Design and Implementation of a Multi-scale Object Detection Algorithm on TensorFlow

Chen Yanbin a, Wang Huai b and Han Zhuo c
Space Star Technology CO., LTD., Beijing 100086, China
Email: acyb0475@sina.com, bwanghuai4210474@163.com, c1099535825@qq.com

Abstract. The technology of object detection is a very important subject in computer vision task. It has been widely used in intelligent transportation, face detection, aerospace and medical image equipment. In this paper, a kind of object detection algorithm based on regression and region proposals is studied. We use a popular deep learning framework—TensorFlow, as the experimental platform. We propose a multi-scale object detection algorithm based on RPN region proposal network, which can extract the features of large and small objects by using the level of feature map. In addition, this paper also improves the classification regression model and proposes a two-dimensional loss function, which makes the region proposals closer to the groundtruth boxes in the final training, which makes the training process of the improved network easier. The experimental data set used in this paper is PASCAL VOC data set, the accuracy and speed of each category in 20 detection objects are calculated and analyzed. A number of experiments have proved that the proposed multi-scale object detection algorithm based on RPN have improved in accuracy and speed. The average detection precision on the PASCAL VOC dataset increased to 74.4%.

Keywords: object detection, RPN, deep learning, convolutional neural network, multi-scale detection

1. INTRODUCTION

For human beings, visual image information is a kind of vivid and accurate description of objective things, and it is also a mainstream form that human beings can receive external information. Therefore, object detection is always a more difficult research topic in the field of computer image processing. Scientists hope that computers, like human beings, can transform the input image into the understanding and cognition of the external world, which will promote the development of human science and technology and society, and create higher scientific research value.

Object detection has been paid more attention and become a research hotspot of artificial intelligence and machine learning. Its main task is to identify and locate the object of human interest. Although object detection is widely used in real life and scientific research, it is still a challenging research topic because of the interference of low resolution, different light intensity, shooting angle, self posture position, shape change, occlusion and other factors[1].

There are two main development directions of the object detection model based on convolutional neural network. One is Region-based methods, and the other is the object detection model based on regression[2]. In 2014, Ross B. Girshick and others proposed a region based convolutional network (R-CNN). It uses the algorithm based on selective search to generate thousands of region proposal at the
input, uses convolutional neural network instead of the traditional way to extract image features in the region proposal, then trains the classifier separately for each kind of object, and uses SVM[3] to complete the classification of the object. In 2015, Ross B. girshick and others proposed an optimized version of R-CNN: Fast R-CNN[4]. It only extracts single feature from image, realizes multi task loss function, and uses softmax classifier instead of SVM classifier. The training speed is nearly 8 times faster than R-CNN. Although these methods continue to improve the detection speed, they still can not meet the requirements of real-time detection.

In order to achieve the requirements of real-time object detection as much as possible, the object detection algorithm based on regression idea came into being. This algorithm is no longer used in the process of extracting candidate boxes[5]. In 2016, Redmon et al. proposed the YOLO (You Only Look Once) algorithm that can detect the object in real time. Instead of extracting the candidate box, it directly divided the image into several image blocks. The biggest feature of YOLO is to realize end-to-end training[6]. Taking an image as the input, the output can directly get the position and category of the object bounding box. However, in the face of small object detection, its accuracy is lower than the first object detection algorithm based on region proposal.

In this paper, an object detection algorithm based on region proposal and regression will be studied respectively. We use TensorFlow, a popular deep learning framework, as the experimental platform, and propose a multi-scale object detection algorithm based on RPN region generation network[7], which can use the level of feature map to extract the features of large and small objects. This detection method combines the training idea of SNIP algorithm[8], trains several classification networks with different structures to adapt to various small-scale objects[9]. It is beneficial to the detection output of object objects.

2. RELATED WORK
Almost all of the current state-of-the-art object detectors are based on convolutional neural networks architectures. In this section, we review some CNN architectures and deep-learning-based object detection methods[10].

In 2012, Professor Hinton led a team in the ImageNet Large Scale Visual Recognition Challenge, ILSVRC. Alexnet[11], which was proposed by krizhevsky Alex, got the first place. Compared with the traditional detection algorithm, Alexnet reduced the top-5 error rate to 15%, and also proposed dropout method to reduce the over fitting problem in the training process.

Compared with the Alexnet network structure, ZFNet[12] has only made a small part of fine-tuning. The specific difference is that Alexnet uses a sparse connection structure of two GPUs, while ZFNet uses only one GPU. The former layer of ZFNet uses smaller convolution kernel, and the step size is reduced by 2 times, which can retain more features.

VGG-Net(Visual Geometry Group-Net), proposed by the visual geometry group at the University of Oxford, finished second on the image classification task in 2014[13]. Its innovation is to use a very small 3 * 3 convolution, and it is proved by experiments that increasing the depth of the network model will greatly improve the effect, and VGG-Net has excellent generalization ability for different types of data sets.

Ross B. girshick has made corresponding improvements on the problem that R-CNN needs to repeatedly calculate candidate boxes, and proposed Fast R-CNN, The main innovation is that Fast R-CNN[14] introduces the multi task loss function, synthesizes the object detection and boundary box regression into a network, cancels the step of step-by-step training network, and releases a lot of memory for storing feature data.

Faster R-CNN is an algorithm proposed in 2016. It aims at the problem that Fast R-CNN takes a long time to extract features using selective search algorithm, and selects a more effective method to extract candidate boxes.

You Only Look Once (YOLO) object detection model[15] is the first regression-based object detection model proposed by Joseph Redmon et al. in 2016. The YOLO model uses a single network and can output the probability of the boundary box and the category directly from the whole image.
PASCAL VOC Challenge[16] is an important test competition for visual object recognition and detection. It provides a standard image annotation data set and evaluation system to detect the performance of algorithms proposed by domestic and foreign researchers. PASCAL VOC dataset contains the most common objects in life, which can reflect the practicability of the algorithm. Therefore, we use this dataset for experimental training and test.

3. NETWORK ARCHITECTURE
The detection accuracy of Faster R-CNN on small targets has more advantages, while the detection efficiency is obviously insufficient. The YOLO method has a high detection efficiency, but the detection of small objects is not accurate enough. The main reason is the contradiction between the semantic information extracted by convolutional neural network and the object resolution. In this chapter, a multi-scale object detection algorithm based on RPN network, which is implemented by TensorFlow platform, is proposed.

3.1 Region Proposal Network
Region proposal network was proposed in Faster R-CNN. The emergence of RPN solves the problem of high calculation cost of candidate region generated by selective search model in Fast R-CNN. Its input is the original image, and its output is a series of rectangular candidate boxes containing scores. RPN is a full convolution network[17], which shares convolution characteristics with Fast R-CNN detection network framework. RPN network proposes a mechanism of anchor box, which is mainly reflected in translation invariance. When the position of the object to be detected is shifted, the position of the generated candidate box will shift accordingly, and the loss function selected by the prediction candidate box will be consistent.

3.2 Scale Normalization for Image Pyramids
SNIP algorithm[18] is a special multi-scale training method. At present, there are four methods to face multi-scale changes: 1. Deep and shallow feature fusion. 2. Change the size of convolution kernel to improve the sensitivity of the detector to the resolution, and then the large target can be identified. 3. Direct independent prediction on the characteristic map of shallow and deep layer. 4. Multi scale training test.

For the impact of scale on classification, the settings of the three networks are as shown in Figure 1:

Figure 1. Comparison of snip multiscale detection experiments.

CNN-B uses high-resolution image training to classify the down sampled and up sampled images[19]. CNN-B is a model trained on 224 * 224 scale. We down sample the test image to (48 * 48, 64 * 64, 80 * 80, 96 * 96, 128 * 128), and then enlarge it to 224 * 224 for test.

CNN-S uses low-resolution image training to classify the down sampled images. CNN-S is based on the above principles, we do a training test scale matching experiment. We choose 48 * 48 as the scale.

CNN-B-FT uses high-resolution image training, and then fine tunes the low-resolution image to classify the down sampled and up sampled images.

Extended to object detection, only when the scale of the object is close to the scale of the pre training data set, we can use it as the training sample of the detector. In the training, only those candidate frame gradients whose size is within a predetermined range are returned at a time, while those which are too large or too small are ignored; in the test, image pyramids with different sizes are established, and such
a detector is run on each image, and only those output results whose size is within the specified range are retained. Finally, they are not maximally suppressed together. In this way, the network can always be trained on the same scale objects.

3.3 Principle of multiscale object detection

The multi-scale object detection framework proposed in this paper is shown in Figure 2, and VGG-Net is selected as the backbone network.

![Figure 2. Principle diagram of multiscale object detection.](image)

Firstly, we input image, real object category and border information are sent to RPN network for training. Anchor box of fixed size will be defined at each position of the feature map. The length width ratio of anchor frame is consistent with that of Faster R-CNN (1:1, 1:2, 2:1). The size of low level feature map is defined as $H$, and the size of each high level feature map is reduced by half. In contrast to the feature map, the bottom anchor box has the smallest scale, and the scale of anchor box increases layer by layer from bottom to top.

The region generating network can predict a coordinate offset $\Delta \mu$, $\Delta \nu$, $\Delta W$, $\Delta h$ for the anchor box, that is, the difference between the front and back of the center coordinate of the anchor box and the difference between the width and height. If the central coordinate of anchor box is $(\mu, \nu)$, the central coordinate value and width height of the region of interest are shown in formula 1:

$$
\begin{align*}
\mu' &= \mu + \Delta \mu \cdot w \\
\nu' &= \nu + \Delta \nu \cdot h \\
W' &= W \cdot e^{\Delta w} \\
h' &= h \cdot e^{\Delta h}
\end{align*}
$$

We define the number of anchor boxes as $m$ and the number of aspect ratio as $n$. In the region generation network, the convolution kernel size is $3 \times 3$, the sliding window scans on the feature graph, and the dimension of the output vector is $512$. Then two fully connected layers are connected with it, one of which outputs the background probability of the region of interest, the second of which outputs the coordinate displacement difference and the width height scaling difference of the anchor frame for each region of interest. The border range of the region of interest is formula 2:

$$
4 \times H \times W \times M \times N
$$

Here, $H \times W$ represents the size of the feature map sent into the area generation network.

According to the SNIP algorithm in Section 3.2, when the convolution neural network extracts features from the first several layers and pools them, it often loses some information. Especially for the small object information, the loss after pooling is greater. Therefore, the region generating network needs to slide on the feature map of different depth and scale to get the background probability and coordinate information of the region of interest. The best detection effect is to detect small object information in shallow feature map, and deep feature map is more used to detect large object. After that, the detection results of each layer are connected in series to get the final detection results.
3.4 Improvement of classification regression model

One of the contributions of Fast R-CNN network is that each region of interest (ROI) can be sent to the classifier for classification after passing through the full connection layer; at the same time, it can also be sent to the branch of candidate box regression without occupying additional memory.

Because the output of the region of interest only contains the probability of the background and does not detect the classification information, the output of the region of interest needs to be sent to a classification regression model for operation, as shown in Figure 3.

![Figure 3. Schematic diagram of improved classification regression model.](image)

Every region of interest needs to be pooled. Considering that the backbone network is VGG-Net, the output size after pooling is set to 7 * 7, which is conducive to sharing the trained parameters in the full connection layer of the VGG model. The classification output is background probability and category. Suppose there are k categories in total, then the score output is k categories plus 1 pre background probability, i.e. K + 1; the regression output is Δ μ, Δ ν, Δ W, Δ h, i.e. the number is (K + 1) * 4, indicating the coordinates of different categories of border regression.

3.5 Training methods and Strategies

This multi-scale object detection algorithm will choose the same positive and negative sample sampling method as Fast R-CNN network. The region of interest with the intersection division union value (IOU) higher than 0.75 is defined as a positive sample, and the region of interest less than 0.25 is defined as a negative sample. Because the number of more than 2000 regions of interest is still relatively large, we sample these prospective regions of interest, and the final number of positive and negative samples is about 200. These samples will participate in the calculation of RPN loss function in the region generation network.

The traditional loss function generally uses one-dimensional coordinate loss function, as shown in formula 3:

\[
\text{Loss} = \sum_{l=1}^{\text{smooth}} (\Delta \mu - w)
\]

It is to add the loss of each coordinate point and calculate the loss of global offset. Because this method is to separate the coordinates of the real boundary box and the region of interest, it can not reasonably predict the overall offset loss of the two. Therefore, a two-dimensional area overlap loss function based on the idea of IOU is proposed in this experiment, as shown in formula 4:

\[
\begin{align*}
\mu_{\text{min}} &= \text{Min}(\mu, \Delta \mu) \\
\mu_{\text{max}} &= \text{Max}(\mu', \Delta \mu') \\
\nu_{\text{min}} &= \text{Min}(\nu, \Delta \nu) \\
\nu_{\text{max}} &= \text{Max}(\nu', \Delta \nu') \\
I &= (\mu_{\text{max}} - \mu_{\text{min}}) \times (\nu_{\text{max}} - \nu_{\text{min}}) \\
A &= (\mu - \mu') \times (\nu - \nu') + (\Delta \mu - \Delta \mu') \times (\Delta \nu - \Delta \nu') \\
U &= A - I \\
\text{New Loss}_{\text{RPN}}(I,U) &= -\log \frac{I}{U}
\end{align*}
\]
4. RESULTS

In this paper, we evaluate our network on PASCAL VOC challenges and compare the detection performance of object detectors with different object detection models. We train our network on a single Nvidia GeForce GTX 1080 Ti GPU.

In this paper, an RPN based multi-scale target detection algorithm based on tensorflow platform is proposed to verify the effectiveness and feasibility of Pascal VOC data set. Among them, the average accuracy of each target category AP (average precision) is used to measure the accuracy of detection and recognition of each object; the average accuracy of all categories mAP (mean average precision) is used. To measure the overall performance of the algorithm, we use FPS (frame / second) to view the object detection speed results of the algorithm; recall rate represents the proportion between the real positive samples and the detected positive samples.

Table 1. Experimental results of SS and RPN object candidate box algorithm in Pascal VOC dataset

| Method | Number of candidate boxes | Training set | mAP | IoU       |
|--------|---------------------------|--------------|-----|-----------|
|        |                           |              | 0.55| 0.65 | 0.75 | 0.85 | 0.95 |
| SS     | 2000                      | 2007         | 67.8| -     | -    | -    | -    |
| SS     | 2000                      | 07+12        | 70.0| 91.3  | 86.1 | 74.2 | 53.8 | 26.7 |
| RPN    | 200                       | 2007         | 69.3| -     | -    | -    | -    |
| RPN    | 200                       | 07+12        | 74.4| 95.7  | 91.9 | 75.3 | 38.8 | 4.7  |

From the experimental data showed in Table 1, we can also find that the region proposal network can still obtain higher detection accuracy than the selective search method when the number of candidate boxes is small, and also we can see that the recall rate is relatively high when the IoU is greater than 0.75; it can also be explained that the position deviation will be reduced when the RPN region is used to generate the network, and the quality of ROI in the region of interest will be improved high.

Table 2. Object detection speed based on Pascal VOC data set

| Method         | AP (%) | Time (ms) | FPS | Number of candidate boxes |
|----------------|--------|-----------|-----|---------------------------|
| Fast R-CNN     | 71.1   | 1790      | 0.5 | 2000                      |
| Faster R-CNN   | 72.5   | 133       | 8   | 300                       |
| Our method     | 74.4   | 84        | 10  | 200                       |

This experiment also compares the detection speed indexes of all image categories of three methods in Pascal VOC 2007 data set, as shown in Table 2. It can be found that compared with Fast R-CNN, the time consumption of FPS is significantly reduced. This is because the selective search method is removed and RPN is selected to extract the object candidate area. At the same time, the number of candidate boxes after sampling is reduced to about 200 in this paper, with small number and high precision. It also proves that the algorithm proposed in this paper is superior to the first two object detection algorithms in detection efficiency.
Table 3. Object detection accuracy based on Pascal VOC data set

| Method       | mAP  | aeroplane | bike | bird | boat | bottle | bus | car | cat | chair | cow | dining table | dog | horse | mbike | motorbike | person | potted plant | sheep | sofa | train | tv |
|--------------|------|-----------|------|------|------|--------|-----|-----|-----|-------|-----|--------------|-----|-------|-------|------------|--------|--------------|-------|------|-------|----|
| R-CNN        | 71.6 | 72.2      | 79.2 | 70.8 | 63.3 | 47.5   | 76.1| 73.5| 65.3| 54.5  | 78.6| 71.9         | 78.1| 71.5  | 72.2  | 67.2       | 49.7  | 73.9         | 78.0  | 73.4| 74.2  |
| Fast R-CNN   | 73.2 | 72.7      | 73.9 | 76.3 | 72.8 | 52.6   | 76.8| 76.2| 76.2| 57.3  | 72.1| 70.9         | 82.6| 72.1  | 74.9  | 79.8       | 51.3  | 78.6         | 72.4  | 82.7| 73.3  |
| Faster R-CNN | 73.1 | 77.8      | 77.5 | 76.8 | 62.3 | 57.3   | 74.8| 72.4| 80.6| 64.2  | 73.7| 70.1         | 63.8| 80.8  | 75.8  | 73.1       | 52.5  | 65.9         | 77.2  | 77.1| 68.1  |
| Our method   | **74.4** | **78.6** | **77.0** | **69.6** | **61.9** | **70.8** | **85.3** | **81.7** | **90.1** | **58.7** | **83.9** | **68.3** | **68.1** | **89.8** | **85.1** | **32.8** | **54.8** | **81.2** | **68.3** | **79.7** | **82.3** |

Table 3 shows the average results of the test accuracy of all image categories. It can be seen that the multi-scale object detection algorithm based on RPN network proposed in this paper is superior to other network methods in the comparison experiment in the accuracy index map. In addition, the detection rate of some small object sets is improved more obviously, such as bottles, potted plants, birds, etc.

5. CONCLUSION

In this paper, we propose a multi-scale object detection algorithm based on RPN network, which is implemented by TensorFlow platform. The improved algorithm and network structure can use RPN to extract features from multiple feature maps with different depth and scale. At the same time, inspired by the SNIP algorithm, more information about small objects is collected in the low-level feature map.

In the multi-scale object detection algorithm, the role of RPN network and the subsequent classification regression model are two core parts. At the same time, a two-dimensional area overlapping loss function is proposed, which makes the candidate boxes closer to the ground truth boxes in the final training regression process. Therefore, in the face of multi-scale object detection, the object scale covered by different depth layer structure receptive field is different, and the analysis of experimental data also proves that the performance of multi-scale object detection can be improved.

REFERENCES

[1] Joseph R, Santosh D, Ross G and Ali Farhadi 2016 You only look once: Unified, real-time object detection. 779–88
[2] Tao K, Anbang Y, Yurong C and Fuchun S 2016 Hypernet: Towards accurate region proposal generation and joint object detection. 845–53
[3] Jan H, Rodrigo B, Piotr D and Bernt S 2016 What makes for effective detection proposal-s. 814–830
[4] Cheng Yang F, Wei L, Ananth R, Ambrish T and Alexander C 2017 Dssd: Deconvolutional single shot detector
[5] Spyros G and Nikos K 2015 Object detection via a multi-region and semantic segmentation-aware cnn model,1134–42
[6] Karen S and Andrew Z 2014 Very deep convolutional networks for large-scale image recognition
[7] Jasper RR U, Koen EA Van De S, Theo G and Arnold W 2013 Selective search for object recognition. 154–71
[8] Ross G 2015 Fast r-cnn,"IEEE international conference on computer vision. 1440–48
[9] Kaiming H, Xiangyu Zh, Shaoqing R and Jian S 2014 Spatial pyramid pooling in deep convolutional networks for visual recognition. 346–61
[10] Jifeng D, Yi Li, Kaiming H and Jian S 2016 R-FCN: Object detection via region-based fully convolutional networks. 379–387
[11] Alex K, Ilya S and Geoffrey E 2012 Imagenet classification with deep convolutional neural networks.1097–105
[12] Gao H, Zhuang L, Kilian W and Lau-rems van der M 2016 Densely connected convolutional networks
[13] Yousong Zh, Chaoyang Zh, Jinqiao W, Xu Zh, Yi W and Hanqing L 2017 Couplenet: Coupling
[15] Sergey I and Christian S 2015 Batch normalization: Accelerating deep network training by reducing internal covariate shift
[16] Wei L, Dragomir A, Dumitru E, Christian S, Scott R, Cheng Yang F and Alexander C 2016 SSD: Single shot multibox detector, "conference on computer vision. Springer. 21–37
[17] Sean B, C Lawrence Z, Kavita B and Ross G 2016 Inside-outside net: Detecting objects in context with skip pooling and recurrent neural networks. 2874–83
[18] Tsung Yi L, Piotr D, Ross G, Kaiming H, Bharath H and Serge B 2016 Feature pyramid networks for object detection
[19] Liu C, Gong S and Chen C 2014 Evaluating Feature Importance for Reidentification Person Re-Identification, 203–28