Airport and Ship Target Detection on Satellite Images Based on YOLO V3 Network

Ren Ying

Abstract Airplane and ship play a very important role in both civil life and military operations. It is a meaningful to detect airplane and ship around the world through remote sensing images. Target recognition algorithms based on deep learning technology are proven to be effective and gradually replacing traditional algorithms. This paper builds a target detection system based on NVIDIA TX2 development platform and YOLO V3 algorithm and focuses on both the ship and airplane targets. The training data comes from image fragments generated by satellites such as Jilin No. 1, DigitalGlobe, and Planet. The label of each target includes a bounding box and category information. The image processing method such as rotation and noise is added to increase the robustness of the trained YOLO V3 network for different sensors and atmosphere. The training of the input image takes about 3 days on an NVIDIA Titan X GPU. At test time, we partition testing images of arbitrary size into cutouts with a fixed size of 1 k × 1 k and run each cutout through our trained model to find ships and airplanes. The experimental results show that the F_1-score values of the airport and the ship are 91.48% and 93.89%, respectively, and the detection speed of one cutout on NVIDIA TX2 development platform is about 0.56 s.

Keywords Target detection · Deep learning · YOLO V3 · Satellite images · Training data

1 Introduction

In recent years, with the development of related technologies such as aerospace, remote sensing, and sensors, the data and information contained in remote sensing images are becoming more and more abundant, which also facilitates the analysis of remote sensing images by researchers. Airplane and ship play a very important role in both civil life and military operations. It is an important technology to identify airplane and ship around the world through remote sensing images.

R. Ying (✉)
Chang Guang Satellite Technology Co., Ltd., Changchun, China
e-mail: renying1009@163.com

L. Wang et al. (eds.), *Proceedings of the 6th China High Resolution Earth Observation Conference (CHREOC 2019)*, Lecture Notes in Electrical Engineering 657,
https://doi.org/10.1007/978-981-15-3947-3_12

Remote sensing image target recognition based on traditional algorithms mainly relies on manual interpretation, which is low in efficiency, high in cost, and poor in timeliness. The traditional model can no longer meet the current operational needs. Fortunately, a variety of target recognition algorithms based on deep learning technology are gradually replacing traditional algorithms. With the introduction of networks such as deep residual networks and deep dense networks, the number of layers of deep convolutional neural networks is getting deeper, and the over-fitting phenomenon brought about by the deepening of the network is greatly reduced, and the recognition effect is also increasing accurate.

At present, the methods of target recognition of convolutional neural networks can be divided into two categories. The first category is region-based target recognition methods such as Faster R-CNN [1] and Mask R-CNN [2]. This type of method works well for small targets, but the detection speed is slow. The other type is regression-based target recognition methods, such as SSD [3] and YOLO [4]. The regression-based target recognition method uses end-to-end target recognition, and the speed is much faster than the region-based target recognition method.

In this paper, an on-board target detection system based on NVIDIA TX2 development platform and YOLO V3 [5] algorithm is built, which is focused on the detection of airplanes and ships. Satellite images with the two targets are collected and labeled. The experimental results show that the F_1-score values of the well-trained network on airports and ships are 91.48% and 93.89%, respectively, and the detect speed of a signal image with a size of 1 k $\times$ 1 k on NVIDIA TX2 development platform is about 0.56 s.

2 Annotation of Dataset

2.1 Motivation

Datasets play an important role in data-driven technology, such as deep learning. There are many conventional target detection datasets, and the front-end target detection algorithms (such as Faster R-CNN, Mask R-CNN, SSD, and YOLO) are basically experimented on these regular datasets. But the classification training based on conventional datasets performs poorly on satellite images, because the satellite images have its particularity, such as scale diversity, particular perspective, and small target problem.

Satellite images are of different orbital altitudes from a few hundred kilometers to thousands of kilometers, and the ground targets are even different in size, such as the aircraft carrier is more than 300 m, and the small ship is only a few 3 m. The perspective of the satellite image is basically a high-altitude view, but the conventional datasets are mostly the horizontal perspective, so the same target mode is different. Even a good detector trained on the regular dataset may have a poor performance on detecting targets of satellite images. Many targets of satellite images are with small

size (tens or even a few pixels), which leads to a small amount of target information. The CNN-based target detection method has a good performance on the conventional target detection dataset, but for small targets, CNN's pooling layer will further reduce the amount of information. A 24 × 24 target has only about 1 pixel after 4 layers of pooling, making the dimension too low to distinguish.

Based on the above reasons, for the target detection task of satellite images, the conventional dataset is often difficult to train the ideal target detector, and a special satellite database is needed.

2.2 Image Collection and Annotation

As mentioned before, two categories are selected and annotated in our dataset, i.e., airplane and ship. In target detection of satellite images, in addition to resolutions of satellite images, the variety of sensors are also effective to the performance of classifier. To ensure that the classifier of targets on satellite images is robust to different resolutions and sensors, and satellite images in our dataset are collected from multiple satellites with multiple resolutions. Training data is collected from small chips of large images mainly from three satellites, i.e., Jilin No. 1, DigitalGlobe, and Planet. To increase the diversity of training data, the satellite images are from different cities. In addition, the dataset is expanded by some image processing operations, such as image noise, rotation, and contrast change.

As for dataset labeling, many visual concepts such as region descriptions, and objects, can be annotated with bounding boxes, as shown in Fig. 1. A common description of bounding boxes is (x_c, y_c, w, h), where (x_c, y_c) is the center location, w and h are the width and height of the bounding box, respectively. That is to say, each label includes a bounding box and category identifier for each object.

Compared with those in regular images dataset (e.g., PASCALVOC and MSCOCO), these satellite images are mostly very large in size. The original size of images in our dataset ranges from about 800 × 800 to about 4 k × 4 k. We make annotations on the original full image without partitioning it into pieces to avoid the cases where a single instance is partitioned into different pieces. And then, the original images are partitioned into pieces with sizes smaller than 1 k × 1 k. The cross-dataset generalization is an evaluation for the generalization ability of a dataset. We randomly select 14,662 for training and 3258 for testing. Our final dataset and the train/test split for each category are shown in Table 1.

3 Network Architecture

You Only Look Once (YOLO) algorithm is a regression-based target recognition method, which is proposed in 2015 [6]. The third generation of YOLO V3 has been developed in 2018. Just like its name, it only needs to do a forward calculation to make

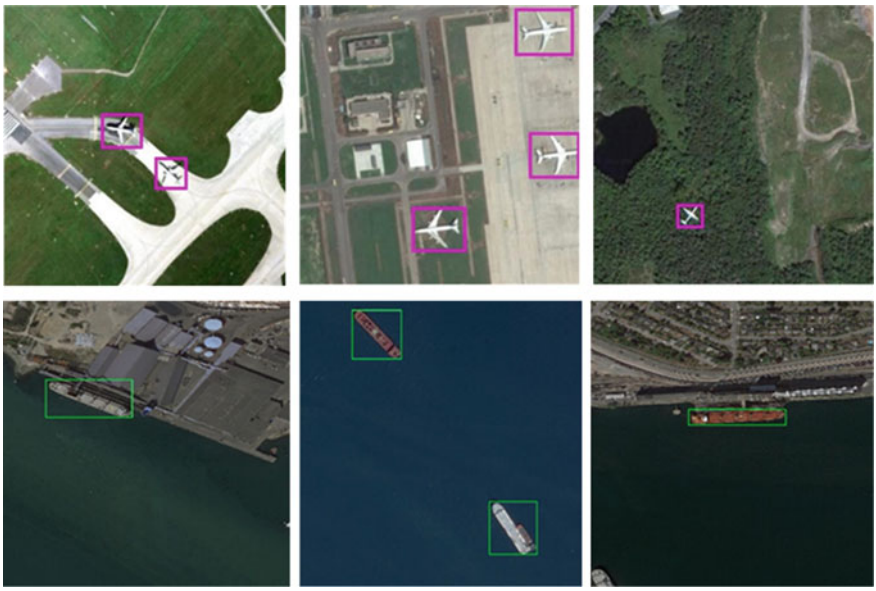

Fig. 1 Samples of annotated images in our dataset

Table 1 Train/test split

Object class	Training examples	Test examples
Airplanes	6686	1486
Ships	7976	1772
All	14,662	3258

a variety of objects are detected, so the YOLO series algorithm detects quickly. The network still maintains the advantages of the fast detection of the YOLO V2 network, and the correct rate of recognition is greatly improved. Especially in the detection and identification of small targets, the accuracy is greatly improved. YOLO V3 network draws on the idea of residual neural network, introducing multiple residual network modules and using multi-scale prediction improves the defect of YOLO V2 network in small target recognition. Because of the high accuracy and timeliness of detection, this algorithm is one of the best algorithms in the field of target detection. The model uses a number of well-formed 3×3 and 1×1 convolutional layer and later uses multi-scale predictions to structure some residual networks. Finally, it has 53 convolutional layers and is called as Darknet-53.

The YOLO V3network introduces the idea of using anchor boxes in Faster R-CNN. Three scales are used for COCO datasets and VOC datasets. Each scale has 3 anchor boxes, and the scale features are large. A small a priori box, so you can select the appropriate a priori box anchors according to the target you want to identify and modify the network structure according to the scale of the prediction.

The YOLO V3 network uses anchor boxes as a priori boxes to detect targets in the image. In the Faster R-CNN and SSD, you need to manually set the a priori box, which will make the selection subjective. If you can choose a suitable a priori box, the deep convolutional neural network will be easier to learn. Hence, the K-means algorithm is used in the YOLO V3 network to cluster the target frame size of our dataset.

4 Evaluations

4.1 Network Training

We train with stochastic gradient descent. Besides the anchor boxes, the main parameters in the training are set as follows: each batch of 32 images, number of iterations of 50,200, an initial learning rate of 10^{-3}, a weight decay of 0.0005, and a momentum of 0.9. Meanwhile, rotating training dataset is used during training to increase the contrast and exposure of the image. Each iteration trains 1 batch and performs 1 scale transformation to realize the expansion of the dataset. In addition, by doing k-means clustering on our dataset, the size of the corresponding prediction box is set to the center of the nine clusters, which are (44, 65), (63, 37), (135, 189), (124, 85), (76, 67), (217, 103), (76, 130), (115, 50), (243, 219). Training takes 3 days on a single NVIDIA Titan X GPU.

4.2 Test Results

If the number of iterations of the training network exceeds a certain number, the phenomenon of over-fitting may occur. In order to select the best weight file, the network weight file is saved every 10,000 iterations during training. The first 500 images of the test images are selected to test all the weights using the network and calculate the recall rate, which is expressed by

$$R = T_{\mathrm{P}}/(T_{\mathrm{P}} + F_{\mathrm{N}}) \tag{1}$$

where T_{P} denotes the number of targets that are correctly detected and F_{N} denotes the number of targets that are not detected.

Part of the test results is shown in Table 2. It can be seen from Table 2, the optimal iteration time for the YOLO V3 network is around 50,000.

We select the weight file with 50,000 iteration times to evaluate the well-trained network on our dataset and calculate the accuracy rate, which is expressed by

$$P = T_{\mathrm{P}}/(T_{\mathrm{P}} + F_{\mathrm{P}}) \tag{2}$$

Table 2 Relationship between recalling rate and iteration times

Iteration times	Average recalling rate (%)
10,000	88.13
20,000	89.61
30,000	90.19
40,000	91.62
50,000	**92.02**
50,200	91.89

Bold represents the maximun value at the iteration number.

Table 3 Relationship between recalling rate and iteration times

Object class	Accuracy rate (%)	F_1-score
Airplane	92.12	91.48
Ship	94.57	93.89

where T_P denotes the number of targets that are correctly detected and F_P denotes number of targets that are miss detected.

In this paper, F_1-score, i.e., balanced F score, is used to evaluate the target detection effect of our well-trained network on satellite images, which can be expressed by

$$F_1 = 2^* \frac{P * R}{P + R} \tag{3}$$

As shown in Table 3, the F_1-score of the well-trained network on airplane and ship is 91.48% and 93.89%, respectively.

Then the well-trained network is transplanted to NVIDIA TX2 development platform to evaluate the detection speed. At test time, we partition testing images of arbitrary size into manageable cutouts and run each cutout through our trained model. Partitioning takes place via a sliding window with user-designed bin sizes and overlap (30% by default). As shown in Table 3, the average detection time of a single image with a size of 1 k × 1 k on NVIDIA TX2 development platform is 0.56 s. The visualization results are shown in Figs. 2 and 3.

5 Conclusion

In this paper, an on-board target detection system based on NVIDIA TX2 development platform and YOLO V3 algorithm is built, which is focused on the detection of airplanes and ships. Satellite images with the two targets are collected and labeled. The experimental results show that the F_1-score values of the well-trained network on airports and ships are 91.48% and 93.89%, respectively, and the detect speed of

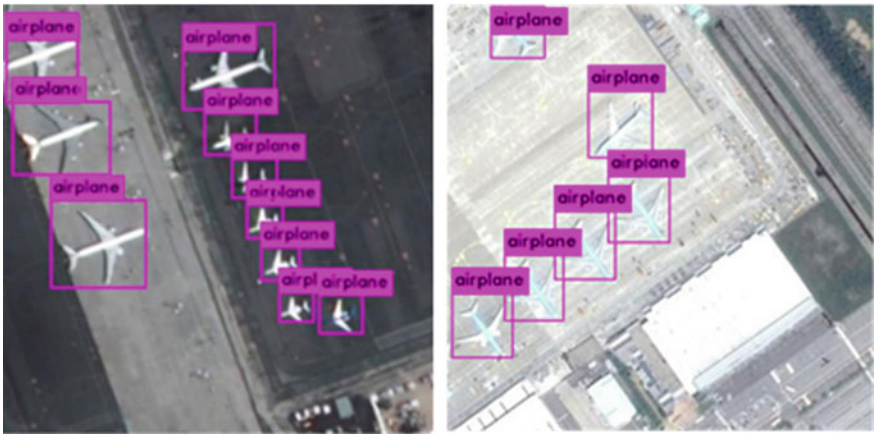

Fig. 2 Visualization results of airplane detection

Fig. 3 Visualization results of ship detection

a signal image with a size of 1 k × 1 k on NVIDIA TX2 development platform is about 0.56 s.

References

1. Girshick R (2015) Fast R-CNN. In: IEEE conference on computer vision and pattern recognition, pp 1440–1448
2. He KM, Gkioxari G, Dollar P et al (2018) Mask R-CNN. In: IEEE conference on computer vision and pattern recognition, pp 1–12

3. Liu W, Anguelov D, Erhan D, et al (2016) SSD: signal shot multibox detector. In: IEEE conference on computer vision and pattern recognition, pp 21–37
4. Redmon J, Farhadi A (2018) YOLO V3: an incremental improvement. In: IEEE conference on computer vision and pattern recognition, pp 1–6
5. Redmon J, Farhadi A (2017) YOLO9000: better, faster, stronger. In: IEEE conference on computer vision and pattern recognition, pp 6517–6525
6. Redmon J, Divvala S, Girshick R, Farhadi A (2015) You only look once: unified, real-time object detection. arXiv:1506.02640