2022 IEEE 4th International Conference on Civil Aviation Safety and Information Technology (ICCASIT) | 978-1-6654-6766-7/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICCASIT55263.2022.9986690

Detection Algorithm Based on Improved YOLOv4 for Pedestrian

Wen Kong[#]
Henan Normal University
Xinxiang, Henan, China
kw888_up@126.com

Yidan Qiao[#]*
Taiyuan University of Technology
Taiyuan, China
qyd1596676108@outlook.com

Ziqi Wei[#]
Henan Normal University
Xinxiang, Henan, China
wzq13693739860@163.com
#These authors contributed equally

***Abstract*—In this paper, the application of the target pedestrian detection algorithm is studied based on the YOLOv4 network. The simulative results show that the improved YOLOv4 algorithm, which integrates the CBAM attention mechanism and Focal Loss function, shows high detection accuracy in target pedestrian detection. By CBAM, The new feature map will get the attention weight of the channel and space dimension, practical features of the target. Besides, focal loss enhanced the training of difficult-to-classify samples. Based on the improved YOLOv4 algorithm, which will have great potential, the problems of low detection accuracy and seriously missed pedestrian detection in realistic, complex visual scenes are solved.**

Keywords—target pedestrian detection, YOLOv4 network, the CBAM attention mechanism, Focal Loss function

I. INTRODUCTION

With the economy's rapid development and the increase in science and technology, pedestrian detection [1]in intelligent video surveillance applications is very extensive. Especially in crowded places, pedestrian detection technology[2], abortion statistics, people tracking, traffic guidance, and security early warning work[3] have become one of the research hotspots in the field of computer vision, and the difficulty in the field of target detection.

As a relatively new model in the YOLO series of algorithms[4], YOLOv4 combines a large number of previous research technologies and integrates appropriately to achieve a perfect balance between speed and accuracy. The influence of the most advanced target detection training optimization algorithm Bag-of-Freebies and Bag-of-Specials target detection methods in detector training was verified by YOLOv4[5]. The most advanced method was modified to make it more efficient and suitable for single GPU training, including CBN, PAN, SAM, etc.. Research on the optimization of the YOLOv4 algorithm is also emerging. YOLOV4 Tiny target detection algorithm is a simplified version of the YOLOV4 target detection algorithm. The convolutional neural network extracts the feature, and the classical deep learning algorithm predicts the category and boundary frame coordinates. Pro-YOLOv4 is a target detection algorithm integrating convolution attention mechanism and lightweight network, which has good application in multi-scale aerial image target detection. In the feature enhancement part of the YOLO4 algorithm, the attention module D-CBAM combining residual connection and cavity convolution is introduced, and the important information in the extracted features can be obtained.

In the research of attention mechanism, attention mechanism has made an important breakthrough in the image, natural language processing, and other fields in recent years, which has been proved to be beneficial in improving the performance of the model. The result is usually displayed as a probability map or feature vector, in which CBAM is the representative network. The current research shows that the convolution block attention mechanism module (CBAM) is added at the end of CSPDarkNet to improve the effect of network feature extraction. The algorithm is improved to effectively reduce the missing detection phenomenon of YOLOX in detecting close ship targets and small-size ship targets.

The focal loss function is proposed to solve the problem of feature information loss and class imbalance in target detection. By reducing the weight of easily classified samples, the model focuses more on difficultly classified samples in training. The research results show that Focal loss can alleviate the imbalance of the proportion of positive and negative samples in the training process of YOLOX and can also improve the detection accuracy.

The detection target of this paper is pedestrians. Aiming at the problems of low detection accuracy and seriously missed detection of pedestrians in complex visual scenes, in reality, the original YOLOv4 algorithm is improved by introducing the CBAM attention mechanism and Focal Loss. Finally, the algorithms of different strategies are compared experimentally. The research contents of this paper are as follows: YOLOv4 Network was introduced in Chapter 2, YOLOv4 Network was discussed in Chapter 3, which is improved based on CBAM attention mechanism and Focal Loss, the experimental comparison and result from the analysis were analyzed in Chapter 4.

II. YOLOV4 NETWORK

Many optimization methods based on the previous YOLOv4 network were added through YOLOv4, including optimizing the backbone network, activation function, loss function, network training, data processing, etc. The YOLOv4 network consists of three parts: Backbone, Neck, and Head. As shown in figure 1.

The backbone network adopts CSPDarknet53. The convolutional networks and ensure accuracy while realizing network lightweight. Compared with traditional ReLU with a negative complex zero boundary, the Mish activation function is used to give the network better generalization and accuracy. The neck network adopts path aggregation network PANet to enhance the feature extraction capability of the network and introduces Space Pyramid Pool (SPP) module to improve

978-1-6654-6766-7/22/$31.00 ©2022 IEEE

network prediction accuracy effectively. The head adopts a multi-scale prediction method similar to YOLOv3 to detect small, medium and large targets. In terms of the loss function, the mean square error (MSE) of YOLOv3 is no longer used as a regression box prediction error, but CIOU is used, as shown in Formula (1) - (3).

$$L_{CIOU} = 1 - IOU_{(a,\ b)} + \frac{\rho^2(a_{ctr}, b_{ctr})}{d^2} + av \tag{1}$$

$$\alpha = \frac{V}{(1 - IOU_{(a,b)}) + V} \tag{2}$$

$$V = \frac{4}{\pi^2}\left(\arctan\frac{w_{gt}}{h_{gt}} - \arctan\frac{w}{h}\right)^2 \tag{3}$$

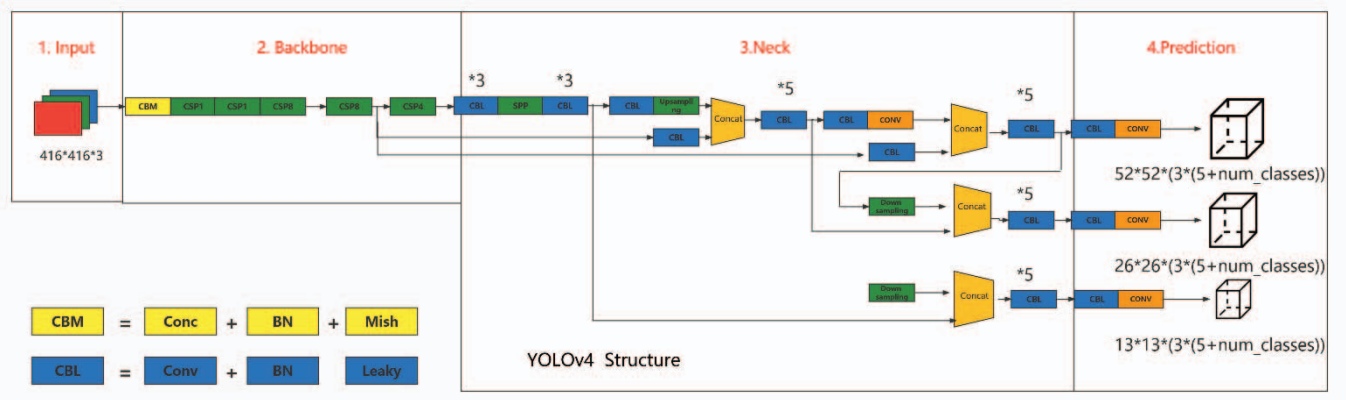

Fig. 1. YOLOv4 Network Structure

Where $IOU_{(a,b)}$ is the intersection ratio of the actual frame and the prediction frame, $p^2_{(Actr,\ bctr)}$ is the Euclidian distance between the center point of the real frame and the prediction frame, and d is distance.

As for the unbalanced number of small, medium, and large objects in the data set, YOLOv4 introduces the Mosaci data enhancement, which uses four images randomly for scaling and stitching, greatly enriching the data set and enhancing the effect of the Batch Normalization layer to improve the performance of small object detection.

III. Improved YOLOv4 Network

Compared with YOLOv4 and YOLOv5, YOLOv3 has the problems of insufficient recall rate, inaccurate positioning, and the failure of the model to meet the real-time requirements in practical application scenarios. YOLOv4 proposes many optimization templates and makes excellent improvements to these shortcomings. YOLOv5 has a certain degree of optimization compared with YOLOv4, but when the detection accuracy is high, the detection speed can not keep up, and the higher detection speed will be at the cost of reducing the detection accuracy. In summary, YOLOv4 not only has many optimizations based on YOLOv3 but also can ensure the detection accuracy and detection speed simultaneously, because YOLOv4 network is selected as the optimization object in this paper.

A. CBAM attentional mechanisms

Aiming at the problem that the feature fusion network of the YOLOv4 algorithm is located behind the trunk network, resulting in the extracted feature redundancy and making the recognition effect of the model worse, this paper adds CBAM attention mechanism in YOLOv4 to make the network pay attention to the more important features and ignore the redundant features in the training process, so as to improve the detection accuracy.

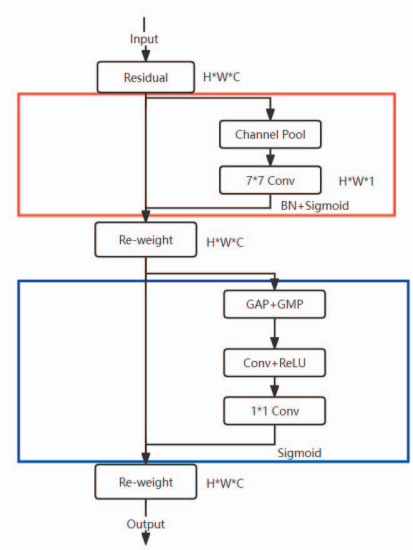

Fig. 2. CBAM Schematic diagram of attentional mechanism module

The figure 2 above is the CBAM attention mechanism module. CBAM attention mechanism is divided into two parts: spatial attention and channel attention. As seen from the figure above, the part in the red box is channel attention, and the part in the blue box is spatial attention. Channel attention is before, and attention in the space is before. After inputting the characteristics of the figure, channels of attention in the space were accessed. GAP and GMP were conducted. Then through the Sigmoid function, the normalized attention weights were obtained, at last, by multiplication by weight to the original input channel characteristic diagram, the re-calibration of channel attention to the original features is completed, as shown in formula 4:

$$\begin{aligned} M_c(F) &= \sigma(MLP(AvgPool(F)) + MLP(MaxPool(F))) \\ &= \sigma(W_1(W_0(F^c_{avg})) + W_1(W_0(F^c_{\max}))) \end{aligned} \tag{4}$$

To obtain the attention features in the spatial dimension, the feature images of attention outputted through channels are also globally maximized and averaged based on the width and height of the feature images, and the feature dimension is transformed from H×W to 1×1. Then, the dimension of the feature images is reduced after the convolution kernel of 7×7 and the Relu activation function. Then, it is upgraded to the original dimension after a convolution. Finally, the feature graph normalized by the Sigmoid activation function is merged with the feature graph of channel attention output to complete the recalculation of the feature graph in both spatial and channel dimensions, as shown in the following formula:

$$\begin{aligned} M_s(F) &= \sigma\left(f^{7*7}([AvgPool(F); MaxPool(F)])\right) \\ &= \sigma\left(f^{7*7}([F^s_{avg}; F^s_{\max}])\right) \end{aligned} \tag{5}$$

In the spatial attention module, global average pooling and maximum pooling obtain spatial attention features and establish the correlation between spatial features through two convolutions while keeping the input and output dimensions unchanged. The convolution operation with a convolution kernel of 7×7 significantly reduces the parameters and computation, which is conducive to establishing high-dimensional spatial feature correlation. After CBAM, the new feature map will get the attention weight of channel and space dimension.

B. Focal loss

In YOLOv4, a large number of prior frames Anchor Boxes will be generated first for locating objects. However, in actual pedestrian detection, in most cases, there are only a small number of targets in an image so many Anchor boxes will be generated in the background area. YOLO algorithm will directly classify these Anchor boxes with uneven positive and negative samples and then use cross-entropy to calculate classification loss and confidence loss, as shown in Formula (6) and (7), where P represents the probability of predicting categories. This process treats all categories indiscriminately and ignores the imbalance of positive and negative samples. To solve this problem, the balanced cross-entropy loss adds a weight factor A1 before each category to coordinate the category imbalance, as shown in Equation (8).

$$CE = -\log\left(p_t\right) \tag{6}$$

$$p_t = \begin{cases} 1-p, & p < 0 \\ p, & p \gg 0 \end{cases} \tag{7}$$

$$Balance\_CE = -a_t \log\left(p_t\right) \tag{8}$$

In addition to positive and negative samples, there are also samples easy to classify samples and difficult to classify. In order to improve the detection ability of the network, the difficult-to-classify samples should be considered in training. Meanwhile, only one weight factor is added to balance the positive and negative samples in the loss of equilibrium cross-entropy, and the distinction of difficult-to-easily samples is not considered. To solve this problem, focal loss function (FL) added a regulating factor (1-p^t)y based on balancing cross-entropy loss to reduce the weight of easily classified samples and focus on training complex classified samples. FL is expressed as Formula (9), where y is the focusing parameter, and the reduction degree of weight (1-pt)y can be adjusted. The larger y is, the greater the reduction degree of weight will be. When p^t is very small, it is indicated that the sample is difficult to classify. At this time, the regulatory factor (1-p^t)y approaches 1, and the weight of the sample in the loss function is not affected. When p^t is very large, the sample is easy to classify. At this time, the regulatory factor approaches 0, and the weight of the sample in the loss function decreases a lot to enhance the training of difficult-to-classify samples.

$$FL_{LOSS} = -a_t \left(1-p_t\right)^y \log\left(p_t\right) \tag{9}$$

IV. Experimental comparison of different improvement strategies

A. Data set Description

This experiment was uniformly trained and verified in the INRIA data set. A pedestrian detection dataset, whose image library is divided into four categories: car only, people only, people with cars, and no cars, was trained and tested on the INRIA Pedestrian dataset. INRIA is the most used static pedestrian detection data set at present. The pedestrian posture and lighting conditions in the image are rich and changeable, and there is a single pedestrian and a packed crowd, so it is suitable for pedestrian detection. There were 614 images in the training set and 288 images in the test set. In order to avoid over-fitting in the training process, the training set was expanded to 3070 images by randomly adding noise, adjusting brightness, rotation, clipping, translation, and cutout.

B. Network training and evaluation indicators

The experimental environment is Windows operating system, NVIDIA GeForce GTX 1050 graphics card. The model input image size was set to 416*416*3 during training for software environments of cudA10.0, CUDN10.0, Tensorflow-gpu1.13.1, Keras2.1.5, and PYTHon3.6. In order to avoid overfitting, cosine annealing attenuation of the learning rate is used to realize the change in learning rate during training. When the learning rate linearly increases to the maximum value, it remains unchanged for some time, and then the downward trend of the simulated cosine function is attenuated. The decay equation of cosine annealing is shown in Equation (10) :

$$\lambda_t = \lambda_{\min} + \frac{1}{2}\left(\lambda_{\max} - \lambda_{\min}\right)\left(1 + \cos\left(\frac{T_{cur}}{T_i}\pi\right)\right) \tag{10}$$

Where λ_{max} and λ_{min} represent the maximum and minimum values of the learning rate. T_{cur} represents how many epochs are currently executed, and T_i represents the total number of epochs trained by the model.

In this experiment, two performance indexes, mAP(mean average precision, equal to AP value) and FPS(Frames Per Second, the number of Frames transmitted Per Second), were used to evaluate network performance. MAP value refers to a p-R (Precise) two-dimensional curve drawn according to precision P (Precise) and Recall rate R (Recall). The larger the area enclosed with the X-axis, the higher the detection accuracy of the model. FPS represents the number of video frames the model can process per second, and the larger the value, the faster the model detection speed. P, R, and mAP are calculated in the following formula. Explanation of variables

in the formula is shown in table I.

$$P = \frac{T_P}{T_P + F_P} \times 100\% \tag{11}$$

$$R = \frac{T_P}{T_P + F_N} \times 100\% \tag{12}$$

$$mAP = \int_0^1 P(R)dR \tag{13}$$

TABLE I EXPLANATION OF VARIABLES IN THE FORMULA

Variable	Meaning
TP	The number of targets correctly detected by the model
FP	The number of targets detected by system error
FN	The number of system error detection and missed detection
AP	The area under the PR curve

C. Experimental comparison of different improvement strategies

In order to verify the influence of the improved strategy proposed in this paper on network detection performance, a series of comparative experiments were carried out. Based on the original YOLOv4 network, improvements were made, respectively, and the Focal Loss function and CBAM module were added to the original YOLOv4 network for experimental comparison.

TABLE II COMPARISON OF EXPERIMENTAL RESULTS OF DIFFERENT IMPROVEMENT STRATEGIES

algorithm	mAP(%)	FPS(frames per)
Original YOLOv4	94.09	33.39
YOLOv4+Focalloss	94.12	33.48
YOLOv4+CBAM	96.48	33.15
Algorithm proposed	97.14	33.54

Fig 3. Visual renderings of YOLOv4 and the proposed algorithm on INRIA

By analyzing the experimental data in Table II, it can be found that the mAP value of detection accuracy can be improved by 3.24% after the proposed algorithm replaces the original YOLOv4 algorithm. The detection speed can reach up to 33.54FPS, which is better than the module alone. In conclusion, the proposed method greatly optimizes the detection accuracy of the YOLOv4 network in pedestrian detection. In addition, the results of the proposed method and YOLOv4 algorithm on the INRIA validation set are visualized, as shown in Figure 3. Through the visual effect diagram, we can see the remarkable effect achieved by the method in this paper.

D. Test results of different algorithms

Different YOLO networks and improved YOLOv4 networks were tested on INRIA data sets, and multi-category average detection accuracy mAP and FPS values were obtained, as shown in Table III.

TABELIII MAINSTREAM ALGORITHM DETECTION COMPARISON RESULTS

Training dataset	Testing dataset	Detecting algorithm	mAP(%)	FPS(frames per second)
The INRIA training set	The INRIA test set	YOLOv3	92.17	24.30
		YOLOv4	94.09	33.39
		YOLOv5	96.26	71.43
		Ours	97.14	33.54

Table III compares the mAP and FPS results of different YOLO network structures and algorithms in this paper. Compared with YOLOv3, both YOLOv4 and YOLOv5 are improved to a degree. YOLOv5 has a significant improvement effect, with a 4.44% increase in mAP and a 47.13FPS increase in speed. Based on YOLOv5 optimization, the mAP value of the proposed algorithm is improved by 0.91% again, reaching 97.14%, but the speed is not fast enough. In summary, the method in this paper greatly improves the detection accuracy of pedestrian detection.

V. CONCLUSION

This paper studies the application of deep learning for target detection. The YOLOv4 algorithm integrated CBAM attention mechanism and Focal Loss function is applied to target pedestrian detection areas. Then, the CBAM attention mechanism is utilized to improve the connection of each feature in channel and space and extract the practical feature of the target. Afterward, the Focal loss function is used to reduce the weight of easily classified samples and focus on training complex classified samples. Numerical results indicate that the method of deep learning by improved YOLOv4 algorithm can be applied to target detection. The accuracy based on the CBAM attention mechanism and Focal loss function can verify the feasibility of handling problems of the bad detection accuracy when pedestrian detection is severe in actual complex visual scenes.

REFERENCES

[1] Dollár P, Wojek C, Schiele B, et al. Pedestrian detection: A benchmark[C]//2009 IEEE conference on computer vision and pattern recognition. IEEE, 2009: 304-311.

[2] Yu J, Zhang W. Face mask wearing detection algorithm based on improved YOLO-v4[J]. Sensors, 2021, 21(9): 3263.

[3] Tian Y, Mao W, Yuan S, et al. A Decision Support System for Power Components Based on Improved YOLOv4-Tiny[J]. Scientific Programming, 2021, 2021.

[4] Redmon J, Farhadi A. Yolov3: An incremental improvement [EB/OL]. (2018-08-08)[2022-06-10]. https://arxiv.org/abs/1804.02767.

[5] Wen H K, Dai F Z, Yuan Y S. A Study of YOLO Algorithm for Target Detection[C]//26th International Conference on Artificial Life and Robotics, Jan 21-24, 2021, ELECTR Network. Japan: OITA, 2021: 622-625.