Analysis of small target detection algorithm based on SSD and YOLOv5

Wen Zhou, Yan Gou, Langlang Chen, Tian Shi, Zisu Yuan

School of Information Engineering, Nanjing University of Finance and Economics, Nanjing, Jiangsu, 210023, China

Abstract: *SSD is a single-stage target detection algorithm, which performs feature extraction by convolutional neural network and takes different feature layers for detection output, so SSD is a multi-scale detection method. In the feature layer to be detected, a 3*3 convolution is directly used to perform the transformation of the channels. ssd uses an anchor strategy with pre-defined anchors of different aspect ratios, and each output feature layer predicts multiple detection frames (4 or 6) based on the anchor. A multi-scale detection approach is used, where a shallow layer is used to detect small targets and a deep layer is used to detect large targets. yolov5 is a single-stage target detection algorithm, which adds some new and improved ideas to yolov4, resulting in a significant performance improvement in both speed and accuracy. We conduct algorithm experiments with SSD and YOLOv5, and analyze the experiments to obtain better improvement ideas for small target algorithm.*

Keywords: *SSD; YOLOv5; target detection algorithm*

1. Introduction

SSD and YOLOv5 algorithms are both one-stage algorithms in target detection algorithms, small target detection is a difficult and painful point in computer vision in recent years, because small targets are not easy to find in samples compared to large targets, such as Figure 1, low resolution, few features, unbalanced samples and other problems. In recent years, many researchers based on the existing target detection algorithm to improve the application of small target detection, in order to more effectively improve the algorithm of small target detection and the needs of realistic applications, this paper selects YOLOv5 from the first-stage target detection algorithm with SSD for experiments, analyzes the experimental results and process, and comes up with suggestions to improve the small target detection algorithm[1-2].

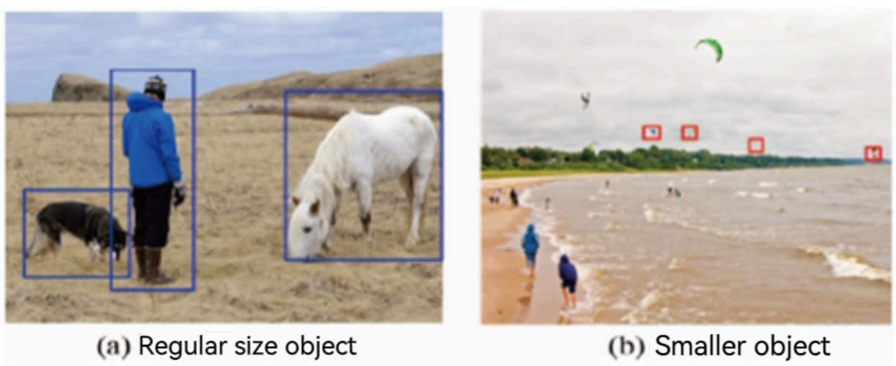


Figure 1: Example of a small target

1.1. Experimental equipment

Ubuntu Server 20.04 LTS 64 + 6-core 56GB Tencent GPU Compute N8.

1.2. Contribution

The research topic aims at evaluating the two models through experiments, processing not only the evaluation of the two models, but also discussing and summarizing them in order to find improvement strategies for the small target detection algorithm[3-5].

2. Related Work

2.1. Previous work

Before the experimental study, we did the following work.

1) Read the SSD and YOLOv5 algorithm papers and refer to related literature.

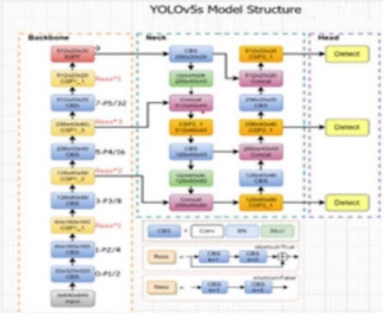
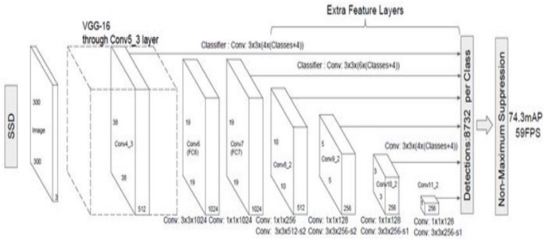
2) Create a code repository, reproduce the code in the papers on the local computer, and upload it to the corresponding code repository.

3) Download the COCO dataset and VOC dataset, and use labelImg for data annotation work.

4) Set up the experimental environment (git, conda, python, C++, pytorch, etc.) in the server and upload the corresponding experimental datasets, as shown in Table 1.

2.2. Network structure analysis of SSD and YOLOv5

Table 1: Algorithm and Network Structure

Algorithm	YOLOv5	SSD
Network Structure	 YOLOv5s Model Structure Backbone Neck Head Detect Detect Detect	 Extra Feature Layers VGG-16 through Conv5_3 layer Classifier : Conv: 3x3x(4x(Classes+4)) Classifier : Conv: 3x3x(6x(Classes+4)) SSD 300 Image 300 3 38 Conv4_3 38 512 19 Conv6 (FC6) 19 1024 19 Conv7 (FC7) 19 1024 10 Conv8_2 10 512 5 Conv9_2 5 256 3 Conv10_2 3 256 Conv11_2 256 Conv: 3x3x(4x(Classes+4)) Detections:8732 per Class Non-Maximum Suppression 74.3mAP 59FPS Conv: 3x3x1024 Conv: 1x1x1024 Conv: 1x1x256 Conv: 3x3x512-s2 Conv: 1x1x128 Conv: 3x3x256-s2 Conv: 1x1x128 Conv: 3x3x256-s1 Conv: 1x1x128 Conv: 3x3x256-s1

3. Experiments

3.1. Introduction to the data set

This experiment is based on the VOC and COCO datasets on:

3.1.1. VOC

Pascal VOC challenge is a very popular dataset for building and evaluating algorithms for image classification, object detection and segmentation.

As a standard dataset, voc-2007 is a benchmark for image classification and recognition capabilities. voc dataset contains: training set (5011 images), test set (4952 images), total 9963 images, containing 20 categories in total. (bicycle bird boat bottle bus car cat chair cow ......). The test data label of VOC2007 has been published, but the subsequent ones have not been published (only images, no label). Faster-rcnn, YOLO are using this dataset as a demo example.

3.1.2. COCO

COCO, whose full name is Common Objects in COntext, is a dataset provided by the Microsoft team that can be used for image recognition. Images in the MS COCO dataset are divided into training, validation, and test sets. COCO collects images by searching 80 object categories and various scene types on Flickr, and its use of Amazon's Mechanical Turk (AMT).

The COCO dataset is an image recognition+segmentation+captioning dataset acquired by the Microsoft team, which has the following main features: (1) Object segmentation (2) Recognition in Context (3) Multiple objects per image (4) More than 300,000 images (5) More than 2 Million instances (6) 80 object categories (7) 5 captions per image (8) Keypoints on 100,000 people.

The dataset addresses 3 main problems: target detection, contextual relationships between targets, and precise localization of targets on 2 dimensions.

3.2. Experimental setup

1) Before starting the experiment, we set the learning rate to 0.01 uniformly.

2) Before we start, we set the training model we need to train the algorithm on, the maximum number of data to train per batch, and the yaml file address of the training dataset.

Experimental setup of YOLOv5:

python train.py --data coco.yaml --cfg yolov5n.yaml --weights " --batch-size 128		
	yolov5s	64
	yolov5m	40
	yolov5l	24
	yolov5x	16

Experimental setup of SSD:

```
python train.py --config-file configs/vgg_ssd300_voc0712.yaml
```

3.3. Model experimental results and analysis

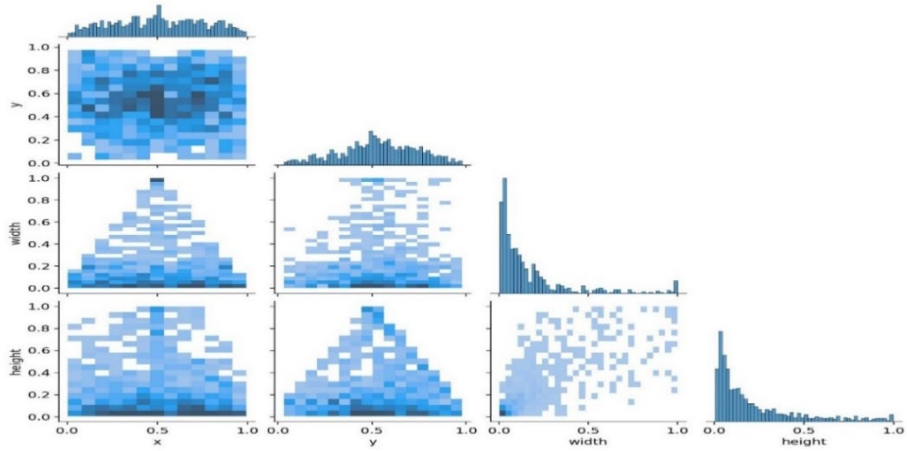


Figure 2: Labels_correlogram

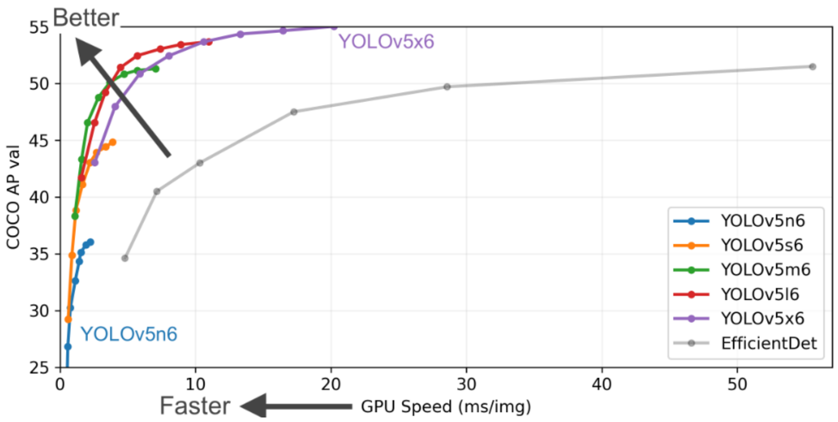
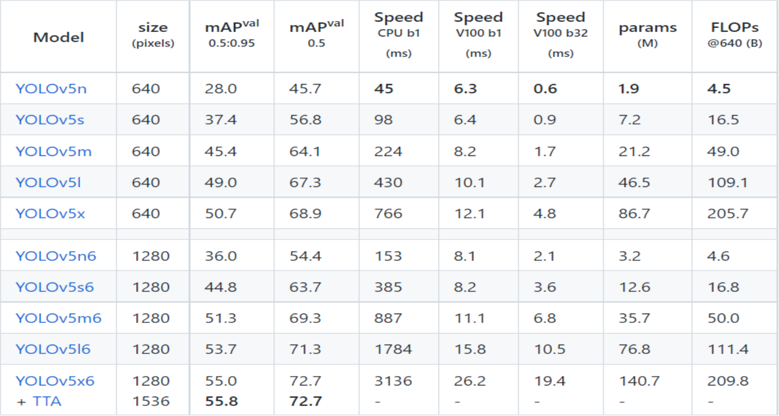


Figure 3: Comparison of GPU reasoning speed of algorithms on COCO data



Model	size (pixels)	mAP^{val} 0.5:0.95	mAP^{val} 0.5	Speed CPU b1 (ms)	Speed V100 b1 (ms)	Speed V100 b32 (ms)	params (M)	FLOPs @640 (B)
YOLOv5n	640	28.0	45.7	**45**	**6.3**	**0.6**	**1.9**	**4.5**
YOLOv5s	640	37.4	56.8	98	6.4	0.9	7.2	16.5
YOLOv5m	640	45.4	64.1	224	8.2	1.7	21.2	49.0
YOLOv5l	640	49.0	67.3	430	10.1	2.7	46.5	109.1
YOLOv5x	640	50.7	68.9	766	12.1	4.8	86.7	205.7
YOLOv5n6	1280	36.0	54.4	153	8.1	2.1	3.2	4.6
YOLOv5s6	1280	44.8	63.7	385	8.2	3.6	12.6	16.8
YOLOv5m6	1280	51.3	69.3	887	11.1	6.8	35.7	50.0
YOLOv5l6	1280	53.7	71.3	1784	15.8	10.5	76.8	111.4
YOLOv5x6 + TTA	1280 1536	55.0 **55.8**	72.7 **72.7**	3136 -	26.2 -	19.4 -	140.7 -	209.8 -

Figure 4: Comparison of relevant detection index results of YOLOv5 series algorithms on COCO dataset

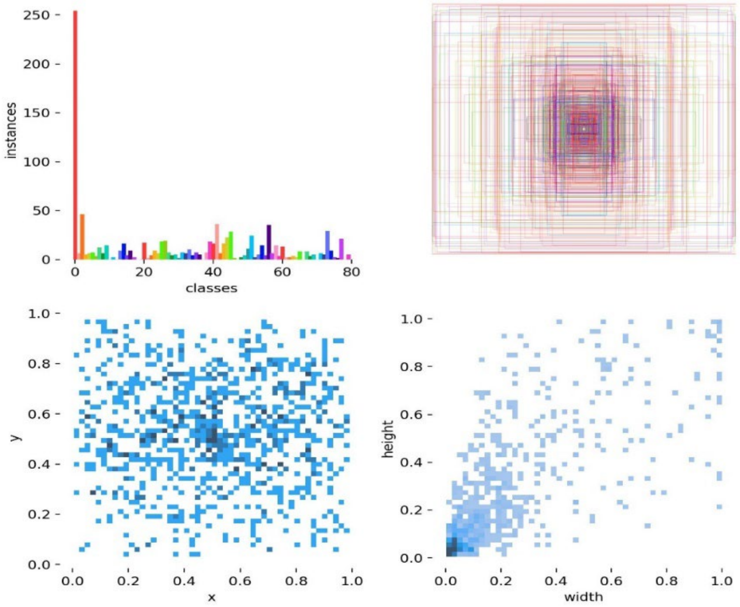


Figure 5: Features labels

Due to the equipment, the SSD algorithm was experimented on the VOC, as shown in Figure 2, Figure 3, Figure 4 and Figure 5. During the experiment, 30 training rounds were conducted, and the maximum number of training batches per round was 50, of which some of the experimental results are as follows shown in Figure 6, Figure 7, Figure 8 and Figure 9:

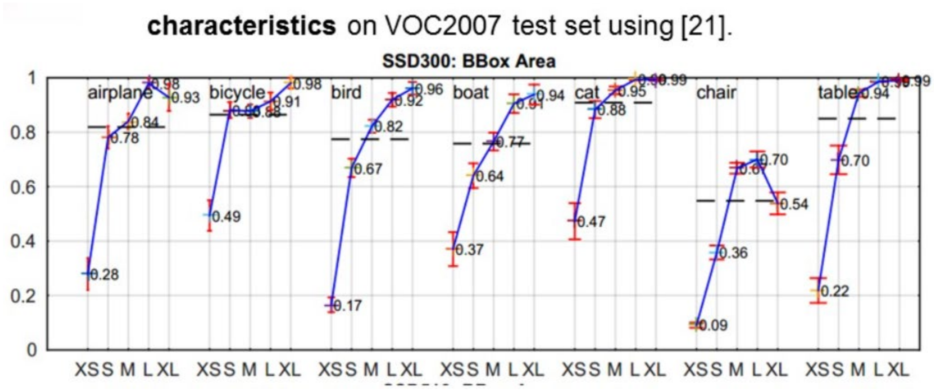


Figure 6: Characteristics on VOC2007 test set using

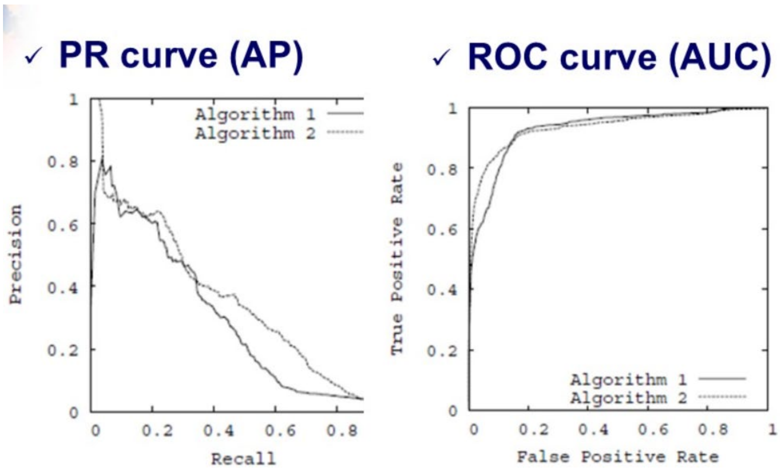


Figure 7: The PR curve and ROC curve (The Algorithm 1 is YOLOv5, and the Algorithm 2 is SSD)

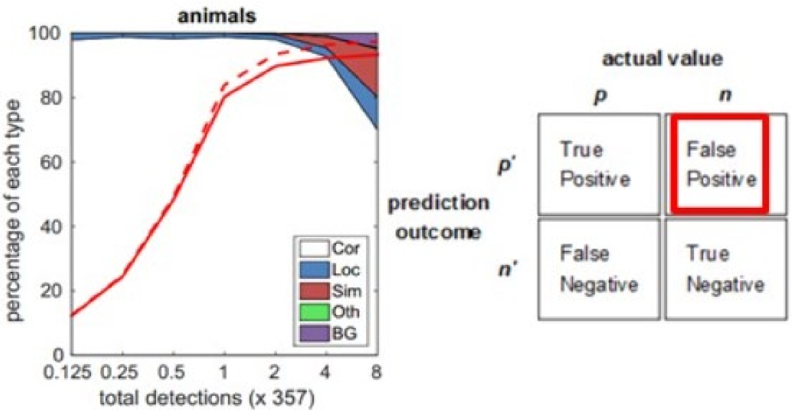
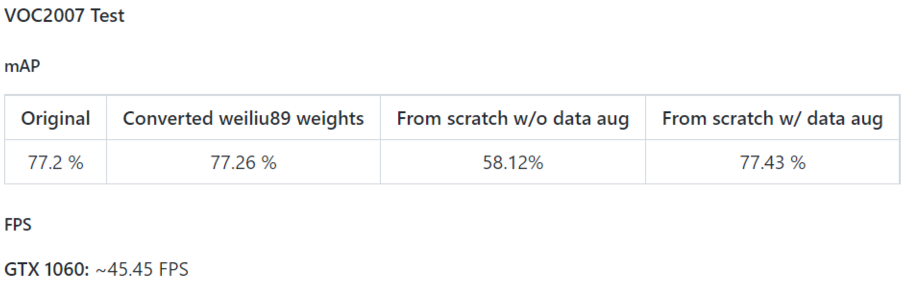


Figure 8: Percentage diagram and confusion matrix diagram of animal detection categories


VOC2007 Test

mAP

Original	Converted weiliu89 weights	From scratch w/o data aug	From scratch w/ data aug
77.2 %	77.26 %	58.12%	77.43 %

FPS

GTX 1060: ~45.45 FPS

Figure 9: Related metrics of SSD algorithm for detection on VOC2007test

3.4. Algorithm Summary

The purpose of this experiment is to seek the improvement direction and research direction of the small target detection algorithm through experiments in the existing target detection algorithm, choose mAP as the evaluation index to see, although YOLOv5 and SSD1 achieved good results, but they are weak for small target detection, small target detection in this experiment we found the following defects:

Small target objects are small in the image with respect to the image, which is not conducive to localization and recognition.

Deep neural networks need a large amount of data to train, and the existing data set, the number of images about small target objects is relatively small, which is not conducive to training a small target detection recognizer that can be widely used.

If there are defects or interference categories around small target objects, it is more difficult to detect small target recognition classifier, and it may even cause problems such as recognition errors.

Small target labeling requires a lot of labor cost, and the existing target detection models are strongly supervised algorithmic detection models.

The image resolution of small target objects is relatively low, which leads to difficulties in recognition and localization of small target algorithm detectors.

Some of the small target objects are extremely small, which existing deep learning and future deep learning networks cannot go to solve, and other knowledge needs to be used to make up for this lack.

In the future theories and methods of deep learning continue to move forward, target detection algorithms have achieved better results on general-purpose data sets. However, small target detection is still a difficult problem, and most of the research on it is based on the improvement and optimization of existing algorithms. For the future development of small target detection, there is an improvement direction and research direction worth exploring[6-7].

3.5. Directions for improvement

1) Data enhancement strategy:

(a) Improving the resolution of image acquisition

The resolution of image acquisition can be improved by resampling at different combinations of multiples such as 4x, 8x, 16x, etc., which helps small target objects to be located and identified more easily in the image.

(b) Increasing the input resolution of the model

By increasing the resolution of the model input facilitates the localization of small target objects and the reduction of recognition difficulty.

(c) Tile your images

The tile not only improves the image resolution and generates multiple data samples, which helps to improve the generalization ability of the model and reduce the difficulty of locating frames.

(d) Generating more data by augmentation

By augmentation strategy, more data is generated, and a large amount of data training will make the algorithm more robust and capable of small target detection.

2) Automatic learning model anchors

By self-learning model anchors, the algorithm can improve the localization ability of small target objects, which is obviously very difficult, but it is believed that the self-learning model anchors can be achieved by neural networks, and the anchor box can be automatically adjusted to make the adaptive target object localization.

3) Filter out irrelevant categories

By filtering out irrelevant categories not only can reduce the difficulty of locating and identifying small target objects, but also can reduce the problems of missing and missing detection of small target objects caused by object stacking.

3.6. At the same time, there are several research directions worth exploring

3.6.1. Designing a network dedicated to feature extraction

Most of the existing target detection models use the classification network pre-trained on ImageNet as the backbone network, and the design principles of the models in the classification task and the detection task are different, and the variability of target size distribution among data sets leads to certain problems in small target detection. Therefore, designing specialized feature extraction networks for small targets is one of the important research directions for small target detection in the future.

3.6.2. Combining other non-deep learning disciplines such as machine learning and mathematics

For very small targets, such as those at the 4×4 pixel level, the depth network model may have inherent defects, and the feature extraction stage leads to further loss of limited information about the target. At this time, it is necessary to consider the idea of combining the depth network model with non-depth methods, for example, by combining with non-depth methods such as saliency detection and super pixel segmentation to extract more effective information and complete the subsequent target localization and recognition tasks.

3.6.3. Establishing a multi-task learning mechanism

A multi-task learning model is established to learn jointly with other types of tasks to obtain more information beneficial to small target detection, and to improve the performance of small target detection through the parameter sharing mechanism of multi-task learning. For example, using the target edge and context information captured by the semantic segmentation task helps to better identify and localize targets; combining the counting task with the target detection task and learning the counting-assisted target detector can eliminate the interference of complex backgrounds to a certain extent and avoid missed detection.

3.6.4. Building a non-strongly supervised target detection model

Accurate target bounding box annotation information is a prerequisite for strongly supervised target detection, but since small target annotation requires a large amount of labor cost and is inefficient, obtaining good detection results and improving detection efficiency through low-cost annotation information is a current research hotspot. Using weakly supervised learning, small sample learning, and self-supervised learning to build a non-strongly supervised detection model to complete small target detection in the absence of calibration data is a direction worthy of in-depth research in the future.

References

[1] Li Hongguang, Yu Ruonan, Ding Wenrui. Research progress of small target detection based on deep learning [J]. Journal of Aeronautics, 2021, 42(7):19. (In chinese)

[2] Zhang R, Wen C. SOD-YOLO: A Small Target Defect Detection Algorithm for Wind Turbine Blades Based on Improved YOLOv5[J]. Advanced Theory and Simulations, 2022(7):5.

[3] Tang C, Ling YS, Zheng KD, et al. Deep learning based multi-window SSD target detection method[J]. Infrared and Laser Engineering, 2018, 47(1):9. (In chinese)

[4] Li Hongyan, Wu Chengke. A small target detection algorithm based on wavelet and genetic algorithm [J]. Journal of Electronics, 2001, 29(004):439-442. (In chinese)

[5] Jiaji Z, Zhengyuan T, Jie Y, et al. Infrared small target detection based on image sparse representation. 2011. (In chinese)

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

[7] Chen Keqi, Zhu Zhiliang, Deng Xiaoming, Ma Cuixia, & Wang Hongan. (2020). A review of deep learning research on multi-scale target detection. Journal of Software, 32(4), 1201-1227. (In chinese)