ROAM: A Recurrent Attention Model for Object Detection on High Resolution Remote Sensing Images

Cai Yuxiang
Student ID:11821038
Zhejiang University,China
Email: caiyuxiang@zju.edu.cn

***Abstract*—Object detection on high-resolution remote sensing images requires expensive computational costs because the amount of computation of the existing model is linearly proportional to the number of image pixels. To overcome this challenge, we propose a novel recurrent object attention model, called ROAM. It consists of a recurrent neural network, and is trained with the help of reinforcement rule. The trained model can point out the location of the object from images by following two steps iteratively. Firstly, it adaptively selects a range of candidate regions, records their locations, and loads the selected regions into memory with high resolution. Secondly, it estimates the probability that candidate regions contain the object. If the probability is large enough, the model outputs current location. Otherwise, it repeats the first step. It should be noted that the amount of calculation executed by the model can be controlled independently of the input image size. To demonstrate the effectiveness and efficiency of our model, we conduct a group of experiments to compare it with state-of-art models, e.g. FasterRCNN, on different datasets. The results show that our model needs less running time and GPU memory, and its computational cost is smaller.**

***Index Terms*—object detection, remote sensing image, recurrent neural network, reinforcement learning**

I. Introduction

Object detection is one of the most fundamental problems in computer vision. It has achieved great success on natural scene images with the help of rapid development on deep learning.However, the object detection on remote sensing images is still a challenging task because the remote sensing images are usually large-scale, massive and high-resolution. the number of pixels of a remote sensing image is much larger than that of a natural scene image. Although through the scanning window technique [1] [2], the object detection problem can be reduced to classification problem for many objects. However, since the classifier must be applied to all hypothetical image regions with different positions, scales, and various positions, the efficiency is low. Applying convolutional neural networks to large images is computationally expensive because the amount of computation is linearly proportional to the number of image pixels, although the amount of computation is reduced by downsampling [3] and sharing some calculations. But the main computational overhead of these models comes from convolving the filter map with the entire input image, so when the number of image pixels becomes large, their computational complexity becomes enormous.

When a person recognizes a thing, if it is relatively large, the person will construct the overall concept through local construction. Human visual attention has a good mechanism when selecting local areas. It will require fewer steps and more direction to judge the direction of this thing. The attention mechanisms are used to classify handwritten digital images [4], focusing attention selectively on certain parts of the visual space to obtain the information we want and to combine information from different gaze over time. Through the established reinforcement learning model, it guides the movement and decision making of future attention, focuses the computing resources on the parts we want on the scene, ignores other irrelevant features, and makes the ”pixels” that need to be processed less.

In this paper, we propose a novel recurrent object attention model, abbr. ROAM, applying the attention mechanism to handle object detection tasks in remote sensing images. Unlike a classification task, an object detection task requires the location of an object in the image. ROAM adds object location function in the process of serializing search objects to realize the detection of objects in remote sensing images.

II. Related Work

Remote sensing image object detection algorithm based on image analysis is mainly divided into two steps: image segmentation and object classification. First, the remote sensing image is divided into regions, and then the region is classified to determine whether it contains an object. For example, Baatz proposed the MRS (Multi-Resolution Segmentation) algorithm, which was the most popular OBIA algorithm at that time. The MRS algorithm used three parameters of shape, density and scale to divide the image into multiple regions. Dragut [5] proposes a new parameter definition tool that can effectively give scale parameters in the MRS algorithm. Although OBIA is flexible, and achieved good results in some areas combined with some contextual semantic information, this method still contains a lot of subjective information on how to define the segmentation region, and its algorithm is not universal.

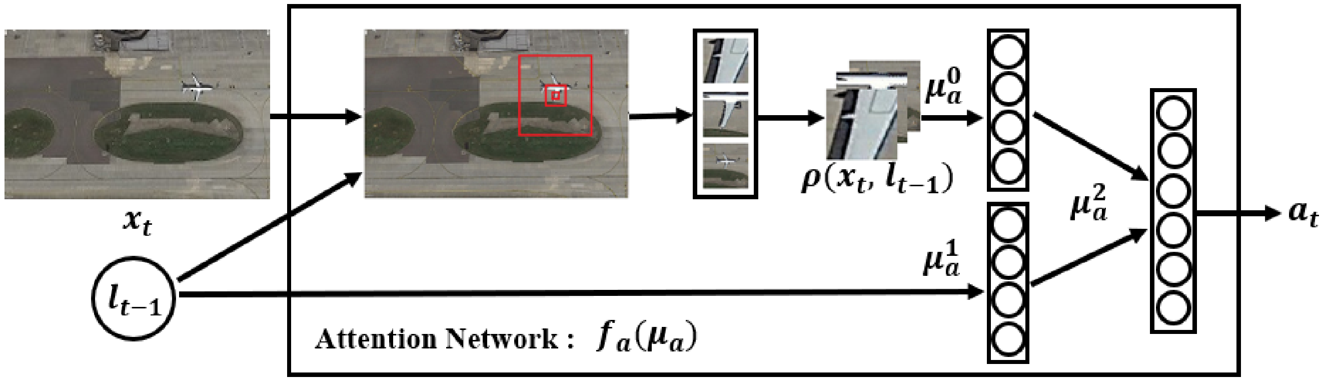

Fig. 1. Attention Network: input the location l_t of the image x_t into attention, and extract $\rho(x_t, l_t)$ containing the three resolution patches as part of the environment for each observation of the agent. For the input image x_t, $\rho(x_t, l_t)$ extracts three square patches centered on location *l*, the first patch is attention, the size is $g_w \times g_w$ pixels, and the other two consecutive patches are for the surrounding environment. The blurry glimpse, the width is k times the previous patch, then all the three patches are adjusted to $g_w \times g_w$ and connected.

Remote sensing image object detection algorithm based on machine learning is a recent research hotspot. The main idea is to obtain the region of interest through sliding window or other candidate box extraction methods, and then extract the layer semantic features in the image(Characteristics of statistical analysis of underlying features, such as HOG (Histogram of Oriented Gridients) features [6], BOW (Bag-of-Words) features [7]). The classifier model is trained with these features, such as the SVM (Support Vector Machines) classifier , and the trained classifier model is used to determine whether the region of interest contains an object. For example, Cheng [8] and Grabner [9] use the sliding window and HOG features for remote sensing image object detection. Shi proposed an algorithm combining circular frequency characteristics and HOG features for ship detection [10]. Zhang [11], Sun [12] and others use the sliding window and BOW features to detect geospatial objects. Aytekin detects airport objects based on texture features [13], and Zhong, Wang et al. also use texture features to detect urban areas [14]. Grabner et al. used the Haar-like feature of face detection feature to detect remote sensing image objects , and achieved good results. Remote sensing image object detection algorithm based on traditional machine learning has better accuracy, stability and universality than the method of template matching and OBIA. However, the sliding window algorithm used in this method brings too much computational cost, and the middle-level semantic feature used in the method is to count the underlying features, and can only effectively express the distribution of different textures, edges and other featuresbut cannot express object features with abstract semantics. For example, airport runway lines with approximate shape structures have completely different abstract semantic information from aircraft objects, but their middle-level semantic features may be very similar.

At present, there are few researches of object detection algorithms on remote sensing image based on deep learning. For example, Xiao et al. use the sliding window algorithm to extract the region of interest [15], and use the improved GoogLeNet network [16] to extract features for airport detection. Although this method exploits the high-level semantic features of convolutional neural networks, the sliding window approach still costs a lot of time. Deep learning models are more common which directly use natural scene images , such as R-CNN [17], Fast-RCNN [18], Faster-RCNN [19], YOL0 [20], SSD [21] and other deep learning models. These deep learning models contain a large number of network parameters, which carry a very large amount of computation, and must use large-scale data and take a lot of time to train. However, the number of remote sensing image object samples that people are concerned with is far less than that of natural scene images. At the same time, these models will carry out a convolution sampling, if the object contains only a small number of pixels, this approach significantly affect the detection accuracy. Another point is that these algorithms process the entire image, but for hundreds of millions of pixels of remote sensing images, hardware graphics memory can become a huge limiting factor.

III. Experiments

We note that [4] uses a visual attention mechanism to classify handwritten digital images. [4] proposes a recurrent neural network that sequentially selects image regions and combines extracted information from these regions for handwritten digit classification tasks. The model is a recurrent neural network (RNN) that processes inputs in sequence. In each step, the model does not process the entire image at once, but only one position in the image is processed at a time,and gradually combines information from those gaze To establish a dynamic internal representation or environment of the scene, and then in the next time step, select the next location to focus on based on past information and the needs of the task. The number of parameters in the model and the amount of computation it processes can be controlled independently of the size of the input image, which is in contrast to a convolutional network where the computational requirements are linearly proportional to the number of image pixels. The attention mechanism can solve the problem of large computational loss

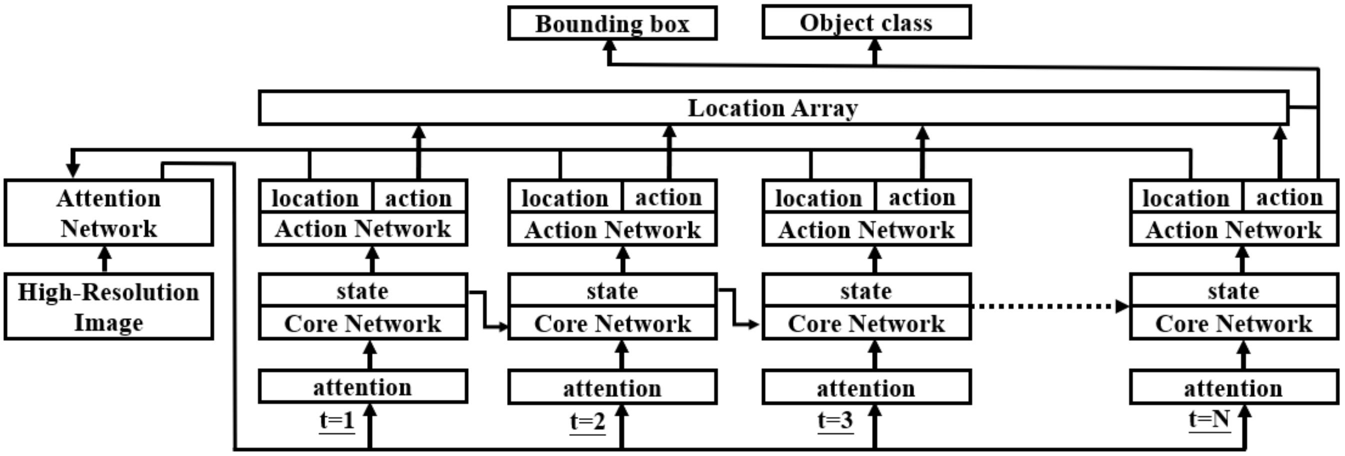

Fig. 2. Model Architecture: In general, the model is a recurrent neural network (RNN). The model consists of three parts: the glimpse network, core network, and action network. The glimpse network extracts the attention and glimpse from the image and converts it into a feature vector, and then sends the feature vector to the core network of the model. The core network combines the attention feature vector with the attention state at the previous time step h_{t-1} to produce a new attention state for the model h_t. The action network contains the environment action and location network, using the model's attention state ht to generate the next location and action to be processed. This basic RNN iteration is repeated for a variable number of steps. The detection results and detection boxes are given in the last step.

caused by the large number of pixels of the remote sensing image.

IV. Recurrent Object Attention Model

A. Model

In this paper, we consider the attention problem as the sequential decision process of the goal-oriented agent interacting with the visual environment. At each point in time, the agent only observes the environment through a fixed-size attention with assisting observation through two glimpses, it does not need to look at the entire environment at once. As shown in Figure 1. It can extract information in the local area through the attention box, and blur the surrounding information through two different sizes of glimpse box. Moreover, after the intensive learning training, the agent can actively select the attention location according to the learned policy. The agent can also execute operations to record the location information of objects in the environment and finally offer the location of the object. Since only a portion of the environment is observed at a time, the agent needs to integrate all of the available information over time to determine how to act and how to give the Most effective location of the attention. In each step, the agent receives a scalar reward (this depends on the action that the agent has executed and can be delayed), and the goal of the agent is to maximize the sum of rewards.

Then we will describe the network structure of ROAM in detail, and the agent is built around the recurrent neural network, as shown in Figure 2. At each time step, it processes the data which obtained by attention, integrates the information over time, and chooses how to operate and how to deploy its attention in the next step:

Glimpse network: At each step *t*, the agent receives the environmental observation in the form of an image x_t. The agent does not need to observe the image completely, but can extract information from x_t through its fixed-size attention $\rho(x_t,l_{t-1})$. By focusing the attention on some areas of interest. In this paper, we assume that a fixed-size attention extracts $\rho(x_t,l_{t-1})$ around the position l_{t-1} from the image x_t. It encodes the area around l with high resolution, but uses progressively lower resolution for pixels farther than distance l, resulting in a vector with a much lower dimension than the original image. We refer to this low-resolution representation as a glimpse to aid observation. The glimpse of the network generates an attention feature vector based on the attention.

Core network: The core network integration attention feature vector generated by glimpse network and the attention state of the previous time step, and then sends the information to the action network to guide the agent's next action. A very important part of core network is the attention state. The attention state is that the agent maintains an internal state that summarizes the information extracted from the history of past observations, it contains the agent's knowledge of the environment and helps determine how to act and where to deploy the attention. The attention state is formed by the hidden unit h_t of the recurrent neural network and is updated by the core network over time.

Action network: In each step, the agent executes two operations: it determines the location of the next attention based on the attention state of the core network. At the same time, it judges whether the smaller glimpse (the second patch) contains the candidate object, and if there is an object, the current attention coordinate is placed in the location array, and in the last time step, it determines whether if there are candidate objects in the image. If the object exists, the agent will traverse the location array, and extract the maximum and minimum values of x, y in the location array to determine the two-point coordinates (x_{min}, y_{min}), (x_{max}, y_{max}) of the diagonal of the detection box.

TABLE I

RESULTS ON THE REMOTE SENSING IMAGE. THE AVERAGE ACCURACY, AVERAGE RUNNING TIME, AND AVERAGE GPU MEMORY OF THE ROAM, SVM, CNN, AND FASTER RCNN MODELS ARE COMPARED.

Model	ROAMv1	ROAMv1	ROAMv1	ROAMv2	ROAMv2	ROAMv2	Sliding Windows	CNN	Faster RCNN
	(20 steps)	(25 steps)	(30 steps)	(20 steps)	(25 steps)	(30 steps)	(NMS)	(VGG-16)	(VGG-16)
AP(%)	85.3	88.6	90.6	86.0	88.2	90.2	82.7	83.9	91.5
run time/img (sec)	0.51	0.61	0.88	19.6	26.9	33.7	876.7	16.0	4.7
GPU memory/img(MB)	7.6	8.0	8.3	1.4	1.6	2.0	11.6	15.5	8.2

Reward: After executing the action, the agent receives a new visual observation of the environmentx_{t+1} and the reward signal r_{t+1}. The goal of the agent is to maximize the sum of the reward signals, which are usually very sparse and delayed:$R = \Sigma_{t=1}^{T} r_t$. In the case of object recognition, for example, if the object is correctly judged after the T step, r_T = 1, otherwise 0.

The above settings are a special example of a partially observable Markov Decision Process (POMDP) known in the RL community. The true state of the environment is unobserved. In this view, the agent needs to learn the policy π((l_t; a_t)$s_{1:t}$; θ) with the parameter θ, which, at each step *t*, maps the history of the interaction with the environment $s_{1:t}$ = x_1, l_1, a_1,... x_{t-1}, l_{t-1}, a_{t-1},x_t to the current time step of the action distribution, subject to attention. In our example, the policy π is defined by the RNN outlined above, and the history s_t is summarized in the state of the hidden unit h_t.

B. Training

The distribution is distributed over the possible interaction sequence $s_{1:N}$. Our goal is to maximize the reward under this distribution by using Eq. (1):

$$J(\theta) = \mathbb{E}_{p(s_{1:T};\theta)}[\Sigma_{t=1}^{T} r_t] = \mathbb{E}_{p(s_{1:T};\theta)}[R], \tag{1}$$

where $p(s_{1:T};\theta)$ depends on the policy.

It is very simple to maximize *J* precisely because it involves the expectation of high dimensional interaction sequences. However, considering the problem as POMDP, we can use the technique in the RL literature: as shown by Williams [23], the sample approximation of the gradient is given by (2),

$$\begin{aligned} \nabla_\theta J &= \sum_{t=1}^{T} \mathbb{E}_{p(s_{1:T};\theta)}[\nabla_\theta \log \pi(u_t|s_{1:t};\theta)R] \\ &\approx \frac{1}{n}\sum_{i=1}^{n}\sum_{t=1}^{T} \nabla_\theta \log \pi(u_t^i|s_{1:t}^i;\theta)R^i, \end{aligned} \tag{2}$$

where s^i is the interaction sequence obtained by the operation. The current proxy π_θ is for *i* = 1... *n* episodes.

The learning rule (2), also known as the REINFORCE rule, involves running the agent and its current strategy to obtain a sample of the interaction sequence $s_{1:T}$, and then adjusting the agent's parameter such that the selected log probability leads to an increase in the action of the high cumulative reward, actions that result in low reward actions are reduced.

Equation (2) requires us to calculate $\nabla_\theta \log \pi(u_t^i|s_{1:t}^i;\theta)$. But this is just the gradient of the RNN, which defines the agent we evaluated at time step *t*, which can be calculated by standard backpropagation [24].

Variance reduction: Equation (2) gives us an unbiased estimate of the gradient, but it may have a high variance. Therefore, usually consider the form of gradient estimation as Eq. (3)

$$\sum_{i=1}^{n}\sum_{t=1}^{T} \nabla_\theta \log \pi(u_t^i|s_{1:t}^i;\theta)(R_t^i - b_t), \tag{3}$$

Where $R_t^i = \Sigma_{t'=1}^{T} r_t^i$, is the cumulative reward obtained after the action u_t^i is executed, and bt is the baseline that may depend on$s_{1:t}^i$(eg, by h_t^i) but not on the action u_t^i itself. This estimate is equal to (2) in expectation but may have a lower variance. It is natural to choose $b_t = \mathbb{E}_\pi[R_t]$ [25]. This form of baseline is called the value function in the reinforcement learning literature. The resulting algorithm increases the log probability of the action, followed by a cumulative reward that is greater than expected, and reduces the probability if the cumulative reward obtained is smaller. We use this type of baseline and learn it by reducing the squared error between R_t^i and b_t.

Using a Hybrid Supervised Loss: The above algorithm allows us to train agents when the "best" action is unknown, and only provides learning signals through rewards. For example, we may not know which previous gaze sequence provides most of the information about the unknown image, but the total reward at the end of the episode will give us an indication of whether the sequence being tried is good or bad.

Dataset: Our training set is divided into two parts: training set and test set. The training set is a remote sensing top view with resolution of 256 × 256. The image only contains the object to be detected. In our experiments, detection object is the airplane. We use the images which includes airplane on the runway as a training set.

C. Network implementation details

ROAM: ROAM has two implementations. ROAMv1 directly reads the entire image into the GPU. ROAMv2 reads the image in the attention box from memory each time.

D. Results

Results are shown in Table 1.

V. Conclusion

In this paper, we apply a novel visual attention model to high-resolution remote sensing images because remote sensing images have high resolution and many pixel points, and the traditional object detection model is computationally expensive on remote sensing images. The model is developed as a single recurrent neural network that takes the attention window as input and uses the internal state of the network to select the next location to focus on and generate action signals. Although the model is indistinguishable, the proposed model architecture is trained end-to-end from pixel input to action using a policy gradient approach. Our experiments show that the model has several good effects on object detection on high-resolution images.

References

[1] Paul A. Viola and Michael J. Jones, *Robust real-time face detection.* In International Journal of Computer Vision,2004.

[2] Ming-Yu Liu, Arun Mallya, Oncel Tuzel, and Xi Chen, *Unsupervised network pretraining via encoding human design.* In Winter Conference on Applications of Computer Vision,2016.

[3] Alex Krizhevsky,Ilya Sutskever, and Geoffrey E. Hinton, *Imagenet classification with deep convolutional neural networks.* In Advances in Neural Information Processing Systems,2012.

[4] Volodymyr Mnih ,Nicolas Heess ,Alex Graves and Koray Kavukcuoglu, *Recurrent models of visual attention.* In Advances in Neural Information Processing Systems,2014.

[5] Lucian Dragut, Dirk Tiede and Shaun R. Levick, *ESP: a tool to estimate scale parameter for multiresolution image segmentation of remotely sensed data .* In International Journal of Geographical Information Science,2010

[6] Navneet Dalal and Bill Triggs, *Histograms of oriented gradients for human detection.* In Computer Vision and Pattern Recognition,2005

[7] Fei-Fei Li and Pietro Perona, *A bayesian hierarchical model for learning natural scene categories.* In Computer Vision and Pattern Recognition,2005

[8] Gong Cheng , Junwei Han, and Lei Guo,*Object detection in remote sensing imagery using a discriminatively trained mixture model .* In ISPRS Journal of Photogrammetry and Remote Sensing,2013

[9] Helmut Grabner , Thuy Thi Nguyen ,Horst Bischof and Barbara Gruber, *On-line boosting-based car detection from aerial images .* In ISPRS Journal of Photogrammetry and Remote Sensing,2008

[10] Zhenwei Shi ,Xinran Yu ,Zhiguo Jiang and Bo Li, *Ship Detection in High-Resolution Optical Imagery Based on Anomaly Detector and Local Shape Feature.* In IEEE Transactions on Geoscience and Remote Sensing,2014

[11] Dingwen Zhang, Junwei Han,Gong Cheng,Zhenbao Liu, Shuhui Bu and Lei Guo, *Weakly supervised learning for target detection in remote sensing images .* In IEEE Geoscience and Remote Sensing Letters,2015.

[12] Yu Li,Xian Sun,Hongqi Wang,Hao Sun and Xiangjuan Li, *Automatic Target Detection in High-Resolution Remote Sensing Images Using a Contour-Based Spatial Model.* In IEEE Geoscience and Remote Sensing Letters,2012.

[13] Örsan Aytekin,U. Zongur and U. Halici, *Texture-based airport runway detection.* In IEEE Geoscience and Remote Sensing Letters,2013.

[14] Ping Zhong and Runsheng Wang, *A multiple conditional random fields ensemble model for urban area detection in remote sensing optical images* . In IEEE Transactions on Geoscience and Remote Sensing,2007.

[15] Zhifeng Xiao,Yiping Gong, Yang Long, Deren Li, Xiaoying Wang and Hua Liu, *Airport detection based on a multiscale fusion feature for optical remote sensing images.* In IEEE Geoscience and Remote Sensing Letters,2017.

[16] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott E. Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke and Andrew Rabinovich, *Going deeper with convolutions.* In Computer Vision and Pattern Recognition,2015

[17] Ross B. Girshick, Jeff Donahue, Trevor Darrell and Jitendra Malik, *Rich feature hierarchies for accurate object detection and semantic segmentation .* In Computer Vision and Pattern Recognition,2014.

[18] Ross B. Girshick, *Fast R-CNN.* In IEEE International Conference on Computer Vision,2015.

[19] Shaoqing Ren, Kaiming He, Ross B. Girshick and Jian Sun, *Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks.* In Advances in Neural Information Processing Systems,2015.

[20] Joseph Redmon, Santosh Kumar Divvala, Ross B. Girshick and Ali Farhadi, *You only look once: Unified, real-time object detection.* In IEEE International Conference on Computer Vision,2016.

[21] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott E. Reed, Cheng-Yang Fu and Alexander C. Berg, *Ssd: Single shot multibox detector.* In European conference on computer vision,2016.

[22] Ronald A. Rensink, *The dynamic representation of scenes.* In Visual Cognition,2000.

[23] Ronald J. Williams, *Simple statistical gradient-following algorithms for connectionist reinforcement learning.* In Machine Learning,1992.

[24] Daan Wierstra, Alexander Förster, Jan Peters and Jürgen Schmidhuber, *Solving deep memory pomdps with recurrent policy gradients.* In International Conference on Artificial Neural Networks,2007.

[25] Richard S. Sutton, David A. McAllester, Satinder P. Singh and Yishay Mansour, *Policy gradient methods for reinforcement learning with function approximation.* In Advances in Neural Information Processing Systems,1999.

[26] James Bergstra and Yoshua Bengio, *Random search for hyper-parameter optimization.* In Journal of Machine Learning Research,2012