

2.3. YOLO_Slim

A leaner, more efficient multi-task learning method based on YOLOv3 is found. As is shown in Figure 3. We add ECA module to the end of every residual block of Darknet-53, and get the backbone called ECA_Darknet-53 which extracts the fine-grained features of the two dimensions. We used the features of the last two scales extracted by ECA_Darknet-53, and then carried out subsequent further feature processing. Because we just need to detect large and medium-sized faces in a practical application scenario. And larger objects can be further addressed to achieve new tasks. For example, in the actual detection task, we adopt thermal imaging camera to realize the multi-task combination of mask detection task and temperature measurement task. The thermal imaging camera can accurately measure the body temperature within a certain distance, and the face images collected within this range are mainly concentrated on the larger scale.

![Figure 3. The structure of YOLOv3_Slim.](image)

The whole network is named YOLOv3_Slim. It’s more suitable for multi-task learning. The usual practice in training tasks is that a sample predicts only one value, such as a category in a classification, a real value in a regression. is an innovative device that simultaneously predicts multiple output structures. Specifically, multi-scale feature extraction was performed on the input image, then 2 feature layers were obtained. Each feature layer predicted 3 prior boxes, and each prior box predicted 4 bounding box offsets, 1 object probability and probability of each category. Since the mask data sets were divided into 3 categories, the tensor scale of each feature layer was \( N \times N \times [2 \times (4+1+3)] \). Through multiple prediction results, YOLOv3_Slim can simultaneously detect faces and classify wearing masks by coordinate regression and class prediction.
YOLOv3_Slim directly uses the feature layer as the output, making full use of the different channels of the feature map to express not only the features of the image, but also the coordinates and confidence of the object.

To better fit the detecting task, we use the bounding boxes of the data sets to re-cluster anchors and get nine different scales of anchors. The smallest three anchors was removed and other six anchors priors were used for the two output dimensions. On the mask dataset the 9 clusters were: \((13 \times 22), (32 \times 47), (47 \times 86), (70 \times 60), (72 \times 138), (104 \times 92), (112 \times 222), (148 \times 136), (227 \times 270)\).

3. Experiment

In this section, we strictly test the detection speed and accuracy of the proposed network with our own data set, and compare the detection results of some excellent detectors on our data set under the same environment.

3.1. Data set selection

After considering the actual wearing conditions, we find it is difficult to distinguish the types of masks, because the variety of masks in real life makes it difficult to collect data sets. Although some masks fail to meet the requirements of disease prevention and it is of certain significance to classify the types of masks, it is not recommended to distinguish the types of masks in consideration of the difficulty of practical operation.

According to the wearing condition of the mask, firstly, data sets in Figure 4 are divided into three classes: Nomask(non-wearing mask), Wrmask(Wrong way wearing mask) and Swmask(Standard wearing mask). Then as is shown in Figure 5, through follow-up experiments, the wrong way wearing mask was finally divided into only exposing the mouth or nose called Wrmask1(Wrong way wearing 1) and exposing the mouth and nose called Wrmask2(Wrong way wearing 2). Because the two types of data sets were show obvious characteristics of difference, separate them helps reduce loss, improve model accuracy.

![Figure 4. Annotation examples of three classes of data sets.](image)

![Figure 5. Annotation examples of three classes of data sets.](image)
We collected more than 8,000 data sets by online fetching, the volunteering help from laboratories and soliciting, including more than 3,500 pictures for wearing masks, more than 800 pictures for non-standard wearing masks, and more than 4,000 pictures for not wearing masks. As is shown in Figure 4, following the format of VOC data sets, we simply annotate face areas for three types of data sets. All of the data sets were pictures of real faces.

3.2. Experimental environment
All of our code is based on pytorch. We train the detector on the server. The CPU is Intel(R) Xeon(R) Gold 6148 CPU @ 2.40GHz. The GPU is Tesla V100-PCIE-32GB. And then we test the model on the notebook computer. The CPU is Intel(R) Core(TM) i7-8750H CPU @ 2.20GHz. The GPU is GeForce GTX 1060 Ti. The GPU memory is 6GB.

The software is operated by Visual Studio Code under Anaconda3.6 platform. Python version is 3.6.10. Cuda10.0.130 and Cudnn7.4.1.5 are installed to help GPU to accelerate computing. Meanwhile, third-party libraries such as Tensorflow-gpu1.13.2 and pytorch 1.2.0 are installed in the software to support code running.

3.3. Training and interpretation of results
The experiment 1 was designed to investigate the three classes: Nomask, Wrmask and Swmask. For each class of the three classes, we first randomly selects 500 images as the training sets and 100 images as the testing sets. Then we trained our datas with many models such as YOLOv3, YOLOv3_Slim, YOLOv4, YOLOv4-tiny on the server, then tested it on a single GPU laptop. Experiment 2 only changed the labels of the third data set into two classes: Wrmask1 and Wrmask2. And then do the same things as experiment 1.

We can find YOLOv3_Slim worked pretty well from table 1. In experiment 1, the ECA module didn’t achieve the desired effect, but reduced the accuracy of the model. Then we examined the data set and found that there were large intra-class differences among those who did not wear masks in a standard way. Therefore, we assumed that these differences would be magnified by the attention mechanism due to the large intra-class errors, which reduced the classification ability of the model. Therefore, we only changed the labels of the training set and test set of the third class of data and divide this data set into two classes.

| Methods       | Backbone      | Experiment 1 Three classes of data sets | Experiment 2 Four classes of data sets |
|---------------|---------------|----------------------------------------|---------------------------------------|
|               | mAP | FPS | mAP | FPS |
| YOLOv3        | 80.83%| 15.89 | 79.34 | 15.67 |
| YOLOv3_Slim   | 87.17%| 18.19 | 91.71 | 17.39 |
| YOLOv3_Sim    | 84.64%| 14.97 | 93.06 | 14.65 |
| YOLOv3_Slim   | ECA_Darknet-53 | 84.28%| 15.46 | 92.50 | 16.89 |
| YOLOv4_Tiny   | CSPDarknet-53 | 77.65%| 24.52 | 74.06 | 27.34 |
| YOLOv4        | CSPDarknet-53 | 85.92%| 12.20 | 89.45 | 11.74 |

After changing the labels of the third data set, YOLOv3_Slim is faster and more accurate than YOLOv3 and YOLOv4. Though the SE_YOLOv3_Slim is more accurate than the ECA_YOLOv3_Slim, the higher computational overhead makes it slower. As is shown in Figure 6, the ability of subclass classification of YOLOv3_Slim has been greatly improved after add ECA module.
In the actual detection experiment, the detection results of YOLOv4 and YOLOv3_Slim were compared. As shown in Figure 7, more overlapping tags exist in the detection results of YOLOv4, which affects the detection results. The YOLOv3_Slim classification is more accurate, even if there is a false result. By balancing speed and precision, ECA_YOLOv3_Slim is the best choice.

4. Conclusion
Our network can accurately detect faces and classify the wearing conditions of masks, which can still achieve good results even under dense population and various lighting conditions. We're using YOLOv3_Slim. By using the two scales of features extracted by ECA_Darknet-53, it’s faster than YOLOv3. Detection speed increased from 15.67 fps to 16.89 fps. In the mean time, we found the effect of the difference of inner class on the classification ability of the model. The large error of inner class will reduce the accuracy of the model and make the attention mechanism ineffective. So after changing the labels of the third dataset, YOLOv3_Slim is more accurate than YOLOv4. The mAP increased from 89.45% to 92.50%. This further verifies that the attention module can effectively improve the classification ability of subclasses of the classification network.

References
[1] Tso R, Cowling B. Importance of face masks for COVID-19: a call for effective public education[J]. Clinical Infectious Diseases, 2020.
[2] Zou Z, Shi Z, Guo Y, et al. Object Detection in 20 Years: A Survey[J]. 2019.
[3] Redmon J, Divvala S, Girshick R, et al. You Only Look Once: Unified, Real-Time Object Detection[J]. 2015:779-788.
[4] Redmon J, Farhadi A. YOLO9000: Better, Faster, Stronger[C]// IEEE Conference on Computer Vision & Pattern Recognition. IEEE, 2017:6517-6525.
[5] Redmon J, Farhadi A. YOLOv3: An Incremental Improvement[J]. arXiv e-prints, 2018.
[6] Bochkovskiy A, Wang C Y, Liao H Y M. YOLOv4: Optimal Speed and Accuracy of Object Detection[J]. 2020: 1-17.
[7] Jaderberg M, Simonyan K, Zisserman A. Spatial transformer networks[C]//Advances in neural information processing systems. 2015: 2017-2025.
[8] Fu J, Zheng H, Mei T. Look Closer to See Better: Recurrent Attention Convolutional Neural Network for Fine-Grained Image Recognition[C]// IEEE Conference on Computer Vision & Pattern Recognition. IEEE, 2017: 4438-4446.
[9] Hu J, Shen L, Sun G. Squeeze-and-excitation networks[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 7132-7141.
[10] Woo S, Park J, Lee J Y, et al. Cbam: Convolutional block attention module[C]// Proceedings of the European Conference on Computer Vision (ECCV). 2018: 3-19.
[11] Wang Q, Wu B, Zhu P, et al. ECA-Net: Efficient Channel Attention for Deep Convolutional Neural Networks[C]// 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2020.
[12] Wang Z, Wang G, Huang B, et al. Masked Face Recognition Dataset and Application[J]. 2020.

Acknowledgments
This paper is sponsored by the National key R & D Program of China(2020YFC2003301), the National Natural Science Foundation of China(61801069) and Chongqing Human Resources and Social Security Bureau's "Residential Development Plan” Innovation Project(cx2017011).