2021 IEEE 2nd International Conference on Big Data, Artificial Intelligence and Internet of Things Engineering (ICBAIE 2021)

2021 IEEE 2nd International Conference on Big Data, Artificial Intelligence and Internet of Things Engineering (ICBAIE) | 978-1-6654-1540-8/20/$31.00 ©2021 IEEE | DOI: 10.1109/ICBAIE52039.2021.9389953

Face mask detection based on Transfer learning and PP-YOLO

Wang Jian *
Software Engineering Institute of Guangzhou
Guangzhou, China
e-mail: wjn@sise.com.cn

Lin Lang
Software Engineering Institute of Guangzhou
Guangzhou, China
e-mail: ll1831@scse.com.cn

***Abstract*—Taking into account the accuracy and reasoning speed of face mask detection tasks, this paper proposes a detection model PP-YOLO-Mask based on PP-YOLO through transfer learning, data augmentation and model compression methods. In a public scene, the mask detection mAP of this model is 86.69%, and the single picture recognition speed is 11.842ms. Experimental results show that compared with YOLOv3 and FasterRCNN, the model has faster accuracy and detection speed.**

Keywords-component; Mask Detection; PP-YOLO; Transfer Learning; Data augmentation; Model compression formatting;

I. INTRODUCTION

Beginning in December 2019, the novel coronavirus pneumonia (COVID-19) epidemic has broken out in various parts of the world, posing a huge threat to human health and social security[1]. Wear a face mask can effectively prevent the spread of respiratory infectious diseases[2], so it is necessary to supervise and manage the wearing of face mask in public places. Although general target detection algorithms have made some progress in other fields [3-5], there are currently fewer relevant algorithms applied to face mask detection[6], For this, peoples have carried out relevant research, Loey Mohamed et al. [7] and others used YOLO V2 to construct the detection model; Wang Yihao et al. [8] and others introduced an improved spatial pyramid pooling structure in the YOLOv3 algorithm to achieve enhancement of feature extraction; Niu Zuodong et al. [9] improved the Retina Face algorithm, added the task of detecting face mask wearing, and introduced an improved self-attention mechanism in the feature pyramid network; Wang Bing et al. [10] proposed a Based on the improved YOLOv4-tiny lightweight network algorithm, the Max Moudle structure is added to obtain the main features of more targets..

This task is generally deployed on edge computing nodes (ECN, EdgeComputingNode, ECN). ECN belongs to low computing resources and low power consumption devices [11], so the mask detection task needs to take into account the requirements of accuracy and inference speed. The above research basically started from the problems of target occlusion, dense crowds, and small-scale target detection, and did not consider the balance of accuracy and reasoning speed. The PP-YOLO has inherent advantages in accuracy and reasoning speed [12]. Based on PP-YOLO, this paper proposes a PP-YOLO-Mask model suitable for mask wearing detection tasks. Using methods such as transfer learning and MixUp, the model is constructed on a small sample data set, and finally the model is lightweighted by methods such as model distillation and clipping. Experimental results show that PP-YOLO-Mask is superior to other mainstream target detection algorithms in detection rate and inference speed.

II. DATA SET AND DATA AUGMENTATION

There is currently no public data set of face mask wearing. This article crawls Internet data and manually annotates it according to the Pascal VOC standard. The data set contains 7959 pictures and 16,635 annotations. The annotations include Mask and NoMask categories. The number of mask labels is 7024, and the number of nomask labels is 9611. The label information mainly includes the center coordinates, length, and width of the target frame. An example is shown in Figure 1, where 70% is used for training, 20% is used for verification, and 10% is used for testing.

Figure 1. An example of a data set image

Deep learning relies on big data. Insufficient amount of data will cause problems such as overfitting of the model. Data Augmentation (DA) and regularization technology can alleviate this problem[13]. In addition to the conventional random scaling, cropping, expansion, contrast and other methods, we also uses Mixup strategy for data enhancement. It is a simple data enhancement principle that has nothing to do with the data[14]. For the input images of a batch of pictures to be tested, it is fused with randomly selected pictures, An example is shown in Figure 2.

978-1-6654-1540-8/21/$31.00 ©2021 IEEE

Figure 2. Mixup example

III. PP-YOLO Model Structure

he main consideration of PP-YOLO is to balance the accuracy and speed of the model to a certain extent, rather than designing a novel detection model. It is implemented based on YOLOv3 and improves the accuracy of the model by adding a large number of tricks.

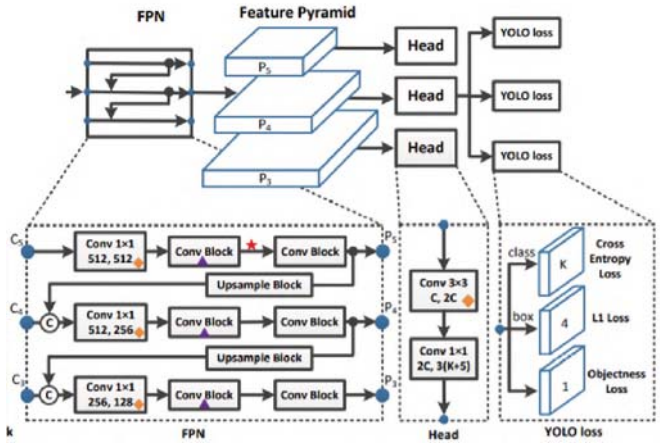

Figure 3. Model structure

The PP-YOLO structure is shown in Figure 3. The backbone network backbone is replaced with ResNet50vd-DCN, because the Deformable Convolution Network (DCN) effect has been proven to have better results in spatially deformed image recognition tasks [15]. FPN is used in Detection Neck, and the output of the 3, 4, and 5 stages of the backbone is used as the input of FPN, and the output size becomes half of the input size. Detection Head contains two convolutional layers: 3x3 convolution followed by 1x1 convolution to get the final prediction. The number of predicted output channels is 3 (K+5), where K represents the number of categories. Three different anchor points are assigned to each position of the predicted feature map. For each anchor point, the first K channels represent the predicted probability of the K category, the subsequent 4 channels represent the position, and the last channel represents the target score. The classification and positioning loss functions are CrossEntropyLoss and L1 Loss respectively.

IV. PP-YOLO-Mask Model Construction

Due to insufficient data samples, this paper adopts the transfer learning strategy for modeling. The process is shown in Figure 4. The PascalVOC data set is used to pre-train PP-YOLO. The data set enhanced by MixUp data is sent to the pre-training model for migration training, and the fully connected layer and related hyperparameters of the network are adjusted to obtain the original training model. The final model is obtained through model compression and conversion.

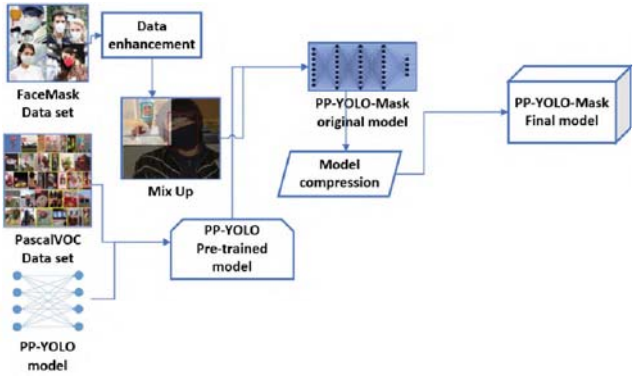

Figure 4. Model building process

A. Transfer Learning Strategies

Fine-tune (Fine-tune) is a common migration learning strategy[16]. By modifying the structure of the pre-training network model, selectively loading the weight of the pre-training network model, and then retraining the model with your own data set is fine-tuning. Fine-tune can quickly train a model with a relatively small amount of data. We respectively freezes the four layers of stage1-stage4 of the PP-YOLO pre-training model, and uses a small number of samples for verification. It is found that when stage1-stage3 is frozen, stage4 and the final fully connected layer are retrained, the mAp value is the best.

B. Model training

We use stochastic gradient descent to train the neural network, and use a Exponential Moving Average, which can improve the performance of the final model on the test data to a certain extent. Compared with the direct assignment of variables, the value obtained by the moving average is smoother and smoother on the image, and the jitter is smaller, and the moving average will not fluctuate greatly due to an abnormal value. Let the weight v be denoted as v_t at time t, θ_t is the value of the weight v at time t, that is, when the moving average model is not used, $v_t = \theta_t$, after using the moving average model, the update formula of v_t is as shown in (1). Show: $\beta \in [0,1)$. But in exponential averaging, when $t = 0$ and β is relatively small, v_0 is a relatively small value, so (2) is introduced to correct this situation, that is, when the number of iterations t is relatively small, the deviation of v from itself is not too large.

$$v_t = \beta \cdot v_{t-1} + (1 - \beta) \cdot \theta_t \quad (1)$$

$$v\_biased_t = \frac{v_t}{1-\beta^t} \tag{2}$$

C. Model Compression

We performed post-training quantization on the weight of the model. Model quantization is the process of approximating floating-point model weights or fixed-point tensor data flowing through the model to a finite number of discrete values with a lower loss of inference accuracy. It is a data type with fewer digits for approximate representation The process of 32-bit limited-range floating-point data achieves the goal of model compression. Under the premise of ensuring mAP, the quantized model only reduces by 56.4mb, and the quantization effect is good.

V. Experiment And Result Analysis

The environment configuration is as follows, GPU: Tesla V100. Video Mem: 16GB, the operating system is Ubuntu 16.04.6, and the PaddlePaddle deep learning framework is used. The training parameters are as follows: The optimizer is Momentum, the number of iterative training samples is 128, and the number of iterations is 120; the learning rate is 0.00008; the learning rate decay rate is 0.001; the evaluation method during training is COCO. And compared with YOLOV3 and FasterCN. Choose average precision (Average Precision, AP) and prediction time (ms/image) to objectively evaluate the effect of this algorithm on face and mask wearing detection, in which the precision rate, precision rate and recall rate are calculated as shown in (3).

$$AP = \int_0^1 p(r)dr \tag{3}$$

$p(r)$ represents the mapping relationship between precision rate and recall rate. The calculation method of precision rate P and recall rate R as (4), (5).

$$P = \frac{T_P}{T_P + F_P} \times 100\% \tag{4}$$

$$R = \frac{T_P}{T_P + F_N} \times 100\% \tag{5}$$

Take the mask label category as an example, where T_P represents the number of masks that the model will detect as masks, F_P represents the number of nomask targets detected as masks, and F_N represents the number of masks detected as nomasks. When IoU=0.5, the corresponding relationship between the accuracy of the mask and nomask of the prediction frame and the recall rate, and the corresponding relationship between the recall rate and the confidence threshold are shown in Figure 5.

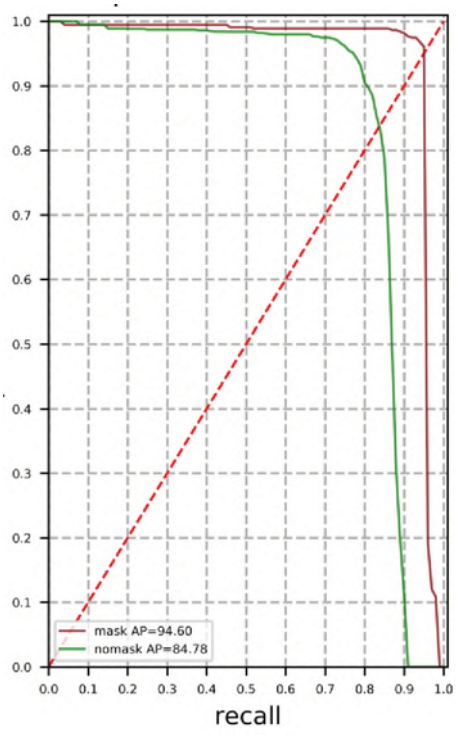

Figure 5. Correspondence between accuracy rate and recall rate

The performance index comparison of different detection models is shown in Table I. Experiments show that PP-YOLO-Mask is better than YOLOV3 and FasterRCNN in terms of detection accuracy and speed. Although the model of PP-YOLO-Mask is relatively large, considering the actual situation of the deployed equipment, this indicator has little effect on the overall performance of the model.

TABLE I. PERFORMANCE COMPARISON OF DIFFERENT MODELS

Model	Size	ms/image	maskAP	nomaskAP	mAp
PP-YOLO-Mask	203.1	11.842	94.60	84.78	89.69
YOLOV3	219.2	20.252	93.65	77.86	85.76
FasterRCNN	247.8	25.292	90.41	74.62	85.52

As shown in Figure 6. The analysis chart shows 7 Precision-Recall (PR) curves. Each curve represents the Average Precision (AP) higher than the one on the left because of the gradual relaxation of evaluation requirements. C75: PR curve when IoU is set to 0.75; C50: PR curve when IoU is set to 0.5, the area of different colors represents the AP gain brought by the relaxation of IoU.

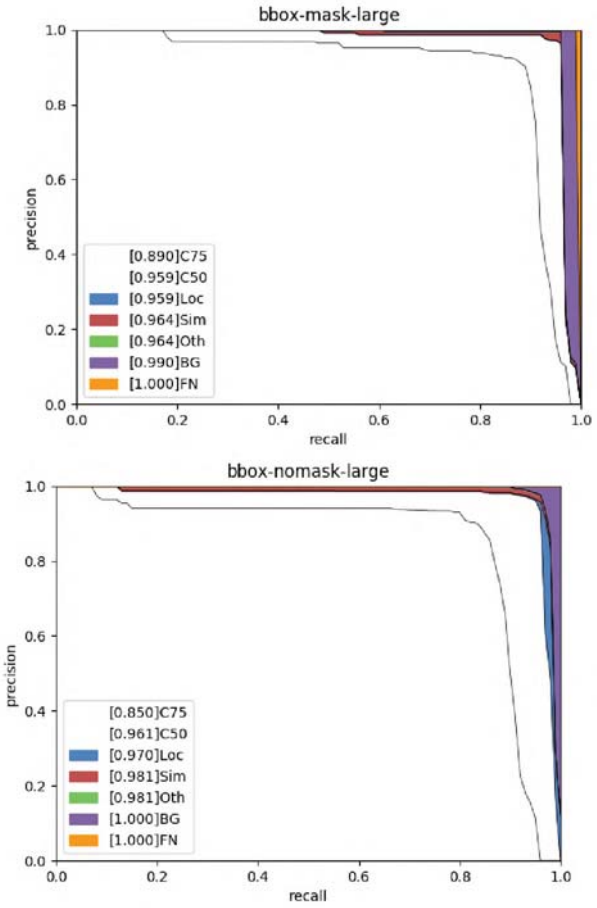

Figure 6. PR curve

VI. CONCLUSION

In this paper we proposes a mask wearing detection method based on the PP-YOLO algorithm, establishes a mask wearing detection data set MaskData, uses the Fine-tune method and the EMA strategy to transfer learning to the original PP-YOLO, and uses quantification to perform the model Compress, and finally get the PP-YOLO-Mask model. The experimental results show that the PP-YOLO-Mask model proposed in this paper has a mAP of 86.69%, and has achieved good detection results in the detection of mask wearing in public scenes. Compared with other mainstream models, the model in this paper has better accuracy and reasoning speed. Has a good application prospect in the prevention and control of the new crown epidemic.

ACKNOWLEDGMENT CONCLUSION

1.Scientific research project of Software Engineering Institute of Guangzhou(ky202009)

2.college students' innovation and entrepreneurship training plan in 2020 (202012618006)

REFERENCES

[1] TANG Long,YAO Liqing. “Progress in clinical research related to coronavirus disease 2019 ”[J/OL]: West China Medical Journal | West Chin Med J, pp. 1–6.

[2] HE Junmei, WEI Qiuhua, REN Zhe, etc. Selection and use of respirators in New Coronavirus pneumonia prevention and control [J]. Chinese Journal Of Disinfection, 2020, 37(2): 137-141.

[3] Girshick R, Donahue J, Darrell T, et al. Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation[C]// CVPR. IEEE, 2014.

[4] SSD:Single shot multibox detector. Liu W,Anguelov D,Erhan D,et al. Computer Vision-ECCV 2016 . 2016

[5] J. Redmon, S. K. Divvala, R. B. Girshick, and A. Farhadi. You only look once: Unified, real-time object detection. In IEEE CVPR, pages 779–788, 2016.1, 2

[6] CAO Chengshuo, YUAN Jie Mask wearing detection method based on YOLO-Mask [J]. Laser & Optoelectronics Progress | Las Optoelect Prog: 1-13.

[7] Mohamed L,Gunasekaran M,N TMH, et al. A Hybrid Deep Transfer Learning Model with Machine Learning Methods for Face Mask Detection in the Era of the Covid-19 Pandemic.[J]. Measurement: Journal of the International Measurement Confederation, 2020, 167.

[8] WANG Yihao,DING Hongwei,LI Bo,et al. Mask Wearing Detection Algorithm Based on Improved YOLOv3 in Complex Scenes [J]. Computer Engineering, 2020, 46(11): 12-22.

[9] NIU Zuodong, QIN Tao, LI Handong, CHEN Jinjun. Improved Algorithm of Retina Face for Natural Scene Mask Wear Detection [J]. Computer Engineering and Applications | Comput Eng Appl, 2020, 56(12): 1-7.

[10] WANG Bing,LE Hongxia,LI Wenjing,et al. Mask detection algorithm based on improved YOLO lightweight network [J]: 1-11.

[11] WANG Jian,Liu Xuehua. Pathological Recognition Of apple Leaves Based On Deeply Separable Convolution [J]. Computer Systems & Applications | Comput Syst Appl, 2020, 29(11): 190-195.

[12] Long X , Deng K , Wang G , et al. PP-YOLO: An Effective and Efficient Implementation of Object Detector[J]. 2020.

[13] 9 Vinyals O, Blundell C, Lillicrap T, Koray K. Matching net-works for one shot learning. In: works for one shot learning. In: Proceedings of the 30thInternational Conference on Neural Informati International Conference on Neural Information ProcessingSystems. Barcelona, Spain:MIT Press, 2016. 3630-3638

[14] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz.mixup: Beyond empirical risk minimization. arXiv preprintarXiv:1710.09412, 2017. 5, 6, 13

[15] Jifeng dai,Haozhi qi,Yuwen xiong. Deformable Convolutional Networks[C]//2017 IEEE International Conference on Computer Vision (ICCV), IEEE, 764-773.

[16] Zongwei, Zhou, Jae, et al. Fine-tuning Convolutional Neural Networks for Biomedical Image Analysis: Actively and Incrementally.[J]. Proceedings IEEE Computer Society Conference on Computer Vision & Pattern Recognition, 2017.