DOI: 10.54254/2755-2721/6/20230833

Research on Yolo-based target recognition detection

Zhang hao

XingYuanHuaYuan Residential Quater, No. 98, ZiBo Road, donggang District,
Rizhao City, Shandong Province, P. R. China, 276800

misaky0721@gmail.com

Abstract. With the development of science and technology, the hotness of target recognition detection has increased year by year in recent years, and the subject has its wide application value in various aspects such as medical, daily life and military. This design on the subject, based on the popular target detection algorithm in recent years - Yolo algorithm, based on the latest Yolov5 algorithm model, the improvement and optimization of the algorithm, combined with the hot topics of current affairs, real-time monitoring of whether the target is wearing a mask, using the dataset provided by Aliyun Tianchi, to conduct experiments. The experimental results are: with Yolov5s base model as the prototype, after optimization and improvement, the fusion rate of target features is improved on the basis of maintaining the original training speed, which effectively improves the accuracy and precision of target recognition.

Keywords: Target Recognition Detection, Deep Learning, Yolov5.

1. Introduction

With the rapid development of science and technology in the 21st actuality, deep learning[1], as one of the popular research directions nowadays, has increased its performance level at a high speed and has been widely used in many different research directions with its excellent ability. Deep Learning belongs to the field of machine learning[2]. The ultimate goal of deep learning is to enable machines to have independent analytical learning capabilities and to be able to think independently like humans.

With its excellent recognition accuracy and excellent computing speed, Yolo algorithm[3] has been loved and respected by many researchers, and research scholars all over the world are devoted to enhance its operation speed or improve its detection performance. In this paper, we will improve the latest version of the algorithm Network model, combined with today's real-time social environment, to detect whether a pedestrian is wearing a mask as the detection target, by comparing the detection rate and accuracy of the algorithm before and after optimization, comparing the differences before and after optimization, so that it can achieve better results in the detection of small and dense targets.

2. Literature review

Target recognition[4] detection is a research topic designed for many fields such as computers and communications. The classical target detection algorithm is implemented in three steps[5]: region suggestion, acquisition of feature information and classification regression. Region suggestion is a multi-scale extraction of feature information from images by sliding windows of different sizes to find regions where a given object may be located. Feature extraction[6] is the conversion of images in candidate regions into feature vectors using manual feature extraction, commonly used methods such as local binary pattern features, gradient histogram features, etc. Classification regression is to predict the class of the target in the candidate region using a pre-trained classifier. The disadvantages are: the traditional target detection algorithm has a large number of redundant calculations in region suggestion, only low-level features can be extracted in feature extraction, and the whole process is divided into three stages, and the algorithm cannot find the global optimal solution.

This design is to study the target recognition detection algorithm with deep learning as the kernel, taking the now rapidly developing algorithm as the research direction, and trying to achieve the goal of improving the performance of the algorithm by modifying the basic structure of the four major components in the network model.

3. Methodology

3.1. Modification ideas of data augmentation methods

In the network model, one of the methods for information augmentation[7] is Mosaic information augmentation that after the scale transformation and color gamut change of four random images in the dataset, the four images are stitched together according to the given anchor points to form a new image for the network model to train, which adds small targets to the network model while enriching the dataset and optimizes the recognition ability of the network. Guided by this idea, this paper decides to improve the method by adopting Mosaic-Nine, which means that nine pictures are stitched together to form a new data picture after random cropping, random scaling, and color gamut transformation, and at the same time, some random noise is introduced into the new data picture, compared with the original Mosaic data enhancement method, Mosaic -Nine data enhancement method further improves the generalization power of the model and adds more small target samples to improve the recognition ability of the network model for small targets.

3.2. Modified idea of feature information extraction

The performance of the backbone network is very good and can do our target work well in most cases, but at some times it cannot meet our needs, for example, when processing small target data, some feature information of small targets is easily lost after convolution sampling. In order to improve the shortcomings of the backbone network, we add Coordinate Attention (CA) mechanism to the backbone network, which can effectively improve the ability to extract feature information of small target data and dense target data, so as to improve the accuracy and precision of network model detection.

As an attention mechanism module, CA attention mechanism module is highly integrated and plug-and-play compared with the pre-existing modules, and it has been widely used in the modification and optimization work of different neural network models. The module is added to the Backbone part of Yolov5s to improve the neural network model's ability to recognize small targets and targets in complex environments.

3.3. Modified idea of feature information fusion

PANet[8] is the network structure now used in Yolov5s Neck layer, i.e., FPN combined with PAN. Considering the shortcomings of the previous network structure, we further modify and optimize the new network structure, which is named BiFPN[9]. Compared with PANet, firstly, we remove the nodes that have and only have one input edge. The reason behind this is simple: 1. If the role of a node is only to receive input information without feature fusion, the Neck layer as a module to achieve feature information fusion in the overall network model, the node will play a minimal role in the realization of the effect of the Neck layer; 2. After removing some nodes, we add a new channel between the original input node and output node in the same layer at the location, sacrificing a smaller training cost to improve the performance of feature information fusion; meanwhile, BiFPN uses Weighted Feature Fusion (WFF) method, which can somewhat distinguish different feature information and learn the importance of different feature information, which has some similarity with attention mechanism. The structures of

FPN, PANet, NAS-FPN and BiFPN are shown in the figure below.

3.4. Loss function modification idea

The loss function part of the network model is composed of several modules: localization loss, confidence loss and category loss[10]. We choose to use the CIoU loss function formula to replace GIoU as the loss function for the target box regression, and the formula is as follows.

$$\text{CIOU}=\text{IOU}-\frac{\rho^2\left(b,b^{gt}\right)}{c^2}-\alpha v$$

$$\alpha=\frac{v}{(1-\text{IOU})+v}$$

$$v=\frac{4}{\Pi^2}(\arctan\frac{w^{gt}}{h^{gt}}-\arctan\frac{w}{h})^2$$

In the above equation, α represents the equilibrium parameter, which is not involved in the gradient calculation process, and what really plays a role is v. The role of v is to measure the consistency of the aspect ratio, and αv can be regarded as a penalty factor, by which the loss function successfully considers the degree of fit of the predicted frame to the real frame size. Compared with GIoU, CIOU is able to better describe the overlap information during regression and optimize the convergence results of the target box, while obtaining better regression results.

4. Result

4.1. Comparison of training results

Comparison experiments are one of the common scientific tests in experiments that can better show the difference in performance before and after model modification. We conducted experiments with the same training data set and training parameters for different models. The experimental results are compared as follows.

Table 1. Model performance comparison test before and after optimization.

Algorithm	precision	mAP@0.5	mAP@0.5：0.95	Recall
Yolov5-basic	0.84	0.85	0.53	0.77
Yolov5-MosaicNine	0.87	0.86	0.60	0.82
Yolov5-BiFPN	0.85	0.87	0.54	0.77
Yolov5-CA	0.90	0.91	0.62	0.82
Yolov5-CIOU_loss	0.85	0.87	0.57	0.81
Yolov5-New	0.95	0.95	0.72	0.87

In Table 1, Precision stands for precision, mAP@0.5 represents the AP(average precision) at a threshold of 0.5, mAP@0.5: 0.95 represents the average value of each mAP in the threshold range from 0.5 to 0.95 in steps of 0.05, and Recall stands for recall. As can be seen from Table 3, compared with the initial Yolov5-basic model, the Yolov5-CA model with the addition of the CA attention mechanism has the best performance in terms of indicators. The remaining several modifications, each indicator also has different degrees of optimization improvement, while the Yolov5-BiFPN algorithm, although the improvement of relevant parameters in the table is small, the algorithm runs the fastest in each epoch during the training process, with a 1/5 improvement in computing speed compared to the initial model.

4.2. Ablation experiment

Based on the afore mentioned network model optimization approach, we add them into the base network model separately to observe the new training effect. As Table 2.

Table 2. Results of ablation experiments.

Model	Modify to Mosaic-Nine	Add CA Attention Module	Modified to BiFPN	Modify the loss function	mAP
Yolov5s	×	×	×	×	0.852
Model 1	√	×	×	×	0.867
Model 2	×	√	×	×	0.902
Model 3	×	×	√	×	0.871
Model 4	×	×	×	√	0.875
Model 5	√	√	×	×	0.914
Model 6	√	×	√	×	0.910
Model 7	√	×	×	√	0.922
Model 8	×	√	√	×	0.913
Model 9	×	√	×	√	0.921
Model 10	×	×	√	√	0.917
Model 11	√	√	√	×	0.931
Model 12	√	√	×	√	0.935
Model 13	√	×	√	√	0.947
Model 14	×	√	√	√	0.942
Model 15	√	√	√	√	0.959

In Table 2, the changes of mAP (mean accuracy) are observed with the four components modified or added as variables, respectively. As can be seen from the table, when adding or modifying the individual modules, the network model performance is most significantly improved after adding the CA attention module. mAP is improved by 5.0% compared to the Yolov5s base model when the attention module is added alone. When all modified methods are used simultaneously, mAP gets a 10.7% improvement when all optimization ideas are added into the base model at the same time, which shows that the performance of the modified Yolov5s network model for target recognition detection is better improved.

4.3. Comparison of picture recognition effects

The following is the effect of different algorithms on the same image recognition, we put the Yolov5 original model and Yolov5-New model on the same target image recognition results together to make a comparative observation and analyze the results.

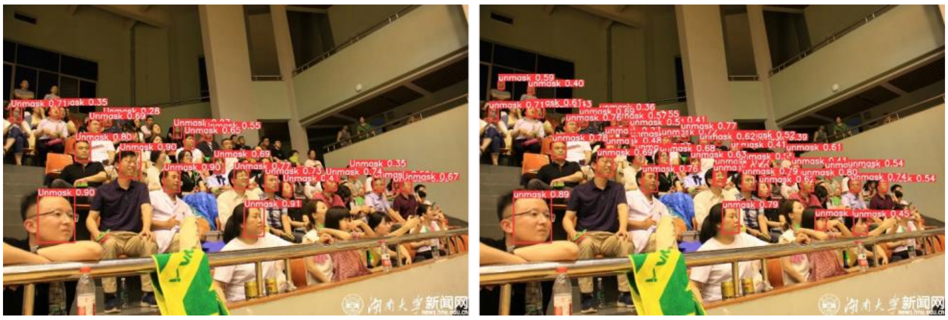

Figure 1. Comparison of the recognition effect of small target objects in bright environment.

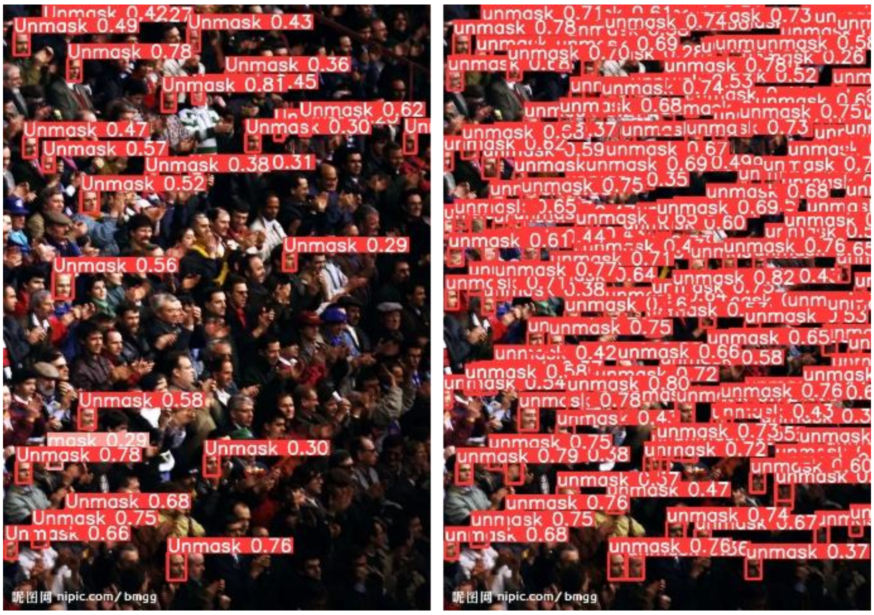

Figure 2. Comparison of the recognition effect of small targets in dark environment.

Figure 1 and Figure 2 show the detection effect of small targets. We can observe that the detection performance of the original Yolov5s model for small targets is acceptable, and it can already detect some small target samples, but the accuracy rate varies, with some targets having an accuracy rate of less than 50% and only some targets having an accuracy rate of more than 70%, and there is also a false detection object. After optimizing the network model structure, the recognition accuracy in the Yolov5-New model is greatly improved. We can observe that the number of detected objects increases very significantly, and most of the objects in the figure are successfully recognized, and the recognition accuracy also obtains very obvious enhancement, and the accuracy rate of almost all targets exceeds 60%, thus it can be seen that the optimization ideas in this paper have helped to improve the recognition performance.

5. Discussion

Based on the topic of target recognition detection with deep learning as the kernel, the following research work is conducted in this paper: we introduced the basic information about the model, and introduced the network structure as well as the advantages and disadvantages of the model based on the Yolov5s model; in terms of data augmentation, Mosaic-Four is modified to Mosaic-Nine, which increases the number of small target samples, improves the robustness of the network model; in the feature information extraction part, the CA channel attention mechanism is added, so that the neural network can obtain the location information of target features while acquiring the target feature information; in the feature information fusion part, we remove the nodes in the FPN structure that play a small role in the overall model and add jump links, which reduces the amount of operations in the overall model and enables more feature information to flow to the same node, making the feature information fusion more complete; in the loss function, in order to better clarify the relative position information of the prediction frame and the real frame when they overlap, we replace the GIoU loss function with the CIoU loss function as the new loss function formula.

6. Conclusion

The optimization ideas proposed in this paper involve data augmentation, attention mechanism, network model complexity optimization, and regressivity loss, etc., and a certain degree of performance improvement is obtained based on Yolov5s. However, the model still has the phenomenon of missed detection and false detection when facing the environment with high complexity; although the detection ability of small targets has been optimized to a certain extent, there are still many unsuccessfully recognized targets; in video detection, only an average 35FPS effect can be achieved, and if we want to obtain higher frame rate and smoother video recognition effect, we need to improve the equipment hardware level, which requires higher equipment performance High.

References

[1] Chen Mu Yen,Lughofer Edwin David,Egrioglu Erol. Deep learning and intelligent system towards smart manufacturing[J]. Enterprise Information Systems,2022,16(2).

[2] Yibin He, Shengzhe Tian, Guilong Lan. Pedestrian Detection Based on Improved Faster-RCNN Algorithm [J]. *Automobile Applied Technology*, 2022, 47(05): 34-37.

[3] S Poonkuntran,Rajesh Kumar Dhanraj,Balamurugan Balusamy. Object Detection with Deep Learning Models:Principles and Applications[M].CRC Press:2022-04-07.

[4] Reja Varun Kumar,Varghese Koshy,Ha Quang Phuc. Computer vision-based construction progress monitoring[J]. Automation in Construction,2022,138.

[5] Xiaolan Wang,Shuo Wang,Jiaqi Cao,Yansong Wang. Data-driven based tiny-YOLOv3 method for front vehicle detection inducing SPP-net[J]. IEEE Access,2020,PP(99).

[6] Li Yunquan,Gao Meizhen,Li Qiangyi. Face Recognition Algorithm Based on Multiscale Feature Fusion Network[J]. Computational Intelligence and Neuroscience,2022,2022.

[7] Shitong Cao, Yulian Jiang. Combining Attention Mechanisms with RPN Networks for Target Tracking[J]. International Core Journal of Engineering,2021,7(10).

[8] Yue Sha,Wang Shengzhe,Cui Yuyong,Guan Wei,Gao Xinyi. Heterogeneous Image Template Matching Based on Region Proposal Network[J]. Journal of Physics: Conference Series,2021,1848(1).

[9] Xiaonan Liu, Debin Wu, Zhenyu Liu, Xue Wei. Small Object Detection Algorithm with Top-down Feature Fusion [J/OL]. *Telecommunication Engineering* ： 1-7[2022-11-05]. http://kns.cnki.net/kcms/detail/51.1267.TN.20220914.1909.006.html.

[10] Zhengyou Liang, Jingbang Geng, Yu Sun. Trafiic Sign Recognition Algorithm Based on Improved Residual Network. *Computer and Modernization*. 2022(04):52-57+64.