Attention-Based Convolutional Networks for Ship Detection in High-Resolution Remote Sensing Images

Xiaofeng Ma, Wenyuan Li, and Zhenwei Shi(✉)

Image Processing Center, School of Astronautics, Beihang University, Beijing, China
{max15,liwenyuan,shizhenwei}@buaa.edu.cn

Abstract. Environmental information, like sea-land distribution, plays an important role in detecting ships from remote sensing images. However, the huge scale difference between environments and ship targets makes current CNN-based detection models hard to learn large-scale geographical information and focus on small targets at the same time. We propose an attention-based method by adding a Fully Convolutional Networks (FCN) to a detection networks as an attention branch to extract environmental features. Within a detection phase, the target detection branch is guided by the attention branch so as to focus on the potential target locations while in a training phase, the losses of other locations are simply ignored. We test our method on a public available remote sensing target detection dataset: LEVIR. By taking the classical Single Shot MultiBox Detector (SSD) as baseline, our method improves its detection accuracy in ship detection task while with an acceptable computational overhead.

Keywords: Ship detection · High-resolution remote sensing image
Attention model · Feature fusion · Convolutional Neural Networks

1 Introduction

With the development of imaging technology in remote sensing field, research work about image content interpretation like ship detection is of great importance for both military and civil applications. Researchers have spent a lot of time in developing detection methods and achieved many brilliant results in last few years. Before CNN-based methods, most detection systems are limited because they could only collect low level features and need a lot of time to locate targets. The most popular way to calculate features of candidate targets is block-wise methods, such as HOG [1] and SIFT [2]. Moreover, there are mainly two kinds of method to locate targets. One kind is morphological methods based on threshold processing. This method is simple but performs poorly in images with complex background. The other kind is sliding window methods. This way is time-consuming but works better than the former. Even so, this location method is still rough and inefficient.

J.-H. Lai et al. (Eds.): PRCV 2018, LNCS 11259, pp. 373–383, 2018.
https://doi.org/10.1007/978-3-030-03341-5_31

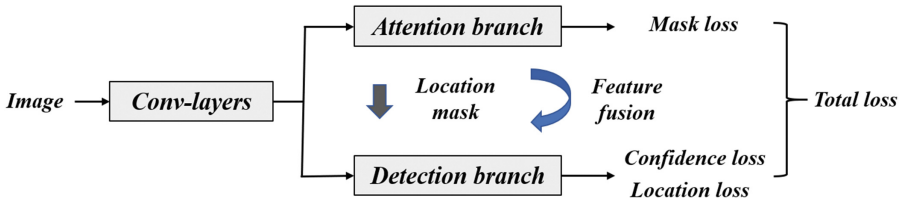

Fig. 1. An overview of our attention-based networks. We add an FCN to a detection networks as an attention branch. The attention branch produces a prediction mask for the locations of ship targets. Mask-loss will be calculated during training and FCN features will be fused into the detection feature maps.

Researchers in vision community developed CNN ways in order to extract more efficient features. The mostly adopted frameworks are R-CNN [3], fast R-CNN [4], faster R-CNN [5] which are two-staged and YOLO [6], SSD [7] which are one-staged. Besides, FCN [8], a special kind of CNN, is popularly used in image segmentation. It can get pixel-wise information and label images pixel by pixel [9].

There are many challenges to apply these models to ship target detection of remote sensing images. Ship targets in remote sensing images, compared with targets in vision images, are highly communicative with the environment. Environmental information, like sea-land distribution and ship wake, plays an important role in detecting ship targets. However, the huge scale difference between environments and ship targets makes current CNN-based detection models hard to learn large-scale geographical information and focus on small targets at the same time. This is a common problem in detecting targets from remote sensing images.

Lei etc. in [10] propose a method to make the detection system focus on the sea-area by extracting a sea-land mask before detecting ships. But this work is unsupervised and is separate from the detection backbone which means that it has little effect in improving the CNN-based model. Inspired by the attention model [11] used in neural network community, we construct our networks which is trained end-to-end to solve the problem mentioned above. Attention models generate attention maps besides feature maps to reveal regions where the following network parts should focus on. Zou etc. in [12] propose a network based on SVD which is more friendly in extracting features of ship targets in spaceborne optical images.

In this paper, we propose an attention-based method to extract environmental features and obtain the potential locations of ship targets which is implemented by adding an FCN to a detection networks shown in Fig. 1. Within a detection phase, the detection branch is guided by the attention branch so as to focus on the potential target locations while in a training phase, the losses of other locations are simply ignored. Besides, we also use a feature fusion strategy where features from the two branches are crossly fused to further improve the detection performance on small targets.

The main contributions of our work are as follows,

(1) By adding an attention branch to a classical network, the networks can focus on the area where ship targets exist with high probability. The detection branch is able to extract cleaner and more accurate features which is of great importance for targets detection in remote sensing images.
(2) By using a feature fusion strategy where features from the two branches are crossly fused, the network can integrate environmental information into detection feature maps. The detection branch achieves improvements in detection performance on detecting small targets.

2 Related Work

CNN is widely used in image related tasks and performs much better than typical methods. Current CNN-based object detection methods are widely used in vision community. Faster R-CNN [4] introduces a region proposal networks (RPN) to get proposal regions. Then features for each region are extracted from feature maps using roipooling. YOLO [6] frames object detection as a regression problem to spatially separated bounding boxes and associated class probabilities. Features are extracted based on the boxes from feature maps. SSD [7] discretizes the output space of bounding boxes into a set of default boxes over different aspect ratios and scales per feature map location. It extracts different features for different default boxes at the same location from different feature layers. Due to the extremely effective performance of CNN-based features, handcrafted features gradually fade out from recent detection systems.

Considering the scale difference between ship targets and remote sensing images and the scale variation of ship themselves [13], researchers proposed many specialized methods to improve the CNN-based detection models performance in detecting ships from remote sensing images. Lin etc. in [9] utilize an FCN to tackle the problem of inshore ship detection and design a ship detection framework that possesses a more simplified procedure and a more robust performance. And in [14], they propose another FCN to accomplish the tasks of sea-land segmentation and ship detection simultaneously. Liu, etc. in [15], mainly focus on the rotation information of the ship targets. They propose a Rotation Region-of-Interest (RRoI) method to project arbitrary-oriented proposals to a feature map. This framework performs better than axis-aligned pooling frameworks due to the large aspect ratio of targets. The framework is also adopted in text detection field and achieves competitive results [16,17]. Besides, new research works beyond detection begin to be studied like in [18], they try to generate humanlike language descriptions for remote sensing images. Chen, etc. in [19], proposed a method based on CNN to detect airports which is very significant.

Attention mechanism used in artificial neural networks simulates the attention model of the human brain so as to make network systems focus on the extraction of key information. Attention mechanism is extensively used in Natural Language Processing (NLP) and models [20] using this mechanism have

achieved the best results on many difficult sequence prediction problems (such as text translation). These approaches train the networks to generate both feature maps that encode the information of the input, and the attention maps that reveal regions of the feature maps where the following parts of the network should focus on. Similarly, we propose a method by using an FCN to guide the detection branch to focus on locations where ship targets exist with high probability during detecting. Losses of other locations are simply ignored during training. The attention branch can also integrate environmental information from high level into target features.

3 Method

3.1 Overview of Our Method

We take SSD as our detection branch. As shown in Fig. 1, we add an FCN branch to a detection networks to apply attention mechanism in the detection system. The attention branch is used to extract environmental features and obtain potential locations of targets.

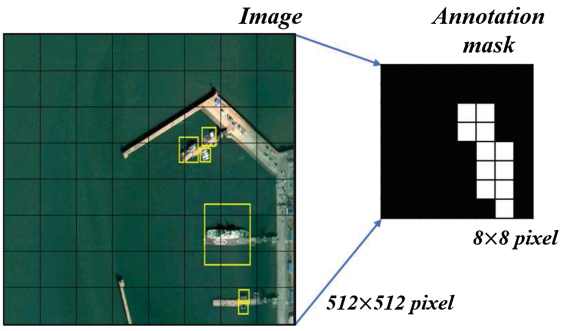

Fig. 2. Annotation mask for the attention branch. Left: ground truth of the input image, yellow bounding boxes denote the locations of each target ship and black grids denote the area of each pixel in the annotation mask. Right: annotation mask for the attention branch which is of the same shape with prediction mask of the attention branch. (Color figure online)

During training, we use heterogeneous annotations in the networks. Figure 2 denotes an annotation mask for the attention branch. Multitasking losses will be calculated to optimize parameters in the networks by adding the location mask loss to SSD losses. The losses of no-ship locations are simply ignored according to the predicted location mask. Moreover, by operating deconvolution, the high-level information about the environment will be fused with features in detection branch. A better feature extracting model will be gained after training. During detection, the attention branch will guide the detection branch to focus on the

potential target locations and provide environment information for the detection feature maps which will improve performance of the detection networks in remote sensing images.

3.2 Network Architecture

This section describes the design details of our Attention-based Networks. Important structural parameters are shown in Fig. 3. In detection branch, we take SSD as the base model. Layers in this branch decrease in size progressively so as to make predictions at multiple scales. At attention branch, we add an FCN at the end of Conv5_4 in detection branch. Then we apply softmax layer to produce prediction of the location mask. The annotation mask will be given when calculating mask loss in the train phase. Moreover, the features of reverse layers will be added to the detection layers to fuse large-scale environment information to single-scale feature maps. Through these improvements, the detection system is able to focus on the target-hot zone and obtain higher-quality feature maps.

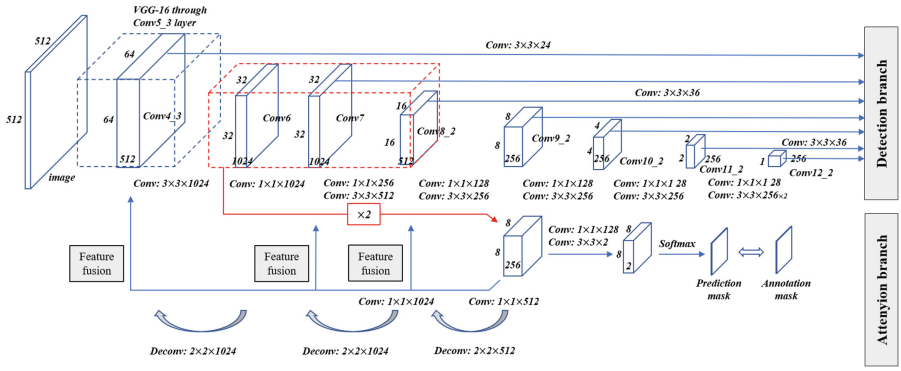

Fig. 3. Network architecture of our Attention-based Networks based on SSD. We achieve an attention-based networks with SSD as our base detection network. Attention branch uses the same structure as marked in the red dotted cube. (Color figure online)

We adopted deconvolution as our up-sampling method in the attention branch. Deconvolution is defined as transpose convolution. The operation of deconvolution can be described as that during backward convolution with stride f, the size of the input gradient map is f times that of the output-map. Thus, up-sampling with factor f can be seen as convolution with a fractional stride of $1/f$, accomplished by transpose convolution. The advantages of deconvolution include high efficiency and learnable convolution kernels.

3.3 Training Objective

The detection branch takes an image X as its input, and outputs a normalized score vector $L(X;\theta)$ for each default box. The objective of the detection branch

is to minimize error between ground-truths and estimated class labels, and is formally written as

$$\min_{\theta} \sum_{i} e_{cls}(y_i, L(X;\theta)) \tag{1}$$

where $y_i \in \{0,1\}$ denotes the ground-truth of the i^{th} example and $e_{cls}(y_i, L(X;\theta))$ denotes the classification loss of $L(X;\theta)$ with respect to y_i.

The attention branch obtains a prediction mask $M(g_i;\theta)$ by applying softmax function after the last layer. z_i denotes the binary ground-truth location mask of the targets shown in Fig. 2. The objective of the attention branch can be formulated as per-pixel regression, which minimizes

$$\min_{\theta} \sum_{i} e_{mask}(z_i, M(g_i;\theta)) \tag{2}$$

We use the following formulas to calculate the confidence loss ($conf$) and the location mask loss ($mask$)

$$L_{conf} = -\sum_{i \in pos} y_i log L - \sum_{i \in neg} log L^0 \tag{3}$$

$$L_{mask} = -\sum_{i} ||M - z_i||^2 \tag{4}$$

In order to get accurate target bounding boxes. We add Smooth L1 loss [5] between the predicted box (l) and the ground truth box (g) parameters to the total loss, with which to regress the offsets for the center ($cx; cy$) of the default bounding box (d) and for its width (w) and height (h).

$$L_{loc}(x,l,g) = \sum_{i \in pos}^{N} \sum_{m \in \{cx,cy,w,h\}} x_{ij}^k smooth_{L1}(l_i^m - \hat{g}_j^m) \tag{5}$$

$$\hat{g}_j^{cx} = (g_j^{cx} - d_i^{cx})/d_i^w,\ \hat{g}_j^{cy} = (g_j^{cy} - d_i^{cy})/d_i^h \tag{6}$$

$$\hat{g}_j^w = log(g_j^w/d_i^w),\ \hat{g}_j^h = log(g_j^h/d_i^h) \tag{7}$$

The overall objective loss function is a weighted sum of the localization loss (loc), a weighted sum of the mask loss and the confidence loss

$$L_{total} = \frac{1}{N}(L_{conf} + \alpha L_{loc}) + \beta L_{mask} \tag{8}$$

where N is the number of matched default boxes. If $N = 0$, we set the loss to 0.

4 Experiments

4.1 Dataset and Evaluation Metrics

We train and test our method on dataset LEVIR introduced by [21]. LEVIR consists of a large number of high resolution Google Earth images with over $22k$ images of 800×600 pixels and $0.2\,\mathrm{m} - 1.0\,\mathrm{m}$/pixel's resolution. There is a total of $11k$ independent bounding boxes including $4,724$ airplanes, $3,025$ ships and $3,279$ oil-pots. The average number of targets per image is 0.5. We just detect ship targets in the images. Table 1 shows the analysis of the ship targets in LEVIR based on scale.

Table 1. Analysis of ship targets in LEVIR based on scale.

Scale (*pixel*)	Scale1 (<100)	Scale2 (100–200)	Scale3 (200–300)	Scale4 (>300)	Total -
Number	1659	870	375	121	3025

We select the target-involved images from LEVIR as our dataset. The train dataset consists of 2844 images with 2325 ship targets and the test dataset consists of 947 images with 700 targets. We evaluate the detecting performance of our networks using average precision (AP) with different preset recalls and detecting speed using the number of frames that the network processes per second (FPS).

4.2 Training Details

We build our attention-based networks using the framework of tensorflow. Training parameters are listed in Table 2. We train base networks SSD with the same parameters. Both networks have been trained about 8 h. Pre-trained VGG model is used in both networks.

Table 2. Preset training parameters.

Parameter	Value
Learning rate	10^{-3}
Batch size	10
Momentum	0.9
Weight decay	5×10^{-4}
α	1.0
β	0.25

4.3 Results and Analysis

Table 3 presents the detected results of the base networks SSD (512×512) and our attention-based networks. We set the value of IoU to be 0.5 which means that the predicted bounding box whose IoU overlap is higher than 0.5 with the groundtruth box is confirmed as a correct detection. Our proposed networks achieve ∼3% improvement on detection AP on testing dataset. Both networks gain a ∼27 FPS speed during processing images.

Our proposed networks achieves almost the same precision as SSD but a much better recall. This indicates that the detection branch gets optimized under the guide of attention branch and is able to focus on the potential target locations. By deploying the mechanism of attention, the networks overcomes the problem of the huge scale difference between environments and ship targets to some degrees. It is able to gain environment information and target information at the same time.

Table 3. Detecting results of SSD and Attention-based Networks (IoU = 0.5).

Results	SSD	Our networks
AP (%) on training dataset	78.58	**84.17**
AP (%) on testing dataset	71.47	**74.44**
Prec (%) on testing dataset	**95.93**	95.76
Rec (%) on testing dataset	71.98	**75.03**
AP (%) of scale_1 targets	57.11	**61.58**
AP (%) of scale_2 targets	75.11	**79.29**
AP (%) of scale_3 targetst	69.83	**75.94**
AP (%) of scale_4 targets	69.23	69.23
FPS	**27.05**	**26.72**

According to the detection results of different-scale ship targets, both networks perform well in detecting large-scale ship targets. But in detecting small-scale ship targets, our attention-based networks performs much better. This indicates that the feature fusion strategy where features from the two branches are crossly fused really improve the detection performance on small targets. This also verifies that environmental information is of great importance in detecting ship targets.

In addition, we use the losses respectively to constraint the ship targets confidence, locations and environment mask. This is useful in improving the networks properties for training. Our whole networks is trained in an end-to-end mode which is helpful in obtaining a better model.

We list some result images on testing dataset using Attention-based Networks in Fig. 4. It demonstrates that our networks performs well in detecting small ships. As shown, ships with tail wakes and ships inshore are well detected.

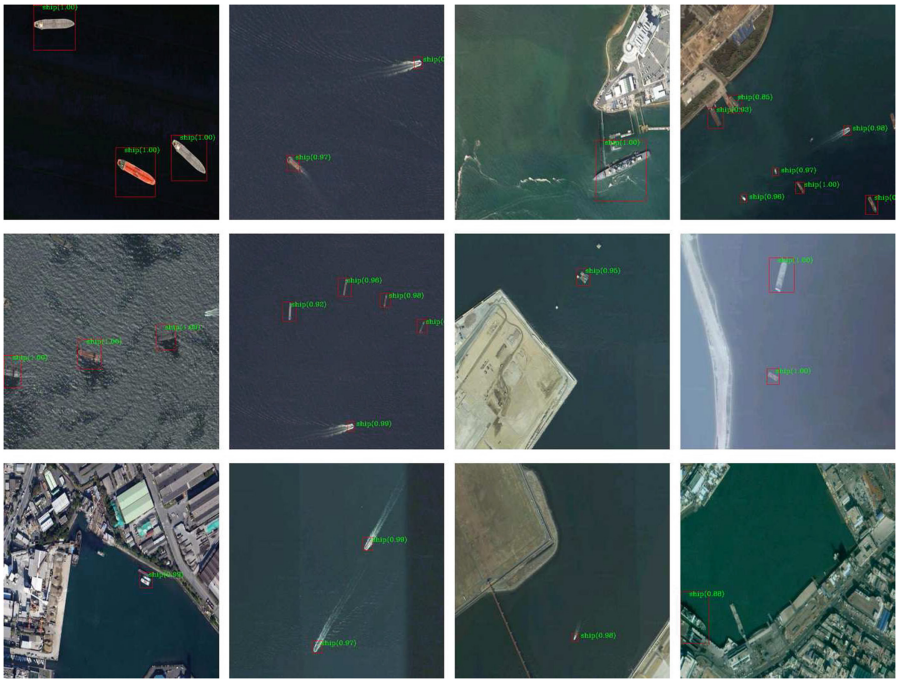

Fig. 4. Detection results on testing dataset using Attention-based Networks. Our attention branch improves the performance of detection branch based on SSD in detecting ship targets in remote sensing images. Especially, small-scale targets are well detected.

5 Conclusion

In this paper, we propose an Attention-based Networks to detect ship targets in high-resolution remote sensing images. By taking SSD as our baseline detection model, the proposed method achieves 3% AP improvement on testing data while with almost the same detecting speed. Our networks overcomes the problem of huge scale difference between environments and ship targets to some degrees which is of great importance for ship detection of remote sensing images. The attention mechanism in our networks can be easily applied to other detecting CNN-based models for targets detection of remote sensing images. Our networks currently cannot detect ships in instance level when ships are connected together. For future work, we aim to modify our networks to detect ships with various complicated backgrounds in instance level and to apply our method to detect other kinds of targets in remote sensing images.

Acknowledgments. The work was supported by the National Key $R\&D$ Program of China under the Grant $2017YFC1405600$ and the National Natural Science Foundation of China under the Grant 61671037.

References

1. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: IEEE Computer Society Conference on Computer Vision & Pattern Recognition, pp. 886–893. IEEE Computer Society (2005)
2. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. Int. J. Comput. Vis. **60**(2), 91–110 (2004)
3. Girshick, R., Donahue, J., Darrell, T., et al.: Rich feature hierarchies for accurate object detection and semantic segmentation. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 580–587. IEEE Computer Society (2014)
4. Girshick, R.: Fast R-CNN. In: IEEE International Conference on Computer Vision, pp. 1440–1448. IEEE (2015)
5. Ren, S., He, K., Girshick, R., et al.: Faster R-CNN: towards real-time object detection with region proposal networks. In: International Conference on Neural Information Processing Systems, pp. 91–99. MIT Press (2015)
6. Redmon, J., Divvala, S., Girshick, R., et al.: You only look once: unified, real-time object detection. In: Computer Vision and Pattern Recognition, pp. 779–788. IEEE (2016)
7. Liu, W., et al.: SSD: single shot multibox detector. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9905, pp. 21–37. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46448-0_2
8. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. IEEE Trans. Pattern Anal. Mach. Intell. **39**(4), 640–651 (2017)
9. Lin, H., Shi, Z., Zou, Z.: Fully convolutional network with task partitioning for inshore ship detection in optical remote sensing images. IEEE Geosci. & Remote. Sens. Lett. **14**(10), 1–5 (2017)
10. Lei, S., Shi, Z., Zou, Z.: Super-resolution for remote sensing images via local-global combined network. IEEE Geosci. Remote. Sens. Lett. **14**(8), 1–5 (2017)
11. Ba, J., Mnih, V., Kavukcuoglu, K.: Multiple object recognition with visual attention, arXiv preprint arXiv:1412.7755 (2014)
12. Zou, Z., Shi, Z.: Ship detection in spaceborne optical image with SVD networks. IEEE Trans. Geosci. Remote. Sens. **54**(10), 5832–5845 (2016)
13. Ding, H., Luo, Q., Zou, Z., Guo, C., Shi, Z.: Object detection with proposals in high-resolution optical remote sensing images. In: Yin, H., et al. (eds.) IDEAL 2017. LNCS, vol. 10585, pp. 242–250. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-68935-7_27
14. Lin, H., Shi, Z., Zou, Z.: Maritime semantic labeling of optical remote sensing images with multi-scale fully convolutional network. Remote. Sens. **9**(5), 480 (2017)
15. Liu, Z., Wang, H., Weng, L., et al.: Ship rotated bounding box space for ship extraction from high-resolution optical satellite images with complex backgrounds. IEEE Geosci. Remote. Sens. Lett. **13**(8), 1074–1078 (2017)
16. Jiang, Y., Zhu, X., Wang, X., et al.: R2CNN: rotational region CNN for orientation robust scene text detection (2017)
17. Ma, J., Shao, W., Ye, H., et al.: Arbitrary-oriented scene text detection via rotation proposals. IEEE Trans. Multimed. **PP**(99), 1 (2017)
18. Shi, Z., Zou, Z.: Can a machine generate humanlike language descriptions for a remote sensing image? IEEE Trans. Geosci. Remote. Sens. **55**(6), 3623–3634 (2017)
19. Chen, F., Ren, R., Van de Voorde, T., Xu, W., Zhou, G., Zhou, Y.: Fast automatic airport detection in remote sensing images using convolutional neural networks. Remote Sens. **10**(3), 443 (2018)

20. Schlemper, J., Oktay, O., Chen, L., et al.: Attention-gated networks for improving ultra-sound scan plane detection (2018)
21. Zou, Z., Shi, Z.: Random access memories: a new paradigm for target detection in high resolution aerial remote sensing images. IEEE Trans. Image Process. **27**, 1100–1111 (2018)