Aircraft Detection in Remote Sensing Images Based on Deep Convolutional Neural Network

Yibo Li
Shenyang Aerospace University
Shenyang, China
liyibo_sau@163.com

Jingfei Zhao
Shenyang Aerospace University
Shenyang, China
me_fan@yeah.net

Senyue Zhang
Nanjing University of Aeronautics
Nanjing, Chins
xuer27@163.com

Wenan Tan
Nanjing University of Aeronautics
Nanjing, Chins
watan@sspu.edu.cn

***Abstract*—Aircraft detection in remote sensing images is always the research hotspot but a challenging task for the variations of aircraft type, pose, size and complex background. The paper proposes a region-based convolutional neural network to detect aircrafts. To enhance the learning ability of the network, a multi-resolution aircraft remote sensing dataset is collected from Google Earth. Then, the detection model is trained end to end by fine-tuning on the obtained dataset and realizes automatic aircraft recognition and positioning. Experiments show that the proposed method outperforms state-of-the-art method on the same dataset and the requirement for real-time can be satisfied simultaneously.**

Keywords—aircraft detection, remote sensing image, convolutional neural network

I. INTRODUCTION

Automatic aircraft detection in remote sensing images has been one of the research focuses due to its high application value in airport dynamic monitoring and military surveillance. With the development of computer vision and image processing technology, different detection methods have been proposed in recent years.

Traditional aircraft detection methods usually consist of three separated stages: region proposal, feature extraction and target classification. The first stage chooses some candidate regions on the given images for further recognition. Sliding window [1] is usually used for object locating. Because millions of windows per image are sent into network to compute gradient, it is rather time-consuming. To solve the problem, some approaches, such as selective search [2], binarized normed gradients (BING) [3], Edge boxes [4] and objectness [5], have been proposed. However, the time consumption of region proposals is still considerable and the process is hardly realizable through GPU. The extracted features from intermediate stage are used to classify and recognize the areas through retrained classifier at last. Conventional methods always apply advanced general features, such as histogram of oriented gradient (HOG) [6] and scale-invariant feature transform (SIFT) [7], to specific target application, but general features have difficulty in distinguishing different types and holding the invariance of target. Researchers also design templates with rotation and scale invariant or corresponding manual features based on the characteristics of the aircraft to detect specific aircraft target [8-9]. Nevertheless, experiments show that those methods can not keep high accuracy faced with new complex scene.

In the paper, we propose a novel detection model based on deep convolutional neural network to detect aircrafts and realize a better performance than the state-of-art method. The method achieves 94.55% detection accuracy that way above the existing methods and the detection rate is about 3fps.

II. REALATION WORK

The CNN [10], first proposed by Yann et al, is a supervised deep learning network and the convolutional layers in the network can learn rich features from the raw data. With the appearance of innovative network models, the CNN has yielded many state-of-the-art performances in the field of image processing. CNN has been used to the aircraft detection in remote sensing images. A FCN model was proposed by Xu et al. [11] for aircraft detection. Wu et al. [12] integrated BING and CNN to recognize aircrafts. Chen et al. [13] introduced multiple thresholds for localization and used CNN for classification.

Object detection based on CNN has achieved major breakthrough in the application of natural images from 2014 and the detection networks can be divided into two types: region-based detection networks and end-to-end detection networks. Faster R-CNN [14], a variant of R-CNN [15] and Fast R-CNN [16], is a typical region-based convolutional neural network. It makes real-time detection to become possibility trough GPU where proposal computation is nearly cost-free. The detection method achieves state-of-the-art object detection accuracy, 73.2% and 70.4% mean Average Precision (mAP) on nature images dataset PASCAL VOC

978-1-5386-7141-2/18/$31.00 ©2018 IEEE

2007 and PASCAL VOC 2012. SSD [17] and YOLO [18] are the representative models of end-to-end detection methods and have advantage in detection speed and achieve a competitive accuracy with Faster R-CNN. However, they are not suitable to detect small targets for the strategy of target localization and the demand of fixed-size images as input. Inspired by Faster-RCNN, we detect aircrafts by modifying the parameters of the model according to the characteristics of our dataset.

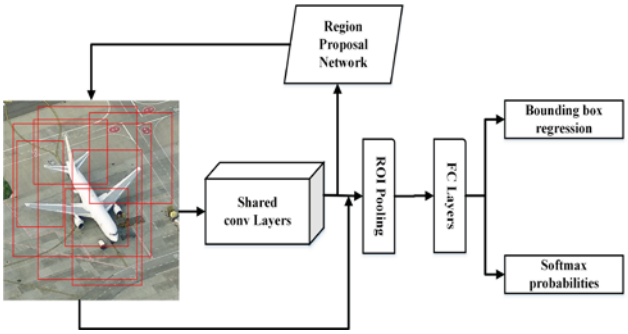

Figure 1. The architecture of Faster R-CNN

The rest of the paper is organized as follows. In section 3, the network model is introduced. Simulation experiment and conclusion respectively arranged in section 4 and section 5.

III. Network Model

A. Model Struction

Faster R-CNN is a region-based object detection network and consists of a Region Proposal Network (RPN) and a state-of-the-art object detection network Fast R-CNN. The architecture of Faster R-CNN is shown in Fig.1. (RPN) shares fully convolutional layers with the object detection network. Among them, the RPN takes an image of arbitrary size as input and output a set of object proposals, each with 4 coordinates (x, y, w and h) of the predicted bounding box and 2 scores to estimate the probability of object/not-object. The cross-boundary anchors are ignored to increase the detection accuracy and the threshold for non-maximum suppression (NMS) is fixed at 0.7 to reduce redundancy caused by the overlapped RPN proposals. Object detection network perform elaborate classification and localization task by calculating softmax probabilities and bounding-box regression offsets for each proposal.

TABLE I. Anchor Boxes Sizes At Each Sliding-window Location

Anchor	$64^2,2:1$	$64^2,1:1$	$64^2,1:2$	$128^2,2:1$
Proposal	93×55	57×60	43×83	188×111
Anchor	$128^2,1:1$	$128^2,1:2$	$256^2,2:1$	$256^2,1:1$
Proposal	113×114	70×90	416×229	261×284
Anchor	$256^2,1:2$	$512^2,2:1$	$512^2,1:1$	$512^2,1:2$
Proposal	174×332	768×437	499×501	355×715

The mechanism of region proposal method is that sliding a 3*3 spatial window on the feature map produced by the last shared convolutional layer. At each sliding-window location, it outputs 9 anchor boxes, 3 scales and 3 aspect ratios, corresponding to 9 region proposals in the raw image. Each sliding window is mapped to a lower-dimensional vector that will be sent into parallel fully connected layers, box-regression layer and box-classification layer. Considering that the aircraft targets in remote sensing image are smaller than the objects in natural images, we extend the anchor boxes to 12 with a smaller box area of 64^2 pixels at the same aspect ratios in the paper. The experimental results also prove the effectiveness. The concrete sizes are shown in Table I.

B. Training Strategy

To training the unified detection network, we assign a binary label to each anchor in RPN. We set a positive label to an anchor when the rules are met: 1) the anchor has the highest Intersection-over-Union (IoU) overlap with a ground-truth box, or 2) the anchor has an IoU overlap that higher than 0.75 with any ground-truth box. Meanwhile, we set a negative label for the anchors that IoU is lower than 0.3 with all the ground-truth boxes. Anchors that are neither positive nor negative do not serve as training samples. The sampling strategy used in RPN follows [16]. A mini-batch comes from one image that contains many positive and negative anchors. To avoid the bias towards negative samples, we randomly sample 256 anchor boxes in an image and the ratio of positive and negative anchors is 1:1.

For each anchor box, we adopt a multi-task loss function for classification and bounding-box regression as used in [14]:

$$L(\mathrm{p}_i,\mathrm{t}_i)=\mathrm{L}_{cls}(\mathrm{p}_i,\mathrm{p}_i^*)+\lambda\,\mathrm{p}_i^*\,L_{reg}(\mathrm{t}_i,\mathrm{t}_i^*) \tag{1}$$

Where the classification loss L_{cls} is log loss and the regression loss L_{reg} is smooth L_1 loss function defined in [16]. Parameter p_i is the predicted probability that anchor i is an aircraft. p_i^* is labeled 1 when the anchor is positive and 0 if the anchor is negative. The term $p_i^* L_{reg}$ indicates that the regression loss is activated only by positive anchors. For locating, each anchor box has 4 coordinates and the parameterizations of coordinates are defined following [15]:

$$\begin{aligned}
&t_x=(\mathrm{x}-\mathrm{x}_r)/\mathrm{w}_r, t_y=(\mathrm{y}-\mathrm{y}_r)/\mathrm{h}_r\\
&t_w=\log(\mathrm{w}/\mathrm{w}_r), \mathrm{t}_h=\log(\mathrm{h}/\mathrm{w}_r)\\
&t_x^*=(\mathrm{x}^*-\mathrm{x}_r)/\mathrm{w}_r, t_y^*=(\mathrm{y}^*-\mathrm{y}_r)/\mathrm{h}_r\\
&t_w^*=\log(\mathrm{w}^*/\mathrm{w}_r), \mathrm{t}_h^*=\log(\mathrm{h}^*/\mathrm{w}_r)
\end{aligned} \tag{2}$$

x, y, w and h denote the center coordinates of the anchor box, its width and height. x, x_r and x^* are for the predicted box, anchor box and ground-truth box (same to y, w and h). The hyperparameter λ is set to 10 to control the balance between the two task losses. With those definitions, the detection network is possible to predict aircraft targets at a wide range of scales and aspect ratios by minimizing the objective function.

IV. Simulation Experiment

A. Dateset

The remote sensing aircraft dataset used in [11] has 76 images in total, too small to train Faster R-CNN. In the paper, we collect 1000 images with different resolution

from Google Earth. Fig. 2 shows some examples of the dataset and the percentage of high-resolution images is about 90%. The image sizes vary from 600*700 to 740*1380 pixels. Research shows that data augmentation can make the network fully study the change of the object and enhance the recognition ability for the complex change in translation and angle. We expand the available dataset with horizontal flip and rotation $(90^{\circ},180^{\circ},270^{\circ})$ to the seed images. With the approach, the dataset is expanded by 8* from 1000 to 8000 images. We select randomly 3000 images for training, 3000 images for validation and 2000 images to test. The network we propose is a supervised object detection network and the other two kinds of documents we need to use the detection network. XML format file is used to indicate the ground-truth locations of aircrafts within the images. We adopt the image annotation tool LabelImg to complete the job manually and only one category in the dataset. Apart from annotation files, the code framework requires 4 TXT files named train.txt, test.txt, val.txt and trainval.txt to indicate the use of those images.

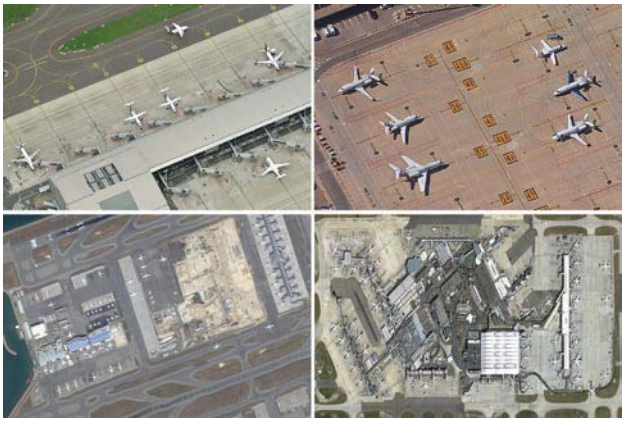

Figure 2. The examples of the image dataset.

B. Network Training

CNN requires a lot of samples to initialize the training of the networks and the number of our dataset is still small compared with millions of datasets. Transfer learning [19] has been successfully applied to small sample training. The training strategy can speed up the convergence rate and avoid local optimum. We use pre-trained VGG-16 [20] model and the approximate joint training which can be trained end-to-end to fine-tune the detection network. The main parameters of the network are shown in table 2. All experiments are done within Caffe and our graphics is NVIDIA GeForce GTX 1070 with an 8GB GDDR5 memory.

TABLE II. NETWORK PARAMETER CONVIGURATION

Parameter	Base_lr	Lr_policy	gamma
Value	0.001	Step	0.1
Parameter	Weight_decay	Momentum	Iter_size
Value	0.0005	0.9	2

C. Rusults and Analysis

Considering that the number of proposals fed into Fast R-CNN has influence on the detection accuracy at test time [14, 21], we analyze the effect of this parameter on the detection performance at first. As showed in Fig. 3, the maximum is achieved when the number of proposals is fixed at 500.

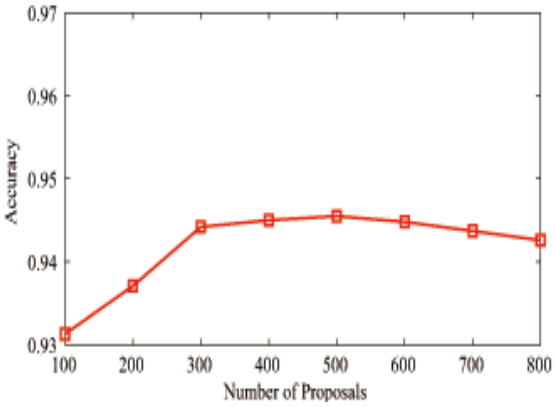

Figure 3. The performance with different proposals number

To evaluate the detection method quantitatively, we test the proposed model and the compared methods on the sane test dataset we collected. The compared methods include the original Faster R-CNN and FCN which achieves the best results in [11]. The detection rate and the average test time are shown in Table 3.

TABLE III. THE DETECION RATE AND AVARAGE DETECTION TIME OF THREE METHODS

Model	**Detection Rate**	**Average Detection time**
FCN	87.41%	4fps
Faster R-CNN	92.76%	3fps
The proposed method	94.55%	3fps

FCN is an end-to-end detection network which can be approximately considered as the RPN part of the Faster-RCNN. We can clearly see that Faster R-CNN and the proposed method have better performance than FCN. The results show that the high-quality region proposals produced by RPN can greatly improve the detection capacity of the network. The proposed method has 1.8 percentages higher than the original Faster R-CNN indicating that the smaller box areas are suitable to the detection of the small targets in large range. The running time is about 3fps on GTX 1070 slower than FCN because of the network structure and is enough fast to meet requirements for real-time.

Fig. 4 shows the detection results of the proposed method on partial test images from which it can be seen that the proposed method can accurately detect multi-scale and multi-direction aircraft targets in complex background.

Figure 4. Aircraft detection results of the proposed method with a score threshold as 0.8

V. Conclusion

The paper proposes a region-based convolutional neural network using smaller anchor boxes area in Faster R-CNN to detect aircrafts in remote sensing images. Experiments show that this method has efficient detection and performs better than the state-of-art method. Aircrafts that are various types and different colors are treated as the same kind of target in the paper. In real application, it always needs to distinguish the type and detect moving target. In view of the results, the future work will be devoted to improving the detection precision and applying it to other target detection tasks.

References

[1] N. Dala and B. Trikks, "Histograms of oriented gradients for human detection," Proc. IEEE Comput. Soc. Conf. Comput. Vision Pattern Recognition, Jun.2005, pp. 886-893, doi:10.1109/CVPR.2005.177.

[2] J. R. R. Uijlings, K. E. A. Van De Sande, T. Gevers, and A. W. M. Smeulders, "Selective search for object recognition," Int. J. Computer. Vis, vol. 104, Sep. 2013, pp. 154-171, doi:10.1007/s11263-013-0620-5.

[3] M.-M. Cheng, Z. Zhang, and W.-Y. Lin, "BING: Binarized normed gradients for objectness estimation at 300fps," Proc. IEEE. Comput. Soc. Conf. Comput Vision Pattern Recognit, Sep. 2014, pp. 3286-3293, doi:10.1109/CVPR.2014.414.

[4] C. L. Zitnick and P. Dollar, "Edge boxes: Locating object proposals from edges," Proc. 10th Eur. Conf. Comput. Vis, Sep. 2014, pp. 391-405, doi:10.1007/978-3-319-10602-1_26.

[5] B. Alexe, T. Deselaers, and V. Ferrari, "Measuring the objectness of image windows," IEEE Trans. Pattern. Anal. Mach. Intell, vol. 34, Jan. 2012, pp. 2189-2202, doi:10.1109/TPAMI.2012.28.

[6] H. Sun, X. Sun, H. Wang, Y. Liu, and X.-J. Li, "Automatic target detection in high-resolution remote sensing images using spatial sparse coding bag-of-words model," IEEE Geosci. Remote Sens. Lett, vol. 9, Aug. 2011, pp. 109-113, doi:10.1109/LGRS.2011.2161569.

[7] X. Bai, H. Zhang, and J. Zhou, "VHR object detection based on structural feature extraction and query expansion," IEEE Trans. Geosci. Remote Sens, vol. 52, Oct. 2014, pp. 6508-6520, doi:10.1109/TGRS.2013.2296782.

[8] J. Inglada and J. Michel, "Qualitative spatial reasoning for high-resolution remote sensing iamge analysis," IEEE Trans. Geosci. Remote Sens, vol. 47, Nov. 2008, pp. 599-612, doi:10.1109/TGRS.2008.2003435.

[9] W. Zhang, X. Sun, K. Fu, C.–Y. Wang, and H, -Q. Wang, "Object detection in high-resolution remote sensing images using rotation invariant parts based model," IEEE Gepsci. Remote Sens. Lett. Vol. 11, May. 2013, pp. 74-78, doi:10.1109/LGRS.2013.2246538.

[10] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," Proc IEEE, vol. 86, Nov. 1998, pp. 2278-2323, doi:10.1109/5.726791.

[11] T.-B, Xu, G.-L. Cheng, J. Yang, and C.-L. Liu, "Fast aircraft detection using end-to-end fully convolutional network," Int. Conf. Dig Signal Process DSP, Mar. 2017, pp. 139-143, doi:10.1109/ICDSP.2016.7868532.

[12] H. Wu, H. Zhang, J,-F, Zhang, and F.-J. Xu, "Fast aircraft detection in satellite images based on convolutional neural networks," Proc. Int. Conf. Image Process. ICIP, Dec. 2015, pp. 4210-4214, doi:10.1109/ICIP.2015.7351599.

[13] X.-Y. Chen, S.-M. Xiang, C.-L. Liu, and C.-H. Pan, "Aircraft detection by deep convolutional neural networks," IPSJ Trans. Comput. Vis. Appl, 2015, pp. 10-17, doi:10.2197/ipsjtcva.7.70.

[14] S.-Q. Ren, K.-M. He, R. Girshick, and J. S, "Faster R-CNN: Towards real-time object detection with region proposal networks," Adv. Neural inf. Proces.Syst, Jun. 2016, pp. 1137-1149, doi:10.1109/TPAMI.2016.2577031.

[15] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," Proc. IEEE. Comput Soc Conf. Comput Vision Pattern Recognit, Sep. 2014, pp. 580-587, doi:10.1109/CVPR.2014.81.

[16] R. Girshick, "Fast R-CNN," Proc. IEEE. Int. Conf. Comput Vision, Feb. 2015, pp. 1440-1448, doi:10.1109/ICCV.2015.169.

[17] W. Liu, D. Anguelov, and D. Erhan, et al, "SSD: Single shot multibox detector," Lect. Notes Comput. Sci, Sep. 2016, pp. 21-37, doi:10.1007/978-3-319-46448-0_2.

[18] J. Redmon, S. Divvala, R.Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," Proc. IEEE. Comput Soc Conf. Comput Vision Pattern Recognit, Jan. 2016, pp. 779-788, doi:10.1109/CVPR.2016.91.

[19] L. Pratt and S.Thrun, "Speical issue on inductive transfer," Mach. Learn, vol. 28, pp. 5.

[20] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," Proc. IEEE. Comput Soc Conf. Comput Vision Pattern Recognit, Apr. 2015, arXiv: 1409.1556 [cs]

[21] Q.-F, Fan, L. Brown, and J. Smith, "A closer look at Faster R-CNN for vehicle detection," IEEE Intell. Vehicles Symp. Proc, Aug. 2016, pp. 124-129, doi: 10.1109/IVS.2016.7535375.