Target detection in remote sensing image based on deep learning

Lianchen Zhao [1,2], Yizhun Peng [1,2*], Di Li[1,2], Yuheng Zhang[12]
[1]*College of Electronic Information and Automation,*
Tianjin University of Science and Technology, China

[2]*Advanced Structural Integrity International Joint Research Centre,*
Tianjin University of Science and Technology, China
*E-mail: * pengyizhun@tust.edu.cn*
www.tust.edu.cn

Abstract

For high-resolution optical remote sensing images, there are still many challenges in target detection. In this paper, deep learning algorithm is used to detect the target in remote sensing image. Improve and optimize the deep learning target detection algorithm. When the selected data set is used for target detection, the AP value is improved, which leads to the concept of multi-scale feature fusion feature pyramid and residual network. By improving the selected Yolov3 network model, the detection effect of the two targets of aircraft and ships in remote sensing images has been significantly improved.

Keywords: Deep learning; Target detection; Residual network; Remote sensing image

1. Introduction

For high-resolution optical remote sensing images, there are still many challenges in target detection. This paper introduces the multi-scale target detection in optical remote sensing images by using Yolov3[1], and compares the detection results of multi-scale targets on the selected remote sensing data sets by using the improved Yolov3 model.

Target detection is an important direction of computer vision and digital image processing[2]. Target detection is also widely used in robot navigation, real-time vehicle monitoring, defective product detection in industry, aerospace ship detection and other target detection. Through the algorithm of target detection in deep learning, we can realize the accurate detection of the target, which not only has a great reduction in manpower, but also has a high efficiency improvement, which has important practical significance. Target detection algorithms in deep learning can be divided into two categories: one-stage algorithm and two-stage algorithm[3].

2. One-Stage target detection algorithm

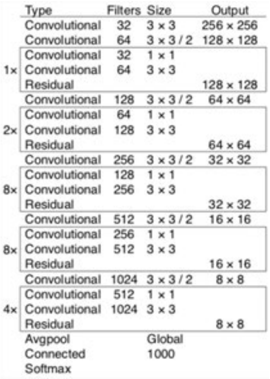
One-Stage, a target detection algorithm based on deep learning, has a faster detection speed when detecting the target[4], because the target detection algorithm discards the process of single region recommendation. This algorithm is first proposed from CVPR 2016, and Yolo (you only look once: unified, real time object detection) is an innovative one stage detection.

The method of generating candidate regions first and then detecting has relatively high accuracy, but it is relatively slow in terms of detection speed.

Yolo did not remove the candidate regions[5], but directly divided the input network image into 49 7 * 7 grids. Any existing grid predicted two boundary boxes, so as to achieve the prediction of 98 boundary boxes. It can be roughly understood as a rough selection of 98 candidate regions on the input image, which cover the entire region of the image. Therefore, regression prediction is used to compare the existing 98 candidate boxes to get the final applicable boundary box.

Yolo network draws lessons from the structure of GoogleNet[6] classification network. Compared with

GoogleNet classification network, the 1x1 convolution layer and 3x3 convolution layers are used to replace the inception module in Yolo. There are 24 convolution network layers and 2 full connection layers in the whole detection network. Compared with the previous version of Yolo, Yolov3 adjusts the network structure, and uses the Darknet-53[7] network structure compared with Yolov2. The specific network structure is shown in Figure 1 below:



	Type	Filters	Size	Output
	Convolutional	32	3 × 3	256 × 256
	Convolutional	64	3 × 3 / 2	128 × 128
1×	Convolutional	32	1 × 1	
	Convolutional	64	3 × 3	
	Residual			128 × 128
	Convolutional	128	3 × 3 / 2	64 × 64
2×	Convolutional	64	1 × 1	
	Convolutional	128	3 × 3	
	Residual			64 × 64
	Convolutional	256	3 × 3 / 2	32 × 32
8×	Convolutional	128	1 × 1	
	Convolutional	256	3 × 3	
	Residual			32 × 32
	Convolutional	512	3 × 3 / 2	16 × 16
8×	Convolutional	256	1 × 1	
	Convolutional	512	3 × 3	
	Residual			16 × 16
	Convolutional	1024	3 × 3 / 2	8 × 8
4×	Convolutional	512	1 × 1	
	Convolutional	1024	3 × 3	
	Residual			8 × 8
	Avgpool		Global	
	Connected		1000	
	Softmax			

Fig. 1. Darknet-53 network structure

In Figure 1, the main structure of the network structure of darknet-53 mainly consists of DBL module, Upsample module, Shortcut module, Res module and Route module. The main components of DBL are convolution network, BN and Leaky relu.

The size of network input image is 416 * 416 * channels. After 5 times of down sampling and 2 times of up sampling stitching through convolution layer, three kinds of feature maps will be output. There are 53 layers of convolutions in the darknet-53 network. Except for the last FC, there are 52 convolutions as the main network model structure. For the low-level convolution layer, the field of vision is relatively small, which is responsible for the detection of small targets. For the deep convolution layer, the field of vision is relatively large, which is responsible for the detection of larger targets.

The activation function adopted by yolov3 is leaky relu. Compared with relu, this activation function sets all negative values to zero. Leaky relu function only gives a non-zero slope to all negative values. The mathematical expression is shown in (1)：

$$y_i = \begin{cases} x_i & if \quad x_i \geq 0 \\ \dfrac{x_i}{a_i} & if \quad x_i < 0 \end{cases} \tag{1}$$

The leaky relu function image is shown in Figure 2:

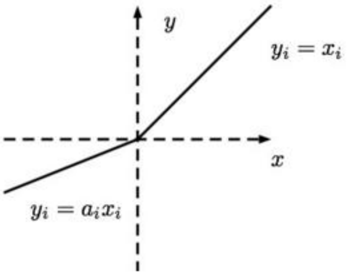


Fig. 2. Leaky Relu function graph

In Yolov3, anchor boxes are needed in bounding box prediction. In Yolov3, K-means clustering algorithm is used to obtain anchors suitable for data sets. The goal of K-means algorithm is to divide n samples into K clusters according to the calculation results[8], so that the similar samples in n samples can be divided into the same cluster. The calculation method used to measure the similarity is to use the size of Euclidean distance (2).

$$d(x,y) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \cdots + (x_n - y_n)^2} = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \tag{2}$$

The loss function in Yolov3 is shown in formula (3) as follows:

$$\begin{aligned} Loss = {} & \lambda_{coord} \sum_{i=0}^{S^2} \sum_{j=0}^{B} I_{ij}^{obj} [(x_i - \hat{x}_i^j)^2 + (y_i - \hat{y}_i^j)^2] + \\ & \lambda_{coord} \sum_{i=0}^{S^2} \sum_{j=0}^{B} I_{ij}^{obj} [(\sqrt{w_j^i} - \sqrt{\hat{w}_j^i})^2 + (\sqrt{h_j^i} - \sqrt{\hat{h}_j^i})^2] - \\ & \sum_{i=0}^{S^2} \sum_{j=0}^{B} I_{ij}^{obj} [\hat{C}_i^j \log(C_i^j) + (1 - \hat{C}_i^j) \log(1 - C_i^j)] - \\ & \lambda_{noobj} \sum_{i=0}^{S^2} \sum_{j=0}^{B} I_{ij}^{noobj} [\hat{C}_i^j \log(C_i^j) + (1 - \hat{C}_i^j) \log(1 - C_i^j)] - \\ & \sum_{i=0}^{S^2} I_{ij}^{obj} \sum_{c \in classes} [\hat{P}_i^j \log(P_i^j) + (1 - \hat{P}_i^j) \log(1 - P_i^j)] \end{aligned} \tag{3}$$

In the above formula, I_{ij}^{obj} is used to determine whether the j anchor box in the i grid is responsible for the current object. If $I_{ij}^{obj} = 1$ is responsible, if $I_{ij}^{obj} = 0$ is not. $\hat{C}_i^j$ parameter is the confidence degree, $\hat{C}_i^j$ is the real value in the training process. The value of this parameter depends on whether the bounding box of grid cell is responsible for the prediction of an object. If it is responsible for $\hat{C}_i^j = 1$, otherwise, $\hat{C}_i^j = 0$.

Central coordinate error:

$$\sum_{i=0}^{S^2} \sum_{j=0}^{B} I_{ij}^{obj} [(x_i - \hat{x}_i^j)^2 + (y_i - \hat{y}_i^j)^2] \tag{4}$$

The formula of wide coordinate error in calculation:

$$\sum_{i=0}^{S^2}\sum_{j=0}^{B} I_{ij}^{obj}[(\sqrt{w_j^i}-\sqrt{\hat{w}_j^i})^2+(\sqrt{h_j^i}-\sqrt{\hat{h}_j^i})^2] \quad (5)$$

Confidence error:

$$\sum_{i=0}^{S^2}\sum_{j=0}^{B} I_{ij}^{obj}[\hat{C}_i^j\log(C_i^j)+(1-\hat{C}_i^j)\log(1-C_i^j)]- \quad (6)$$

$$\lambda_{noobj}\sum_{i=0}^{S^2}\sum_{j=0}^{B} I_{ij}^{noobj}[\hat{C}_i^j\log(C_i^j)+(1-\hat{C}_i^j)\log(1-C_i^j)]$$

3. Experimental results and analysis before improving the model

3.1. Experimental process

The target detection process is as follows:

(1) The experimental data set is preprocessed, and the large-scale remote sensing image is clipped;

(2) Determine whether the remote sensing image is a single channel image, and select the appropriate defogging processing method;

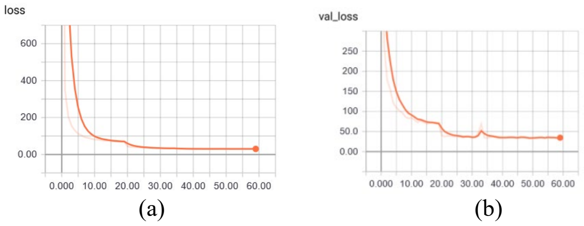
(3) Remove the objects that are not interested in the remote sensing data set, and keep the whole data set only contain the two kinds of objects of interest in this paper;

(4) The processed remote sensing data are input into the network model for training, and the training model is obtained;

(5) After training, the target is detected and classified.

3.2. Analysis of experimental results

The changes of loss value and val_loss value during training are shown in Figure 3 below:



(a)

(b)

Fig. 3. Changes in loss during training

According to the change of loss value in Figure 3 (a), it can be seen that the loss value changes greatly in the first stage, and gradually decreases and tends to be stable in the second stage. According to the change of loss of verification set in Figure 3 (b), it can be seen that the change of loss value tends to decrease during the training process.

After training, you will get trained_weights_final.h5 file is used to test the test set, set the IOU value to 0.5, and then get the final evaluation result of the training result according to the test results. Since the data set used in this experiment is to integrate part of the data in NWPU VHR-10 remote sensing data set and part of the data in DOTA-v1.5 remote sensing data set, and only contains two types of targets, the detection results are shown in Figure 4, and the test evaluation is shown in table 1:

Table 1. Evaluation of test set experiment results

Target category	**Precision**	**Recall**	**AP**
plane	0.64	0.93	83.83%
ship	0.71	0.43	40.02%

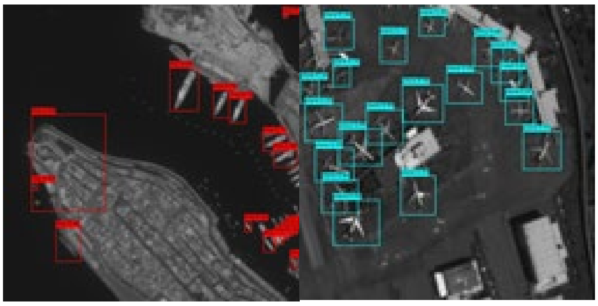


Fig. 4. Test result display diagram

According to the above table 1, the AP value of aircraft is 83.83%, the AP value of ship is 40.02%, and the mAP of two types is 61.93%. From the recall rate, we can see that the recall value for ships is smaller than for aircraft, but the detection accuracy for these two kinds of targets is not high. From the above table, it is not difficult to see that the detection effect is better for large targets such as aircraft, but poor for targets with small ship scale.

4. Experimental results and analysis after improving the model

In this paper, we consider the scale problem to increase the size of the feature map to retain more semantic information on the resolution. By increasing the scale of the input image, more semantic information can be retained. In this paper, the size of the input image is modified to 608 × 608 and 672 × 672.

The prior frame was originally proposed because of Faster R-CNN[9]. According to the size of the ground_truth mark box in the training set, the width and height of the frequently appeared label box are counted, and these mark boxes are taken as the prior box. In order to make the prior frame suitable for the remote sensing data set used in this experiment, therefore, K-means clustering operations are carried out on the dataset for several times.

We use ten times clustering to get ten groups of prior frames, and get the most suitable prior frame by means of average value. As shown in Table 2 below:

Table 2. A priori box improvement

Characteristic map	13*13	26*26	52*52
Perception vision	large	medium	small
Initial prior box	(116x90) (156x198) (373x326)	(30x61) (62x45) (59x119)	(10x13) (16x30) (33x23)
Clustering mean prior frame	(49x37) (65x67) (125x123)	(31x40) (35x26) (43x30)	(14x27) (21x23) (24x23)

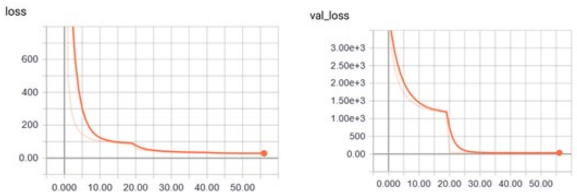
For IOU, IOU is the intersection ratio of detection frame and real frame. In case of non-intersecting border, the gradient will become 0 and cannot be optimized. But for GIOU, this situation can be avoided, so GIOU will be used instead of IOU in this paper. The formula is as follows (7):

$$GIOU = IOU - \frac{|C \setminus (A \cup B)|}{|C|} \tag{7}$$

The loss function of GIOU is shown in formula (8):

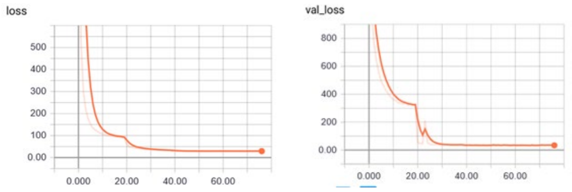
$$L_{GIOU} = 1 - GIOU \tag{8}$$

The evaluation index of the improved model is the same as the evaluation index in the previous section. After the improvement of the model, the loss value of the training set constructed by the partial data of NWPU VHR-10 data set and DOTA-v1.5 data set is constantly decreasing.



(a) (b)

Fig. 5. Change of loss value of input image scale 608×608 after network improvement



(a) (b)

Fig. 6. Change of loss value of input image scale 672×672after network improvement

As shown in Figure 6, the change trend of the loss value in the network training process is basically consistent. Compared with the input image 608 × 608, the change of the loss value of the verification set with the input image scale of 672 × 672 fluctuates, but the overall trend is that the loss value is gradually reduced.

Combined with the change of prior box, the model is changed to input scale of 672 × 672 and 608 × 608. The number of layers is changed from 52 layers to 56 layers due to the addition of 1 to the layers repeated 8 times in convolution layer. With the increase of input scale and the deepening of network layers, not only more semantic information is retained, but also more semantic information can be extracted in feature extraction.

Table 3. Comparison of improved network experiment results mAP

network model	Ship AP	Plane AP	mAP
Yolov3	38.60%	86.77%	61.93%
Yolov3_608	42.85%	86.48%	64.69%
Yolov3_672	40.07%	88.94%	64.84%
Yolov3_608_GIOU	40.02%	83.83%	61.93%
Yolov3	38.60%	86.77%	61.93%

Table 4. The improved network Precision and Recall comparison

network model	Ship Precision	Ship Recall	Plane Precision	Plane Recall
Yolov3	0.814	0.411	0.85	0.92
Yolov3_608	0.82	0.448	0.865	0.918
Yolov3_672	0.806	0.433	0.838	0.939
Yolov3_608_GIOU	0.711	0.438	0.648	0.935

The Yolov3_608_GIOU and Yolov3_672_GIOU versions in the above table refer to the replacement of IOU with GIOU based on the Yolov3_608 and Yolov3_672 versions. Due to the poor detection effect of Yolov3 on small targets, the recall rate is not very ideal. It can be seen from the above table that the recall rate of ships with more small targets has been improved in several improved versions, and the precision will not be greatly reduced. According to the analysis of the mAP value, only the mAP value of the improved network Yolov3_608_GIOU is lower than that of the initial network by 0.75%, and the detection effect of the improved network Yolov3_672_GIOU is the best, and the map value is 3.26% higher than that of the initial network.

5. Conclusion

This chapter first introduces the characteristic pyramid and the loss function of Yolov3, and then optimizes the Yolov3 network. Firstly, K-means clustering method is used to cluster according to the training set, and then the average value is used as the adjustment of prior frame. In order to retain more semantic information, adjust the scale of input data, extract more semantic information and increase the depth of the network, Yolov3 is optimized according to the data set in this paper. Finally, IOU is replaced by GIOU. Compared with several improved networks, the detection effect of the improved network on the test set is improved. Finally, the detection effect of the improved network Yolov3_672_GIOU is the best, and the mAP value has been significantly improved.

References

1. Redmon J, Divvala S, Girshick R, et al. You Only Look Once: Unified, Real-Time Object Detection [M]. 2016 IEEE Conference on Computer Vision and Pattern Recognition. 2016:779-88.
2. Felzenszwalb P F, Girshick R B, Mcallester D, et al. Object Detection with DiscriminativelyTrained Part-Based Models[J]. IEEE Transactions on Software Engineering, 2010, 32(9):1627-1645.
3. Chen L-C, Zhu Y, Papandreou G, et al. Encoder-decoder with atrous separable convolutionfor semantic image segmentation[C]. Proceedings of the European Conference on ComputerVision (ECCV). 2018: 801 – 818.
4. Liu Chang, Wang Pengjun, Zhang Meiling, et al Research on sparse video detection technology based on IOU analysis [J]. High technology letters, 2019(10).
5. Zhang R, Yao J, Zhang K, et al. S-cnn-based ship detection from high-resolution remote sensing images. [J]. International Archives of the Photogrammetry, Remote Sensing & Spatial Information Sciences, 2016, 41.
6. Walk S, Majer N, Schindler K, et al. New features and insights for pedestrian detection[C].IEEE Conference on Computer Vision and Pattern Recognition, 2010: 1030-1037.
7. Ok A O, Başeski E. Circular oil tank detection from panchromatic satellite images: a new automated approach[J]. IEEE Geoscience and Remote Sensing Letters, 2015, 12(6): 1347 –1351.
8. Li Weijun. Overview of K-means clustering algorithm [J]. Modern computer (Professional Edition), 2014 (8): 31-32
9. M. Kang, X. Leng, Z. Lin, et al. A modified faster R-CNN based on CFAR algorithm for SAR ship detection[C]. 2017 International Workshop on Remote Sensing with Intelligent Processing (RSIP), Shanghai, China, 2017, 1-4.