2021 20th International Symposium on Distributed Computing and Applications for Business Engineering and Science (DCABES)

SSD small object detection algorithm based on feature enhancement and sample selection

Zhipeng Liu
School of Artificial Intelligence and Computer
Jiangnan University
Wuxi, China
952738565@qq.com

Wei Fang
School of Artificial Intelligence and Computer
Jiangnan University
Wuxi, China
fangwei@jiangnan.edu.cn

Jun Sun
School of Artificial Intelligence and Computer
Jiangnan University
Wuxi, China
junsun@jiangnan.edu.cn

***Abstract*—SSD has a poor detection effect on small objects. The first reasons is its insufficient feature extraction for small objects. To solve this problem, a feature enhancement module is proposed to make better use of the information around the object to improve the identification ability of small objects. The second reason is that the division of positive and negative samples is unreasonable. The threshold is unfriendly for small objects. To solve this problem, an adaptive training sample selection algorithm is adopted to select the threshold. To improve SSD by the above two methods, and experiments on the PASCAL VOC data set. The mean accuracy precision is increased by 2.6% compared to the SSD algorithm. Compared with the series of SSD improved algorithms such as DSOD, RSSD, DSSD, FSSD, the mAP of our method increased by 2.1%, 1.3%, 1.2%, 1.0%. Our method significantly improved the detection effect of small objects, surpassing SSD and its improved algorithms.**

Keywords-object detection; SSD; small object; feature enhancement; sample selection

I. INTRODUCTION

object detection [1] is one of the important tasks of computer vision. It contains two sub-tasks, one is to accurately identify the object, and the other is to accurately locate the object. In recent years, with the rapid development of deep learning [2], neural networks can extract object features well. Object detection methods based on convolutional neural networks have become a hot issue in the field of object detection. The feature obtained through the deep learning method is stronger than the traditional method. In the wave of deep learning, two types of object detection algorithms have emerged. The one-stage detection algorithms are represented by SSD [3], DSSD [4], DSOD [5], RetinaNet [6], YOLO [7] series, etc. The two-stage detection algorithms are represented by Faster R-CNN [8], Cascade R-CNN [9] and so on. The one-stage detection algorithm has a speed advantage, and the two-stage detection algorithm has a precision advantage, and both have their own advantages.

Small object detection is a difficult task of object detection. The small object occupies few pixels, its image resolution is low, the information is insufficient, and there are fewer features for learning, which leads to the model's poor feature expression ability for small objects. To solve this problem from the perspective of scale, there are FPN [10] algorithm, SNIP [11] algorithm, etc, which integrate high-level feature map with low-level feature map, and make full use of the semantic information of high-level features and the resolution feature of low-level features to improve the expression ability of low-level features; To solve this problem from the perspective of learning the surrounding information of small objects, such as SSH [12,13], etc, which increases the receptive field of the convolutional layer, and better obtain the small object's information through context information information to improve the ability of expressing small objects; To solve this problem from the perspective of the default box, representative of S3FD [14], etc. The author has done a detailed default box experiment, and designed the default box more reasonably, so the default box can match small objects better. To solve this problem from the perspective of the matching strategy, such as Cascade R-CNN, etc, which do not set too strict IoU threshold to ensure the number of small objects' default boxes. The above-mentioned solutions have improved the detection effect of small objects, which has greatly inspired the research work of this paper.

In order to improve the detection effect of SSD on small objects, this paper starts from two perspectives. Firstly, this paper proposes a feature enhancement module, which can better learn the surrounding information of small objects and make up for the shortcomings of insufficient features of small objects; Secondly, this paper adopts a positive and negative sample selection strategy. The original SSD's determination threshold for positive samples is too strict. The number of default boxs corresponding to the object is originally relatively small. After the hard threshold filtering, the number of positive sample default boxs corresponding to the remaining small objects will be less, resulting in insufficient training of small objects, so this paper adopts an adaptive training sample selection method. Choose an appropriate threshold for each object, and use this threshold for sample selection. The experimental results show that the improved SSD algorithm in this paper has significantly improved the detection effect of small objects. From the results of the PASCAL VOC2007 test set, the detection accuracy of the three categories such as bottle, pottedplant and chair has been significantly improved. The three categories have a large number of small objects. Our method is better than SSD and a series of SSD improved algorithms.

2473-3636/21/$31.00 ©2021 IEEE
DOI 10.1109/DCABES52998.2021.00031

II. SMALL OBJECT DETECTION ALGORITHM

This paper improves SSD from two aspects. Firstly, a feature enhancement module is proposed to supplement the feature of the small object by fully learning the edge information of the small object, and improve the detection ability of the SSD detector for the small object. Secondly, adopt an adaptive positive and negative sample selection strategy to replace SSD's original strategy. The strategy adaptively select the IoU threshold to determine the positive which can ensure the number of positive sample boxes for small objects. The improved SSD algorithm proposed in this paper significantly improves the detection effect of small objects.

A. Feature Enhancement Module

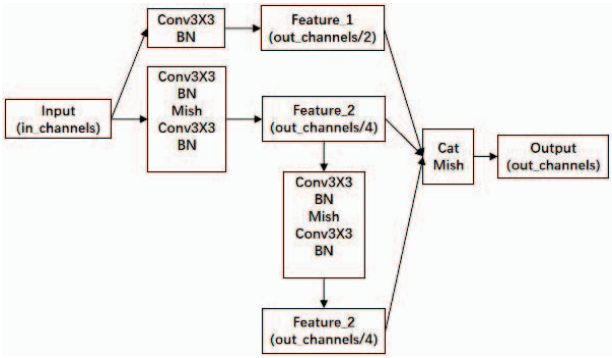

Figure 1. Network structure of feature enhancement module.

Feature enhancement module, we can called it FM. Fig. 1 shows its structure.

The size of the convolution kernel used by the FM is 3. If the convolution kernel is 1, rich surrounding information cannot be extracted. If the size of the convolution kernel is larger, it will increase a lot of parameters and increase the computational cost. After convolution, the FM uses BatchNorm to normalize the data, adjust the data distribution, and accelerate training. The FM module uses the Mish function as the activation function. In view of the long-term dominant position of ReLU in the activation function of deep learning, we compares the Mish function and the ReLU function. The formulas of Mish and ReLU are respectively as follows.

$$g(z) = z * \tanh\left(z * \ln\left(1 + e^{z}\right)\right)$$

$$g(z) = \begin{cases} z & z > 0 \\ 0 & z \le 0 \end{cases}$$

The function image of the two function is shown in Fig. 2. It is not difficult to find that ReLU is directly set to zero for negative values, and Mish [15] has a better gradient flow for negative values, so this module uses Mish as the activation function.

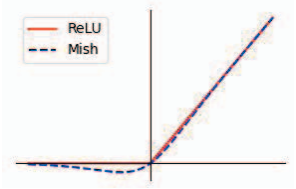

Figure 2. Image function of ReLU and Mish.

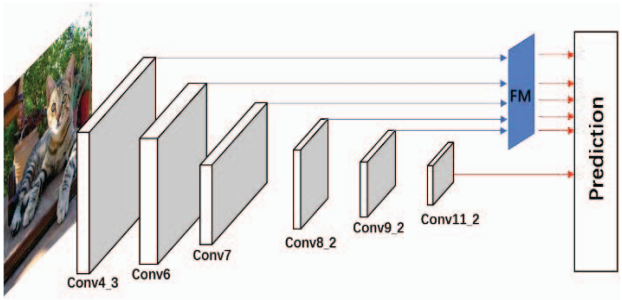

Figure 3. Network structure of SSD embedded FM.

The structure of FM embedded in the original SSD is shown in Fig. 3. In this paper, after the feature map of Conv4_3, Conv6, Conv7, Conv8_2, Conv9_2, the FM module is connected. Then put the five layer's output and Conv11_2 feature map into prediction branch. We only uses the FM module for the first five layers of feature maps, and the last layer of feature maps does not use the FM module. The reason is that the size of the last layer of feature maps Conv11_2 has been 1, and the kernel size 3 in the FM module is no longer applicable.

B. Adaptive Training Sample Selective

The adaptive training sample selection [15] method can better determine the IoU threshold. It selects positive and negative samples according to the statistical characteristics of the ground truth boxes. The IoU threshold can be determined adaptively for each real object. This method doesn't increase any computational overhead to improve the performance of SSD detector. The algorithm steps are as follows.

Input: G、L、A_i、A、k .Output: P、N . G represents the set of all ground truth boxes in the image; L represents the number of feature maps; A_i is the set of default boxes for the i-th feature map; A is the set of all default boxes; k represents the number of default boxes selected from each feature map; P represents the set of positive samples; N represents the set of negative samples.

a) For each ground truth g , $C_g \leftarrow \Phi$.

b) For each feature map i , $i \in [1, L]$,select k default boxes whose center closest to g from A_i based on L2 distance. These selected boxes as S , $C_g = C_g \cup S$.

c) Compute IoU between C_g and g , as D_g .

d) Compute mean of D_g , as m_g .

e) Compute variance of m_g , as v_g .

f) For each g , set threshold as t_g , $t_g = m_g + v_g$.

g) Compute mean of D_g , as m_g .For each candidate box o , $o \in C_g$, if IoU between o and g larger than t_g and its center in g , it will be viewed as positive sample, written as P , otherwise as negative sample, written as N , $N = A - P$.

h) Return P、N .

The algorithm is not sensitive to the value k . In this paper, experiments are conducted on the PASCAl VOC dataset, and the final value selected is 11, because the results obtained at this time are slightly better than other values.

III. EXPERIMENT

A. Experienmental Steps

a) Build this paper's network.

b) Train VGG16 on the ImageNet, get pretrained VGG16 .

c) Make migration learning, remove the fully connected layer of the VGG16 model, load its weights as the initial parameters of the backbone.Use kaiming distribution to initialize the category prediction branch and location prediction branch, and initialize the rest of the parameters randomly.

d) Load the VOC2007 and VOC2012 training set pictures, set the picture size to 300×300, use random cropping, flipping and other methods for data augmentation, and input the network for training.

e) Use stochastic gradient descent method as the optimizer, the initial learning rate is set to 0.001, the batchsize is set to 32, and the maximum number of iterations is set to 200.

f) Save the training model.

B. This paper's algorithm compared with original SSD

As shown in Table. 1. Compared with the original SSD algorithm, the algorithm in this paper has significantly improved the detection effect on bottle, potted plant, and chair which have a large number of small objects. AP increased by 10.8%, 4.0%, and 2.4% respectively. These three categories' mAP increased by 5.7%, and the rest 17 categories' mAP increased by 2.1%. The mAP of the overall twenty categories of objects increased from 77.2% to 79.8%.

TABLE I. COMPARISION OF OURS AND SSD ON ALL CLASSES

	SSD300/%	**ours/%**
mAP	77.2	79.8
aero	82.2	86.4
bike	84.7	87.9
bird	74	78.8
boat	68.8	73.9
bottle	50.1	60.9
bus	84.4	87.5
car	86.2	86.8
cat	87.9	86.6
chair	61.7	64.1
cow	82.0	84.0
table	74.5	81.1
dog	85.4	86.5
horse	87.2	85.8
mbike	83.2	85.2
person	78.1	79.6
plant	51.3	55.3
sheep	77.5	79.4
sofa	80.7	80.4
train	87.7	85.8
tv	76.6	79.7

Fig. 4 is a comparison chart of the recall and precision of the above three categories.The solid line represents the algorithm in this paper, and the dotted line represents the original SSD algorithm. It can be seen that the algorithm in this paper has a significant improvement in the recall rate of small objects.

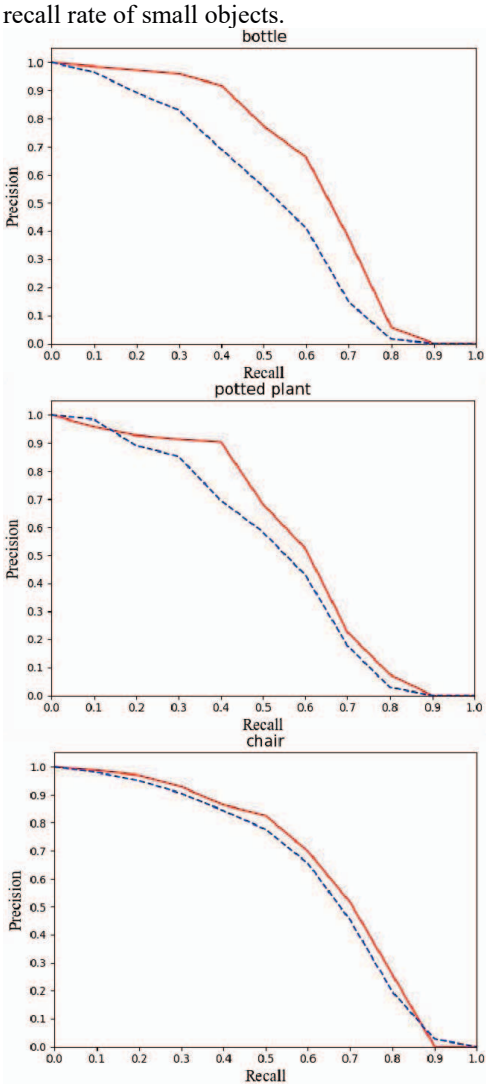

Figure 4. Comparison of recall and precision.

C. This paper's algorithm compared with SSD's improved algorithm

TABLE II. COMPARISION OF OUR METHOD AND OTHER METHOD BASED ON SSD

Method	**Backbone**	**Input size**	**Boxes' num**	**Speed (FPS) Titan X**	**mAP**
SSD300	VGG	300×300	8732	46	77.2
DSOD300	VGG	300×300	---	17.4	77.7
RSSD300	DS/ 64-192-48-1	300×300	8732	35	78.5
DSSD321	ResNet-101	321×321	17080	9.5	78.6
FSSD300	VGG	300×300	8732	35.6	78.8
ours	VGG	300×300	8732	38.3	79.8

As shown in Table. 2, compared with the SSD's improved algorithms DSOD, RSSD, DSSD, FSSD, the algorithm in this paper has the lowest input image resolution, the smallest number of default boxes, the fastest speed, and the highest accuracy.

The intuitive comparison of speed and accuracy is shown in Fig. 5.

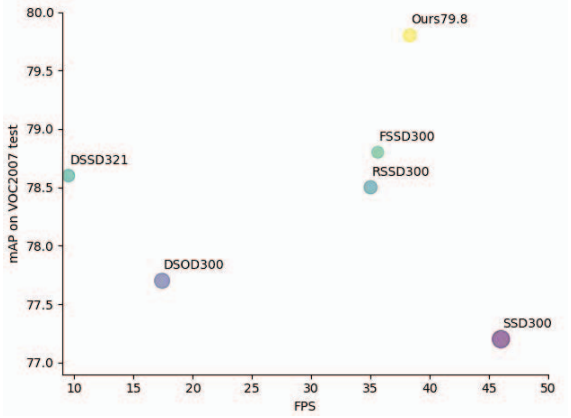

Figure 5. Comparison of speed and mAP.

D. Pictures shows

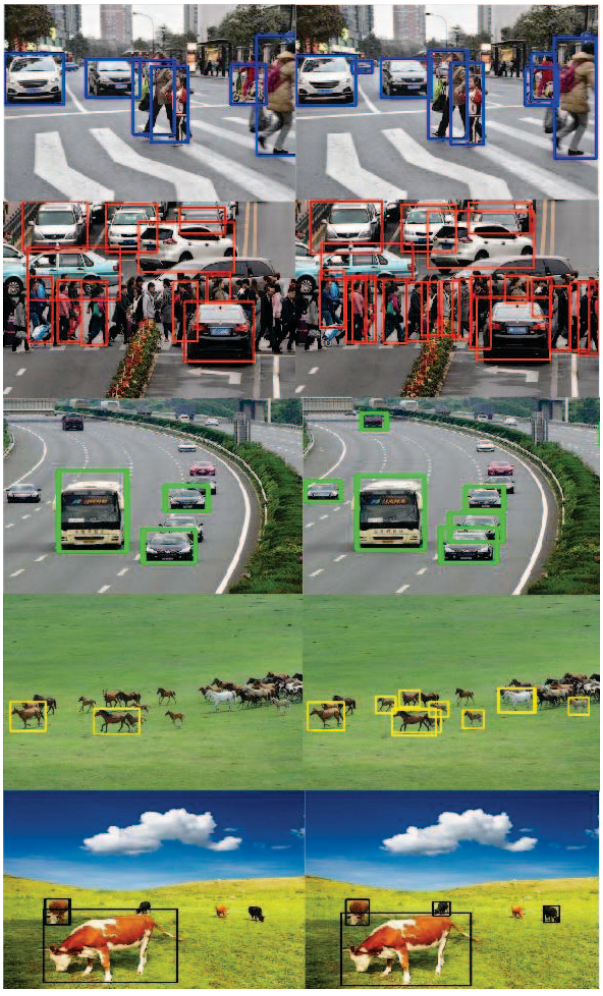

Figure 6. Picture shows of our method and SSD.

In order to verify the effectiveness of the improved SSD algorithm in this paper, we download some pictures with small objects from the Internet and compare the experimental results with the original SSD algorithm. As shown in Fig. 6, the labels and scores are removed for better observation. For the same picture, the left one is the detection result of the original SSD algorithm and the right one is the detection result of the improved SSD algorithm in this paper.

IV. CONCLUSION

The results of comparative experiments show that compared with the original SSD algorithm, the improved algorithm in this paper has significantly improved the detection effect on the VOC2007 test set, and surpasses a series of improved algorithms such as DSSD, DSOD, and FSSD in accuracy and speed. This proves the efficiency and advantages of the algorithm in this paper.

REFERENCES

[1] Wu X, Sahoo D, Hoi S C H. Recent advances in deep learning for object detection[J]. Neurocomputing, 2020.

[2] Bengio Y, Courville A, Vincent P. Representation learning: A review and new perspectives[J]. IEEE transactions on pattern analysis and machine intelligence, 2013, 35(8): 1798-1828.

[3] Liu W, Anguelov D, Erhan D, et al. Ssd: Single shot multibox detector[C]//European conference on computer vision. Springer, Cham, 2016: 21-37.

[4] Deconvolutional Single Shot Detector[J]. 2017.Fu C Y, Liu W, Ranga A, et al. Dssd: Deconvolutional single shot detector[J]. arXiv preprint arXiv:1701.06659, 2017.

[5] Shen Z, Liu Z, Li J, et al. Dsod: Learning deeply supervised object detectors from scratch[C]//Proceedings of the IEEE international conference on computer vision. 2017: 1919-1927.

[6] Lin T Y, Goyal P, Girshick R, et al. Focal loss for dense object detection[C]//Proceedings of the IEEE international conference on computer vision. 2017: 2980-2988.

[7] Redmon J, Divvala S, Girshick R, et al. You only look once: Unified, real-time object detection[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 779-788.

[8] Ren S, He K, Girshick R, et al. Faster r-cnn: Towards real-time object detection with region proposal networks[C]//Advances in neural information processing systems. 2015: 91-99.

[9] CaiCai Z, Vasconcelos N. Cascade r-cnn: Delving into high quality object detection[C]//Proceedings of the IEEE conference on computer vision and pattern recognition.2018:6154-6162

[10] Lin T Y, Dollár P, Girshick R, et al. Feature pyramid networks for object detection[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 2117-2125.

[11] Singh B, Davis L S. An analysis of scale invariance in object detection snip[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 3578-3587.

[12] Najibi M, Samangouei P, Chellappa R, et al. Ssh: Single stage headless face detector[C]//Proceedings of the IEEE international conference on computer vision. 2017: 4875-4884.

[13] Deng J, Guo J, Zhou Y, et al. Retinaface: Single-stage dense face localisation in the wild[J]. arXiv preprint arXiv:1905.00641, 2019.

[14] Zhang S, Zhu X, Lei Z, et al. S3fd: Single shot scale-invariant face detector[C]//Proceedings of the IEEE international conference on computer vision. 2017: 192-201.

[15] Misra D. Mish: A self regularized non-monotonic neural activation function[J]. arXiv preprint arXiv:1908.08681, 2019.