

Research of Image Main Objects Detection Algorithm Based on Deep Learning

Liyan Yu

School of Computer Science and Technology
Wuhan University of Technology
Wuhan, China
e-mail: 1078419389@qq.com

Xianqiao Chen, Sansan Zhou

School of Computer Science and Technology
Wuhan University of Technology
Wuhan, China

e-mail: chenxq@whut.edu.cn, 1193364293@qq.com

Abstract—Images have many objects in complex background. How to identify these objects and identify the main objects therein and understand the relationship between the main objects and other objects are the focus of this paper. There are many ways to object recognition, but most of them cannot mark the main objects of the image. In this paper, we use improved RCNN [1] network to detect and recognize multi-object in the image. Then we put forward the main objects scoring system to mark the image main objects. The experimental results show that the algorithm not only maintains the superiority of RCNN, but also detects the main objects of the image. We found that the image main objects were related to the size of candidate region and the rarity of objects.

Keywords—object detection; convolution neural network; scoring system; selective search; deep learning

I. INTRODUCTION

Object detection and recognition are not difficult for humans. Humans can easily understand the main objects of the image and the meaning of the image. But for the computer it's a very difficult task. Usually, there are more objects in the complex background. It's difficult for computer to identify the main objects and understand the relationship between the main objects and the other objects. As shown in Figure 1, we randomly selected two images in the flickr30k dataset. On the left image, the main object is a man marked with red border, and the other objects are a police officer, room and railings. On the right image, the main object is a robot marked with yellow border, the objects are a group of people and televisions. We want to identify the main objects, like the man and the robot.



Figure 1. An example of the main objects of the image

Image object detection is the current research hot spots, and there are already some research results. Dalal proposed a pedestrian detection method based on HOG and linear SVM [2]. The main idea is to use HOG features and linear SVM to learn positive and negative sample templates. Felzenszwalb proposed a DPM (Deformable Part Model)

based on the multi-scale deformation component detection model [3]. DPM is the best model for object detection beyond deep learning. DPM usually uses a sliding window detection method, which searches on each scale by constructing a scale pyramid. Region Convolutional Neural Network (RCNN) [1, 4, 5] is the earliest convolution neural network applied to object detection in the detection model. RCNN selects the candidate set based on the traditional selective search and establishes the detection idea of CNN's feature extraction network to detect the object. YOLO is a neural network based object detection system proposed by Joseph Redmon and Ali Farhadi et al in 2015 [6]. YOLO is the first CNN-based detection algorithm used for object detection end-to-end model. YOLO takes the whole image as the input of the network model and divides an image into $S \times S$ grids [6]. If the center of a object falls on the grid, the network is responsible for detecting this object and then outputting the object of the region and the confidence of the region.

II. MULTI-OBJECT DETECTION BASED ON RCNN

RCNN [1] is the earliest convolution neural network applied to object detection in the detection model. We select the candidate regions based on the traditional selective search. And then we establish RCNN to extract image features. It takes four steps for RCNN model to detect objects. There are selective search [7], CNN feature extraction, classification and bbox regression, as shown in Figure 2.

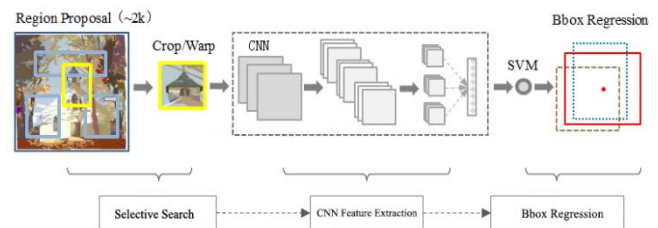


Figure 2. The process of RCNN detection and recognition the multi-object

A. Selective Search

The traditional sliding window method uses different scales of windows to scan the image at a certain step size, and it often obtains more redundant windows with a long time and low efficiency. Uijlings et al. [7] proposed a selective search algorithm to get regions proposal, through

continuously merging neighbouring regions with the largest similarity to obtain fewer target regions with high recall and narrow the search scope.



Figure 3. Selective search detection candidate region

RCNN extracts about 2,000 area candidates from the original image by selective search. Then all the regions are

zoomed to a fixed size by region normalization. We used the selective search test and the result is shown in Figure 3.

B. CNN Feature Extraction

In this paper, we use VggNet16 [8] to train each candidate region to get a 4096 dimensional eigenvector. The model's convolution part directly applies the convolution structure of Fast RCNN [4]. The initial weights are also assigned according to their pre-training weights on ImageNet. The network structure is shown in Figure 4.

Network input is 224×224 , and there are 5 layers of convolution and pooling operations, and the output is $6 \times 6 \times 256$ features in the fifth floor.

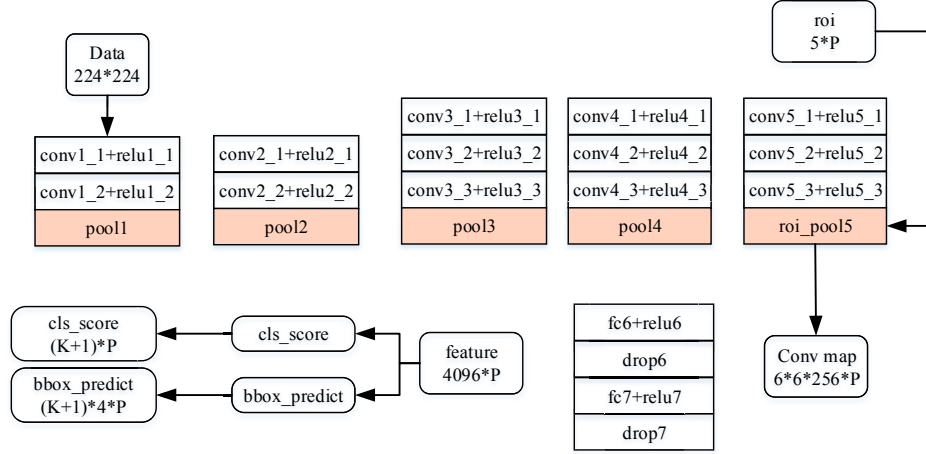


Figure 4. RCNN network structure

The first layer and the second layer use two convolution operations. After each convolution operation, the relu function is used to solve the gradient disappearance phenomenon, and then the convolution feature graph is subjected to the maximum pooling operation. The 3rd, 4th, 5th layers all use three convolution operations, the difference between them is that the pool's fifth layer does not use the maximum pooling, however it uses the RoI [4] pooling operation. RoI pooling evenly divides each candidate area into $M \times N$ blocks and performs the maximum pooling operation on each block. We can transform feature maps of different sizes into uniform-sized data with RoI. RoI layer is different from the other pooling layer training. We set x_i as the input layer node, y_j as the input layer node, the mapping function is:

$$\frac{\partial L}{\partial x_i} = \begin{cases} 0 & \delta(i, j) = false \\ \frac{\partial L}{\partial y_j} & \delta(i, j) = true \end{cases} \quad (1)$$

The decision function $\delta(i, j)$ indicates whether node i is selected as the maximum output by node j . There are two possibilities if not selected: x_i is not in the y_j range and x_i is not the maximum. For RoI pooling, one input node may be interconnected with multiple output nodes. Let x_i be the input layer node and y_{rj} be the j th output node for the r th candidate region. The mapping function is:

$$\frac{\partial L}{\partial x_i} = \sum_{r,j} \delta(i, r, j) \frac{\partial L}{\partial y_{rj}} \quad (2)$$

The decision function $\delta(i, r, j)$ denotes whether the node i is selected as the maximum output value by the j th node of the candidate area r . The cost for x_i is equal to the sum of all the related layers of gradients.

As shown in Figure 5, we visualize the extracted image features of the network. The features of the image will gradually abstract, and we use it to detection and recognition.

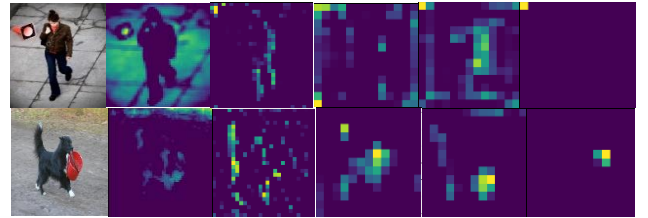


Figure 5. RCNN features visualization of various layers

C. Classification and Bregion Regression

Firstly, the feature vector output by CNN is used for classification, and a linear dichotomous support vector machine (SVM) [9] is used for each category as a classifier for determining whether the candidate window is a target object. Then through the boundary regression to get the exact target area.

SVM [9] is a kind of supervised learning method in machine learning, which is used to solve the problem of classification and regression, and it has strong superiority in

high dimension and nonlinear classification. Comparing with the artificial neural network, the neural network can reduce the misclassification rate by increasing the training samples, and the over-fitting problem is more complicated in the process of increasing the sample, that makes the network not universal. The support vector machine in the control of the error rate and its risk, the combination of SVM and CNNs, can avoid overfitting problem.

Non-maximal suppression (NMS) [10] is mainly used to remove redundant candidate regions in the target detection, extract the highest scores of the candidate regions, and find the best position of object detection. Features of selective search generated candidate regions are extracted from RCNN network, through SVM each region will get a score. But many regions are overlapped, and NMS needs to select the highest scores in those neighborhoods and suppress those low scores. After the classification score is obtained by SVM, the result is processed by NMS, and the final detection result is obtained by using the bounding-region regression to correct the candidate region position.

III. DESIGN OF IMAGE MAIN OBJECTS SCORING SYSTEM

There are many objects in complex images, and it is difficult to distinguish these objects whether the main objects of images. For example, two images in Figure 5, we can easily discern the main object of the left image is the girl, and the other object is the red road sign. So the score of the girl is higher than the red road sign. The main object on the right is the elderly, the other object is the hat, so the score for the elderly is higher than the hat. Usually, the main objects of the image have two main features. First, the ratio of the object's volume to the image. For example, the volume of the girl and the elderly in Figure 6 is larger than the other objects, so they are the main objects. The second is the rarity of all the objects in the image. If there is a person and a panda in the image, the panda can be considered as the main object of the image. So for the main objects in the image this section will propose a set of criteria for scoring.



Figure 6. An example of the main objects of the image

A. Object Rare Level Setting

First of all, we build a Object Rare Level Setting System based on expert knowledge to identify the main object.

We chose the 1000-class ImageNet [13] detection dataset. Part of the categories are shown in Table I. We evaluated these 1,000-class for their rarity. We evaluate with two criteria: One is expert knowledge of the rareness of objects. When people and the class co-exist in a picture, the subjective human will think which object is the main object of the image. The other is what we need. when the

experimental need what kind of class appears high probability, we can set the rare level of this class higher. This setting allows the filtered region to meet our expectations.

TABLE I. SOME CATEGORIES IN IMAGENET DATASET

chicken, ostrich, owl, gecko, toad, frog, turtle, lizard, crocodile, dinosaur, snake, fossil, spider, scorpion, centipede, bird, peacock, duck, goose, conch, elephant, hedgehog, platypus, kangaroo, rabbitfish, jellyfish, coral, sea creatures, sea snake, snail, conch, crab, lobster, hermit crab, penguin, bird, whale, walrus, sea lion, dog, wolf, fox, cat, leopard, leopard, leopard, lion, tiger, leopard, bear, bear, bear, mongoose, mongoose, dragonfly, starfish, sea urchin, rabbit, mouse, rat, squirrel, horse, zebra, pig, fish, abacus, suit, pen, muslim, bachelor, aircraft, airship, church, chapel, ambulance, bell, cabinet, clock, gun, backpacker, hot air balloon, amphibious, cabins, tables, casks, gymnastics instruments, stairs, seats, barbershop, trolleys, baseballs, basketball, babies.

Using the appropriate expert knowledge [11, 12], 1000 categories are divided into four levels. The results of the classification are shown in Table II. This table is a dynamically updated table, each time a new category added to the need to recalculate its rarity. Due to the large amount of data, we give a partial ranking result.

TABLE II. THE LEVEL OF CATEGORIES

Rare Level	Class
Level 1	dinosaurs, etc.
Level 2	peacock, alpaca, fossil, etc.
Level 3	ostrich, eagle, owl, turtle, koala, wolf, fox, leopard, tiger, bear, raccoon, pangolin, sloth, baboon, orangutan, monkey, elephant, raccoon, panda, hot air balloon, lighthouse, beaker, tank, computer, fire truck, survival boat, bulldozer, tank, tractor, etc.
Level 4	other class

B. Image Annotation Evaluation System

The purpose of the image annotation system is to recalculate the weights of the multi-objects recognized by RCNN and to find the main objects in the image. The meaning of an image is often limited, so we use the two features described in the previous section as a standard for the image annotation evaluation system. Design the evaluation system shown in Table III.

TABLE III. IMAGE ANNOTATION EVALUATION SYSTEM

Index	Score
$x > 0.85$	1
$0.35 < x \leq 0.85$	3
$0.15 < x \leq 0.35$	2
$x \leq 0.15$	1
$y = 1$	4
$y = 2$	3
$y = 3$	2
$y = 4$	1

where "x" represents the proportion of the object candidate region, y represents the degree of rareness of the object. "y=1" means that the degree of object rareness reaches level 1. The formula (3) can calculate "x".

$$x = (x_2 - x_1) \times (y_2 - y_1) / (a \times b) \quad (3)$$

$(x_1, y_1), (x_2, y_2), (x_1, y_2), (x_2, y_1)$ are the coordinates of the four vertices of the candidate region. The coordinates of the four vertices of the candidate region are calculated by selective search. a, b are the pixel value of the original image. The result of the mark score is:

$$\text{record}_{\text{sroce}} = \mu x_{\text{sroce}} + (1 - \mu) y_{\text{sroce}} \quad (4)$$

When the definition of the image annotation evaluation system, we need to truly mark the image of the weights. After the RCNN network predicts and classifies the candidate regions, the candidate class and the prediction score cls_{sroce} can be obtained.

First, candidate regions are sorted by descending order of cls_{sroce} and we obtain the top 5 candidate regions. Then we calculate the top 3 scores of these 5 candidate regions based on Equation 4. So the last five candidate regions are the main object of the image we need.

IV. ANALYSIS OF RESULTS

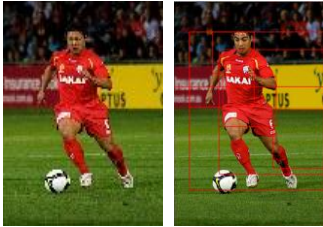










The experiments are based on the caffe learning framework [14] and the python language on a lab equipped with a TiTAN X graphics card with a processor of i7-4790K

and a memory of 32G. We use ImageNet 2012 and Flickr8k [15] as experimental datasets. ImageNet [13] is used to initialize the parameter values of the convolutional neural network, using its 1000 classes of a total of 1.2 million high-definition images to train the establishment of the network model. Flickr8k [15] is a dataset with more than 8,000 images and image descriptions. Image description is a sentence describing image information, which contains the main objects information in the image. So we compare the image description with the main objects we calculated.

A. Experimental Results

As shown in Table IV, we show the results of three randomly selected images in the Flickr8k [15] dataset. The result of the first image multi-object detection is "soccer ball", "football player", "sweatshirt", "miniature pinscher" and "dust jacket". "football player" scored a maximum of 2.6 ($x = 3, y = 1$), so "football player" was the main object of the first image. We take $\mu = 0.8$. The second image of the multi-object detection results are "pedestrian", "sunglass" and "safety pin". "pedestrian" scored a maximum of 1.8 ($x = 2, y = 1$), so "pedestrian" was the main object of the second image. The third image of the multi-object detection results are "Border collie" and "barrel". "Border collie" scored a maximum of 1.8 ($x = 2, y = 1$), so "pedestrian" was the main object of the third image.

TABLE IV. IMAGE ANNOTATION EVALUATION SYSTEM

Image	Multi-object Recognition		Score	Is Main Object
	multi-object	object recognition		
		0.83 soccer ball	1.8	no
		0.67 football player	2.6	yes
		0.38 sweatshirt	1.0	no
		0.07 miniature pinscher	1.0	no
		0.03 dust jacket	1.0	no
		0.30 pedestrian	1.8	yes
		0.30 sunglass	1.0	no
		0.04 safety pin	1.0	no
		0.76 Border collie	1.8	yes
		0.54 barrel	1.0	no

Flickr8k [15] data set description of the three images are:

- a).A man in a red uniform runs towards a soccer ball on a field.
- b).A woman in black and red listens to an Ipod , walks down the street.
- c).A black dog is carrying a red bucket in its mouth .

It can be easily seen from the image description that the first and third image object recognition results are correct, and the main object scoring result is also correct. The second image multi-object recognition results are incorrect, but inferred the main goal is correct. There are two reasons for the incorrect object detection. First, ImageNet does not have a woman's category, there is not enough data to support, so the model can not identify this type of goal. Second, the image features are not obvious. However, for the recognition of the main object, the algorithm proposed in this paper has high accuracy and can mark the main object of the image with high probability.

B. Validation and Evaluation

At present, there is not a very good method for identifying the main objects without distinguishing the background and the foreground. There are many ways to detect objects, such as YOLO [6], SSD [16] Fast RCNN [4], Faster RCNN [5], YOLO [6] and SSD [16] are all improvements based on the RCNN model. We compare the network structure of different object recognition with the main object algorithm.

We compared mAP and mOA for different network structures. The mAP [4] (Mean average precision) is an indicator of the accuracy of the algorithm in object detection algorithms. The mOA which defined in this paper is the accuracy of the main objects. If there are m images to be correctly detected and n images are correctly marked main objects, the mOA is n/m . Since each method's model has been trained, we only need to use the model to test Flickr8k data. The result of the test is shown in Table V.

TABLE V. COMPARISON OF MAP AND MOA FOR EACH METHOD

Method	mAP	Main object accuracy (mOA)
Fast RCNN[4]	52.7	0.91
Faster RCNN(VGG16)[5]	73.2	0.94
YOLO(VGG16)[6]	66.4	0.93
SDD300[16]	74.3	0.94

No matter what kind of structure of the network, mOA is about 90%. So if we want to find the main objects, we only need to focus on the rarity of the category and the size of image candidates. And the rare degree of expert knowledge is especially important and needs to be constantly updated.

V. CONCLUSION

In this paper, we proposed an image main objects scoring model. The model can extract image main objects. And we found that there is no relationship between the image object extraction and the image detection model, only

related to the size of the image candidate region and the rareness of the category. We reconstructed the network structure of RCNN and trained the ImageNet dataset with selective search, svm and bbox regression. Then we put forward the scoring system for recognizing image main objects. Finally, we use the scoring system to testing Flickr8k.

However, our method still has the following problems. Although bregion regression is used to adjust the image's candidate region, the size of the candidate region does not completely replace the area of the target object. So we need to calculate the outline size of the object. We have no remedy for inaccurate images and no error alerts. Follow-up work will be carried out in accordance with these two issues.

ACKNOWLEDGMENT

This work was partially supported by the National Natural Science Foundation of China (51179146) and completed under the help of my teaceher and schoolmates. Thanks them for selfless assistance and support.

REFERENCES

- [1] R. Girshick, J. Donahue, T. Darrell, et al. "Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation," Computer Vision and Pattern Recognition. IEEE, 2014, pp. 580-587.
- [2] N. Dalal, B. Triggs, "Histograms of oriented gradients for human detection," Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on. IEEE, 2005, pp. 886-893.
- [3] P. Felzenszwalb, R. Girshick, D. McAllester, and D. Ramanan, "Object detection with discriminatively trained part based models," IEEE Trans Pattern Analysis and Machine Intelligence, Sep. 2010, pp. 1627-1645.
- [4] R. Girshick, "Fast R-CNN," IEEE International Conference on Computer Visio. IEEE Computer Society, 2015, pp. 1440-1448.
- [5] S. Ren, K. He, R. Girshick, et al. "Faster R-CNN: Towards real-time object detection with region proposal networks," IEEE Transactions on Pattern Analysis & Machine Intelligence, 2017, pp. 1137-1149.
- [6] J. Redmon, S. Divvala, R. Girshick, et al. "You only look once: Unified, real-time object detection," Computer Vision and Pattern Recognition, June 27, 2016, pp. 779-788.
- [7] J. Uijlings, K. van de Sande, T. Gevers, "Selective search for object recognition," International Journal of Computer Vision, 2013, pp. 154-171.
- [8] K. Simonyan, A. Zisserman, "Very deep convolutional networks for large-scale image recognition," Computer Vision and Pattern Recognition, Sep. 2014.
- [9] V. Vapnik, C. Cortes, "Support-vector networks," Machine Learning, vol. 20, 1995, pp. 273-297.
- [10] N. Bodla, B. Singh, R. Chellappa, et al. "Soft-NMS Improving Object Detection With One Line of Code," IEEE International Conference on Computer Vision. IEEE, 2017, pp. 5562-5570.
- [11] J. J. Lennon, P. Koleff, J. J. D. Greenwood, et al. "Contribution of rarity and commonness to patterns of species richness," Ecology Letters, 2010, pp. 81-87.
- [12] M. Zuzana, "Determinants of species rarity: population growth rates of species sharing the same habitat," American Journal of Botany, 2005, pp. 1987-1994.
- [13] A. Krizhevsky, I. Sutskever, G. E. Hinton, "ImageNet classification with deep convolutional neural networks," International Conference

- on Neural Information Processing Systems. Curran Associates Inc. 2012, pp. 1097-1105.
- [14] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, et al. "Caffe: Convolutional architecture for fast feature embedding," the 22nd ACM International Multimedia Conference, vol, 2014, pp. 675-678.
 - [15] A. Karpathy, F. F. Li, "Deep visual-semantic alignments for generating image descriptions," Computer Vision and Pattern Recognition. IEEE, 2015, pp. 3128-3137.
 - [16] W. Liu, D. Anguelov, D. Erhan, et al. "SSD: Single Shot MultiBox Detector," European Conference on Computer Vision. Springer, Cham, 2016, pp. 21-37.