SMALL OBJECT DETECTION ALGORITHM BASED ON YOLOV5 AND ATTENTION MODEL

Yuandong Yao[1], Sergey Ablameyko[1,2]

[1]*Belarusian State University, 4, Nezavisimosti av., Minsk, 220050, Belarus,* farawayeast@qq.com

[2]*United Institute of Informatics Problems of the National Academy of Sciences of Belarus, 6, Surganova str., Minsk, 220012, Belarus,* ablameyko@bsu.by

In order to solve the problems of small object and dense object in complex environment in object detection, such as low amount of object feature information, difficult positioning, false detection and missed detection, this paper proposes a YOLOv5 detection method with optimizes clustering and introduces CBAM attention mechanism. It improves the object feature extraction ability of the algorithm backbone network and captures small object features more accurately. The self-built helmet dataset is used for training and comparison experiments. The experimental results show that the algorithm has improved accuracy and speed, and has strong real-time performance.

Keyword: small object detection; clustering algorithm; attention mechanism.

Introduction

Small object detection is a difficult problem for a long time, which aims to accurately detect small objects with few visible features in images. In real scenarios, due to the large number of small objects, small object detection has broad application prospects and plays an important role in many fields such as autonomous driving, smart medical care, defect detection, and aerial image analysis. Relative to regular-sized objects, small objects usually lack sufficient appearance information, making it difficult to distinguish them from background or similar objects.

Convolution neural networks are widely used for small object detection. CNN-based detectors mostly adopt two architectures: one-stage architecture represented by SSD and YOLO series and two-stage architecture represented by the faster RCNN series and its improved version. In the object-detection tasks with a high requirement for reasoning performance but a low requirement for detection performance, the one-stage detector architecture is preferred [1].

On the object detection public dataset MS COCO [2], there is a significant gap between the detection performance of small objects and large objects. It can be seen that small object detection is still full of challenges, such as few available features,

high positioning accuracy requirements, small proportion of small objects in existing dataset, and small object aggregation.

This paper is intended to extend possibilities of one-stage architecture and we selected YOLOv5 [3] as the basic CNN. We improved the YOLOv5 detection algorithm to achieve better results in small object and dense object detection. Compared with traditional detection methods, the algorithm has improved accuracy and speed, and has strong real-time performance, but there are still shortcomings in the detection of small objects and dense objects.

1. Dataset

This paper collected a total of 12,314 images by autonomous collection. It includes large-scale objects (object pixel area exceeds 96×96), medium-scale objects (object pixel area is between 32×32 to 96×96), small-scale objects (object pixel area is less than 32×32) and dense helmets image of object (including objects at multiple scales). There are about 6800 images with helmet and about 5200 images without helmet, and the number of objects exceeds 80,000.

2. YOLOv5 and attention model

YOLOv5 algorithm uses Mosaic data augmentation at the input to optimize the dataset. Four original pictures in dataset are read each time, pictures are randomly scaled, flipped, changed color gamut and other operations, and then the changed pictures are joined together in four directions respectively. The backbone network of YOLOv5 model adopts Focus+CSP structure. Before picture enters the Backbone, it needs to go through the Focus structure to slice the picture. The slicing operation is shown in Fig.1, An image of size 4×4×3 is sliced into a 2×2×12 feature map.

In the YOLOv5s model, the size of the original graph was 608× 608×3, which became 304×304×12 after the slicing operation. After through a convolution with a 32 convolution kernel, the size of the feature graph became 304×304×32. After adding the SPP module to the CSP module, the multi-scale maximum pooling layer greatly improves the receptive field, reduces the possibility of information loss caused by directly scaling the image, and improves the model accuracy. The network structure of YOLOv5 draws on the design idea of CSPNet [4], add the CSP structure to the network. The backbone network uses CSP1_X structure, and Neck uses CSP2_X structure.

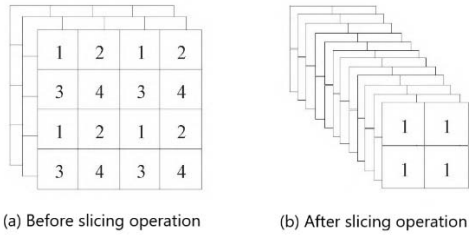


Fig.1. – Slicing operation

The Attention Model [5], first introduced in 2015 for machine translation, has become a predominant topic in the literature on neural networks. Attention models have become extremely popular in the artificial intelligence community as an important component of neural architectures for a large number of computer vision applications [6].

Attention models are methods of processing input data of neural networks that allow the network to focus one at a time on each of the parts of complex inputs until the entire dataset is categorized. The goal is to break down complex tasks into smaller areas of attention, which are processed sequentially, just as the human mind solves a new problem by dividing it into simpler tasks and solving them one by one. The mechanism of the attention model is trained during network training and should help the network focus on the key elements of the image. Architecture of Convolution Block Attention Module is shown in Fig. 2.

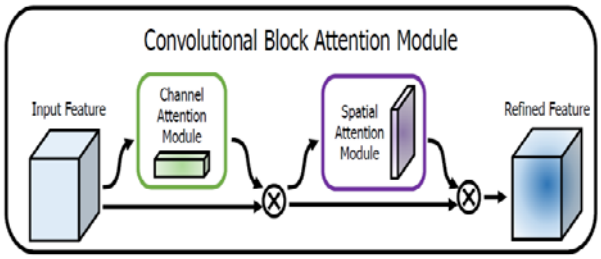


Fig.2. – CBAM structure

3. The proposed algorithm

Based on the YOLOv5 we propose the following algorithm to detect small objects. The object box is optimized by the improved DPC density peak clustering algorithm and introduce CBAM attention mechanism.

Rodrigunz et al. [7] proposed a clustering algorithm DPC based on fast search and find of density peaks, which can effectively solve the problem of K-means clustering algorithm. Xue Xiaona et al. [8] proposed the IDPCA clustering algorithm in view of the shortcomings of the DPC algorithm. When the density

difference of the dataset is large, the local density ρ and the minimum distance δ between this point and other points with higher density in the DPC algorithm cannot cope with complex density differences. DPC algorithm has a domino effect on the allocation strategy of the remaining points, which will lead to error propagation. In this experiment, IDPCA density peak clustering algorithm is used to cluster object boxes. IDPCA clustering selects cluster centers according to the following steps:

(1) Calculate the local density ρ_i of each point x_i and the minimum distance δ_i between this point and other points with higher density, the formula is

$$\rho_i = \sum_j e^{-\frac{1}{r}(\frac{d_{ij}}{\sigma})^2} \tag{1}$$

$$\delta_i = \min_{j:\rho_j>\rho_i} d_{ij} \tag{2}$$

in equation (1) (2): σ is 2% of the amount of data; d_{ij} is the distance between point x_i and point x_j; r is the similarity coefficient, if the value is larger, points that are closer to point x_i have more weight ρ_i.

(2) Obtain m cluster centers through the γ decision diagram, where γ = ρδ;

(3) Extract the core points, and use the global search allocation strategy to classify the classification points. The way to extract the core points is to calculate the average distance u_m between all points in the local class C_m and the class center cen_m, The calculation formula of u_m is shown in equation (3). If $x_i \in C_m$ is in the εu_m neighborhood of cen_m, then x_i is the core point, separation threshold ε = N‰, N is the number of datasets;

$$u_m = \frac{1}{|C_m|} \sum_{i=1}^{|C_m|} d_{im}^{cen_m} \tag{3}$$

(4) Using statistical learning strategy to assign the remaining points;

(5) The last unprocessed point can be regarded as a noise point, which is classified into to the class of its nearest neighbor.

The IDPCA clustering algorithm analyzes the bounding box of the marked helmet, and generates multiple a priori box sizes of different sizes, so that the matching degree between the a priori box and the actual box is higher, thereby improving the accuracy of object detection. The size of the prior frame based on IDPCA clustering algorithm used in experiment is shown in Table 1, the size of the

prior frame based on the K-means clustering algorithm is shown in Table 2.

Table 1 Prior box size based on IDPCA clustering algorithm

object type	priori box size
small object	55,23,33,46,80,36
medium object	62,70,147,59,137,91
big object	92,170,250,103,167,165

Table 2 Prior box size based on K-means clustering algorithm

object type	priori box size
small object	54,23,32,42,78,35
medium object	58,69,152,60,135,90
big object	80,202,158,166,175,100

The architecture of the channel attention module is shown in Fig.3, at input of the input feature map F (F$\in R^{C\times H\times W}$), after average pooling and max pooling, the feature map of size C×H×W become C×1×1, and then send them into the neural network MLP.The number of neurons in the first layer is C/r, r is the decline rate, the activation function is Relu, the number of neurons in the second layer is C, add the results after completion, and then through a Sigmoid function to get the weight coefficient M_c, calculation method of the weight coefficient is shown in equation (4), and then multiplied by the original input, the new feature after scaling can be obtained.

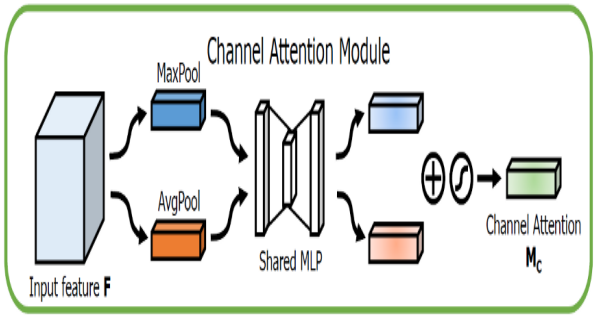


Fig.3. – Channel attention module structure

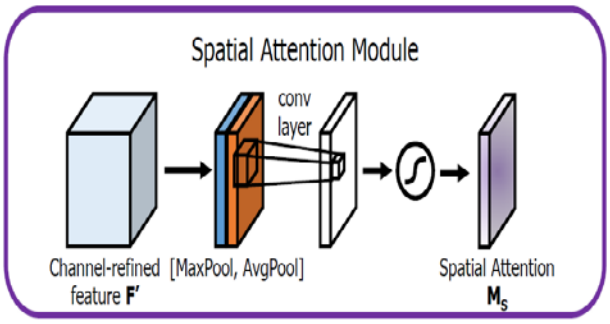


Fig.4. – Spatial attention module structure

$$M_c(\mathrm{F}) = \sigma[W_1(W_0(F_{avg}^C)) + W_1(W_0(F_{max}^C))] \quad (4)$$

in equation (4): σ represents the Sigmoid function; avg represents the global average pooling; max represents the maximum pooling; $W_0 \in R^{C*\frac{c}{r}}$;$F_{avg}^{C}$ represents the average pooling feature of size 1×1×C;F_{max}^{C} represents a max-pooled feature of size 1×1×C.

The architecture of the spatial attention module is shown in Fig.4. The results obtained in the previous step are divided into two channel descriptions with a size of 1 × H × W through maximum pooling and average pooling, and then the tensors are stacked together through the connection operation, get the weight coefficient M_s through the convolution operation and a Sigmoid. The calculation method of the weight coefficient is shown in equation (5). Finally, multiply the input of the previous step by the weight coefficient to obtain the new feature after scaling, and complete the spatial attention operation.

$$M_s(\mathrm{F}) = \sigma(f^{7*7}([F_{avg}^{S};F^{\ s}_{\ max}])) \tag{5}$$

in equation (5): avg represents global average pooling; max represents maximum pooling, f^{7*7} represents a 7*7 convolution, σ represents sigmoid function, F_{avg}^{S} represents average pooling feature, size is 1×H×W, $F^{\ s}_{\ max}$ represents Max pooling features, size is 1×H×W.

The attention mechanism of CBAM is mainly added to the backbone network. In the new version of YOLOv5s model, the BottleneckCSP module is converted into a C3 module. Therefore, we chose to combine the C3 module in Backbone with the attention mechanism and improve it into CBAMC3 module.

Our algorithm is YOLOv5+CBAM+IDPCA. The advantage of the IDPCA algorithm is that it provides a local density calculation method suitable for any dataset, as well as two different residual point allocation strategies, which not only reduces the error propagation, but also effectively improves the clustering efficiency. The advantage of the CBAM attention module is that it can be used with any CNN structure without adding extra consumption and achieves end-to-end training.

4. Experimental results

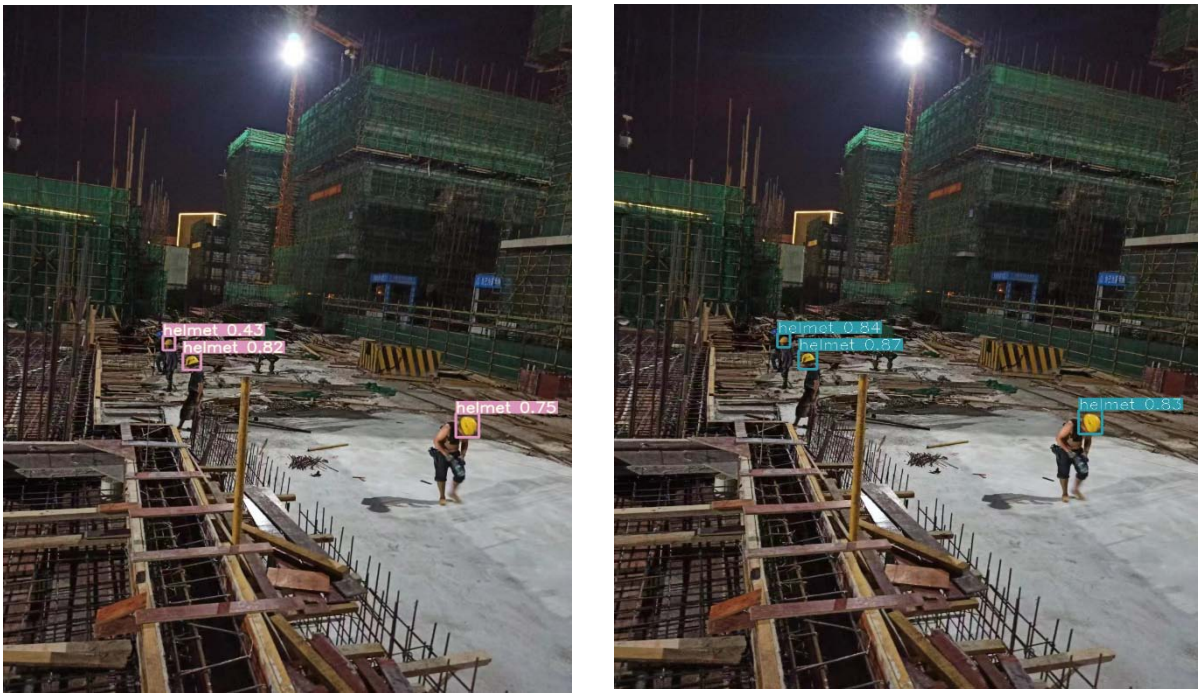
In this experiment, epochs = 200, batch_size = 32, use Warmup method to warm up the learning rate. In the Warmup stage, one-dimensional linear interpolation is used to update the learning rate of each iteration, and after the Warmup stage, the cosine annealing algorithm is used to update the learning rate.

The performance changes caused by network structure changes are gradually verified through ablation experiments. The ablation experiments are divided into test parts such as YOLOv5 + CBAM, YOLOv5 + IDPCA, YOLOv5, and our algorithm. The evaluation indicators P, R, mAP@0.5, mAP@0.5:0.95. The experimental data are shown in Table 3.

Table 3 Analysis of experimental results

Model	Precision	Recall	mAP@0.5	mAP@0.5:0.95
YOLOv5	96.50%	94.21%	96.05%	73.42%
YOLOv5+IDPCA	96.91% (+0.41%)	95.35% (+1.14%)	96.65% (+0.60%)	75.09% (+1.67%)
YOLOv5+CBAM	96.70% (+0.20%)	94.86% (+0.65%)	96.50% (+0.45)	74.91% (+1.49%)
Our algorithm	98.14% (+1.64%)	95.80% (+1.59%)	97.56% (+1.51%)	76.68% (+3.26%)

The comparison of detection accuracy between our algorithm and YOLOv5 is shown in Fig.6. The experiment proves that the improved model has greatly improved the detection accuracy of small objects such as helmets. In some scenes where the environment is complex and the object is not easy to find, the improvement effect is particularly obvious, which basically meets the speed and accuracy requirements of helmet detection in the construction area.



(a) YOLOv5 (b) Our algorithm

Fig.6.– Detection accuracy comparison

Conclusion

Aiming the problems of small objects occupying a small number of pixels, inconspicuous features, and difficulty in distinguishing from the background in object detection, this paper proposes a modified YOLOv5 detection algorithm that optimizes clustering and introduces CBAM. The improved DPC density peak clustering algorithm is used to optimize the size of the prior frame, so that the matching degree between the prior frame and the actual frame size is higher, thereby improving the accuracy of object detection. In addition, the C3 module in the backbone network is combined with the CBAM attention mechanism to improve the model's ability to capture object features and solve the problem that the YOLOv5 algorithm is not good for small object detection.

References

1. Fu Y, Li X, Hu Z. Small-Target Complex-Scene Detection Method Based on Information Interworking High-Resolution Network. Sensors (Basel). 2021 Jul 28. № 21(15). P. 5103. doi: 10.3390/s21155103. PMID: 34372339; PMCID: PMC8348926.
2. Lin T Y, Maire M, Belongie S, et al. Microsoft coco: Common objects in context[C] // European conference on computer vision. Springer, Cham, 2014. P. 740–755.
3. Ultralytics.YOLOv5[EB/OL]. (2020–05–18)[2021–08–12].https://github.com /ultralytics/yolov5.
4. Wang C Y, Liao H Y M, Wu Y H, et al. CSPNet: A new backbone that can enhance learning capability of CNN[C] // Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops. 2020. P. 390–391.
5. 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. P. 3–19.
6. Feng Wang and David MJ Tax. 2016. Survey on the attention based RNN model and its applications in computer vision. arXiv preprint arXiv:1601.06823.
7. Rodriguez A, Laio A. Clustering by fast search and find of density peaks [J] // Science, 2014. № 344(6191). P. 1492–1496.
8. Xie X N, Gao S P, Wu H H. Improved density peak clustering algorithm combining K-nearest neighbors [J][J] // Computer Engineering and Applications, 2018. № 54(7). P. 36–43.