Contents lists available at YXpublications

International Journal of Applied Mathematics in Control Engineering

Journal homepage: http://www.ijamce.com

Research on Object Detection Algorithm Based on Deep Learning

Shengwang Li, Li Chen*

School of Information Science and Engineering, Hebei University of Science and Technology, Shijiazhuang, Hebei, China

ARTICLE INFO

Article history:
Received 10 April 2018
Accepted 10 July 2018
Available online 25 December 2018

Keywords:
Deep Learning
Convolutional Neural Networks
Migration Learning
Object Detection

ABSTRACT

Object detection is a key precondition for solving the problems of recognition, tracking, and semantic analysis. This paper uses a method of object detection based on deep learning. The deep convolutional neural network detect the target object, using its advantages of self-extraction features and learning to avoid the complex feature extraction and manual tagging process. In a relatively complex background, it shows good performance. At the same time, in order to prevent over-fitting, the concept of migration learning is used and integrate the existing network dataset VOC2012 and objects collected under different lighting conditions into new datasets. Through parameter adjustment and experimental verification, the detection rate of this method on the new datasets reaches 92.13%, which greatly improves the traditional target detection method and has a certain degree of robustness to external factors. For the object detection, the method of modifying the SSD algorithm is adopted to replace the VGG16 network, and the deep network structure composed of ResNet and inception is used to detect the objects in the image. The final input image of the experiment is 300*300 and 512*512. By comparing the effect of these two different sizes of image size, using a multi-window detection method to divide the area to optimize the detection accuracy, and finally achieve the effect of mAP 0.644.

1. Introduction

Object detection is an important field in computer vision. With the rapid development of internet, artificial intelligence technology and intelligent hardware, there are a lot of image and video data in human life, which makes computer vision technology, play an increasingly important role in human life, and the research of computer vision is becoming more and more popular. Target detection and recognition, as the cornerstone of computer vision, is attracting more and more attention. It is also widely used in real life, such as target tracking, video surveillance, information security, autopilot, image retrieval, medical image analysis, network data mining, UAV navigation, remote sensing image analysis, national defense system, etc. The main task is to locate the target of interest from the image. Positioning means giving the target a bounding box and then getting the specific category of the object in the bounding box [1]. With the development of deep learning, convolutional neural networks have become more widely used and have achieved excellent results in the field of object detection [2]. In 2012, Convolutional Neural Networks achieved excellent results in the Global Image Contest, so the adaptive image extraction features of Convolutional Neural Networks began to receive attention. The convolutional neural network updates the network parameters through the back-propagation algorithm, and adaptively adjusts the weighted and effective combination features of different features to obtain higher-level features with better robustness [3]. Therefore, if the computer is allowed to actively study the characteristics of the image, compared with the traditional artificial design features, the accuracy of detection can be effectively improved and the effect of the experiment can be improved. The network structure of convolutional neural networks is highly invariant to translation, scaling, tilting, or other distortions. It is particularly prominent in image recognition, speech recognition, and natural language processing problems [4].

The focus of this study is to use neural networks for object detection. Through the design of the network model, the parameters are continuously updated, the complexity of the calculation is further reduced, and the accuracy of the detection is improved. In specific implementation, the existing deep learning development tools will be used and combine with the migration learning method to update the model parameters to extract the depth characteristics of the objects and train the classifier. At the same time, after the sliding window is detected, the area overlapping rate threshold is dynamized to further improve the detection accuracy. For the same datasets in the traditional target detection method has also done a

* Corresponding author.
E-mail addresses: 2394987221@qq.com (L. Chen)

corresponding experiment, through comparison can be drawn, based on the convolutional neural network target detection method compared to the traditional target detection [5-7] method, the speed of The improvement comes from the resampling phase that removes the proposal and pixel features of the bounding box. In the experiment, modify the target detection method VGG16 network structure used in SSD, the size range of the bounding box is converted, the labeling of the target object is labeled with LableImg[8], and the minimum frame is selected for labeling. Finally, the dimensions are reduced by adjusting the size of the pooling layer and the size of the filter layer. This not only reduces the amount of calculation, but also greatly improves the utilization of the parameters.

The paper is mainly divided into three parts. Part II is mainly used for neural network and neural network for target detection. Part III mainly involves the design of the network structure, training method and migration model of this experiment to optimize the network model. The IV part is mainly the analysis of experimental results. By comparing the original SSD object detection method, we can conclude that the object detection method designed in this experiment has greatly improved.

2. Related Work

With the rapid development of the convolution neural network, in recent years, researchers have been working to reduce the overhead of network time, improve the detection speed, and realize fast and good detection. Detection accuracy is the initial and most important index of the object detection task. How to improve the method detection precision mAP is the most basic index [9] of the comparison of various methods. In 2014, for the first time in the paper [10], compared with a simpler HOG[11] class based system, CNN can significantly improve the target detection on the PASCAL VOC, that is, the R-CNN target detection model is proposed: it uses the selective search method to extract a number of candidate frames from the image to be detected and then convert the candidate frame into a sequence. The size is unified, and then the feature is extracted through neural network. Finally, the features are classified by multiple SVM [12]. Fast-CNN model: it also uses the selective search method to extract several candidate frames from the image to be detected, mapping the corresponding feature map of the region of the candidate frame into a fixed length feature vector through the spatial Pyramid pool layer, then classifies it through a fully connected neural network, predicts the coordinates of the boundary frame, and the candidate region. Modify [13]. YOLO: its innovation is that it reforms the regional proposed frame detection framework and divides the whole map into a SXS lattice. Each grid is responsible for the detection of the target in the grid. The bbox, location confidence, and all probability vectors are used to solve the problem at once (one-shot). The advantage of working in identifying efficiency is very obvious[14]. For the SSD target detection method, it uses the method of proposal and multiscale on the basis of YOLO, divides the whole map into 8*8 grid, uses the feature graph of different layers, uses the 3*3 window receptive field as the different scale detection, and uses the convolution layer instead of the full connection layer of YOLO to make the budget, so that the detection speed is made. There was a great promotion. Compared with YOLO's mAP, SSD achieves about 51.3% [15]. However, the lack of SSD is that it uses VGG16 as the training network model, which will reduce the accuracy of the model. It not only affects the accuracy of the selected box when the final target is detected, but also has low detection rate for high complex images. So we want to replace the training network in SSD, use the method of migration learning to carry out the pre training of the new network model, which not only increases the utilization of the parameters, but also trains the network detection model that we need.

To sum up, this paper mainly focuses on three aspects of target detection:

1. Establish a model to extract post constituencies from scenes;
2. Identify the selected objects and classify them;
3. Object classification and location can be optimized by adjusting the parameters of the classification model or adjusting the candidate frame.

3. Network Structure

The SSD network structure model obtains multi-class Boxes with different scales and locations as possible target markers after convolution calculation for a frame of image, which improves the poor accuracy of YOLO target detection location. Finally, the final detection results were screened by NMS (non maximal inhibition) method.

This article uses the SSD network structure for object detection. The general process of the SSD is shown in Figure 1.

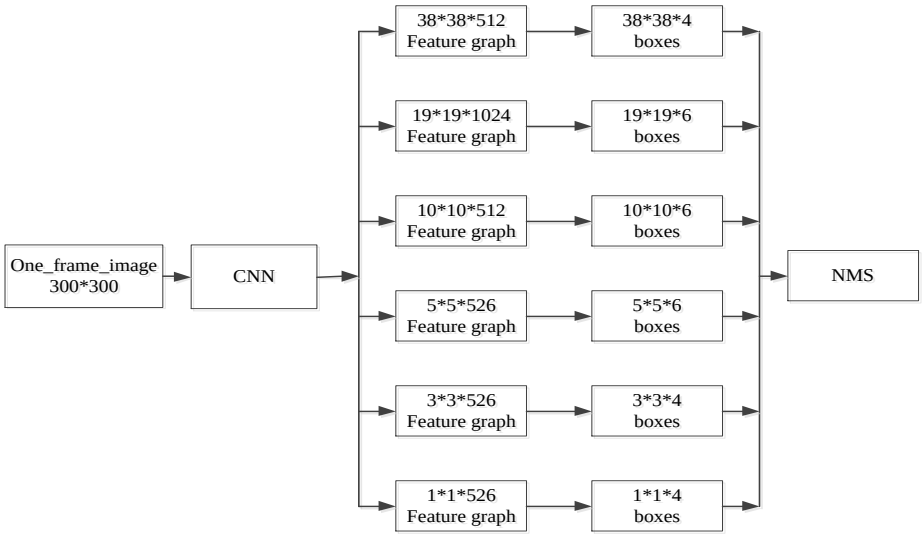


Fig.1 SSD flow chart

For SSD model, by choosing different layers and different levels of size, and different proportion of anchors, we can get the best matching anchor with the ground truth to train, making the whole structure more accurate.

3.1 Datasets

The datasets includes training datasets, validation datasets and test datasets. The training datasets mainly uses the data in the VOC2012 inside. In SSD's original papers, data was also augmented in data. Data augmentation is very helpful to improve the recognition accuracy. It can upgrade mAP 65.5 to 74.3, which plays the most important role in all the innovations. Fast R-CNN and Faster R-CNN use the original image, and 0.5 probability of horizontal flip of the original image for training, but also use the sampling strategy[16].

In order to prevent the over fitting phenomenon, we use camera to collect objects around us under different illumination and different angles. Then the image annotation is carried out by the image tagging software LableImg, and the image files of the XML are obtained. The content is the category of image and the coordinate information of the location of the image. The annotation information is shown in Figure 2:

Fig.2 Image annotation

In order to better train the image, we preprocess the data. In the preprocessing stage, We readjust the minimum edge of the input image, and then cut them into the 300*300*3 and 512*512*3 size patch of size 300*300*3 (cropped central patch of size 300*300*3), subtracting the average from the training set from each pixel. Data processing is done on CPU, and can be trained on GPU, so it doesn't need much computation time. The datasets in this experiment was conducted on a new set of datasets by VOC2012 and its own set of data. 960 objects were tagged and 10 kinds of objects were detected: birds, ships, bottles, cars, chairs, computers, ears, people, potted plants, and mouse.

The original data includes not only the training set and test set, but also some files, such as LabMap file, test_name_size. txt, test. txt, trainval. txt. This article uses its own data and needs to use creat_list.sh to generate test.txt, trainval.txt, and test_name_size.txt files. Parameters include: - anno_type: detection or classificat ion; - label-map-file: map file; - check-label: check for duplicate names or tags; root path: path containing images and annotations; LISTFILE file: trainval. txt and text. txt; outdir: store database file; exampledir: link to database The program calls the trainval. txt and text. txt files in the listfile, reads the image and its annotation information, outputs LMDB data to outdir, and generates a link file under the example file. Run example/ssd/ssd_line.py. This file includes the definition of the model, the automatic generation of. prototxt file, the generation of training scripts and run, save the training model and so on. Major modifications: text_data, train_data, save_dir and other paths, job_name, model_name, gpu, num_classes, etc. Training results in the data / results path, is two TXT files, containing two categories of bounding box information.

3.2 Convolutional neural network framework construction and feature extraction

In the SSD paper mentioned that it uses the VGG16 network structure training model [17]. Its framework structure is shown in Figure 3. In SSD architecture, it is usually divided into two parts: front-end and back-end. The front-end is used to remove the hierarchical classes. Convolutional neural network model is adopted. The back-end is located in the multi-scale feature detection network, which mainly extracts features from the feature layer generated by the front-end through different scale conditions.

VGG16 network model is used in the original paper to extract the initial features of the target. The VGG16 network structure is modified on the classic AlexNet structure. SSD model adds a convolution layer to the AlexNet model to add a feature layer. A series of convolution kernels can be used to produce a series of fixed-size predictions. For a feature with one channel, the convolution kernels are often used.The AlexNet network structure is shown in Figure 4.

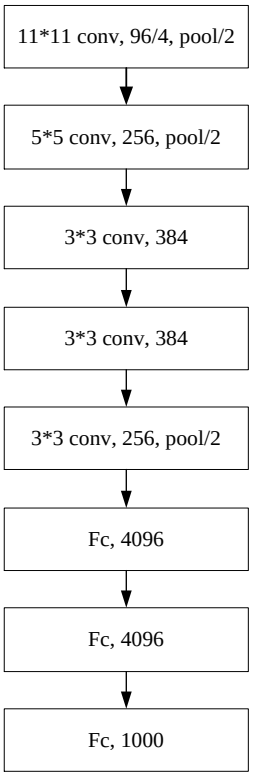


Fig.4 Deep Convolutional Neural Network architecture

It is a depth of 8 layers consisting of 5 layers of convolution layers and 3 layers of fully connected layers. The number of parameters is 60M, the number of neurons is 650k, and the structure won the championship in the ImageNet competition in 2012 [18]. It laid a good foundation for people to study other network structures later.

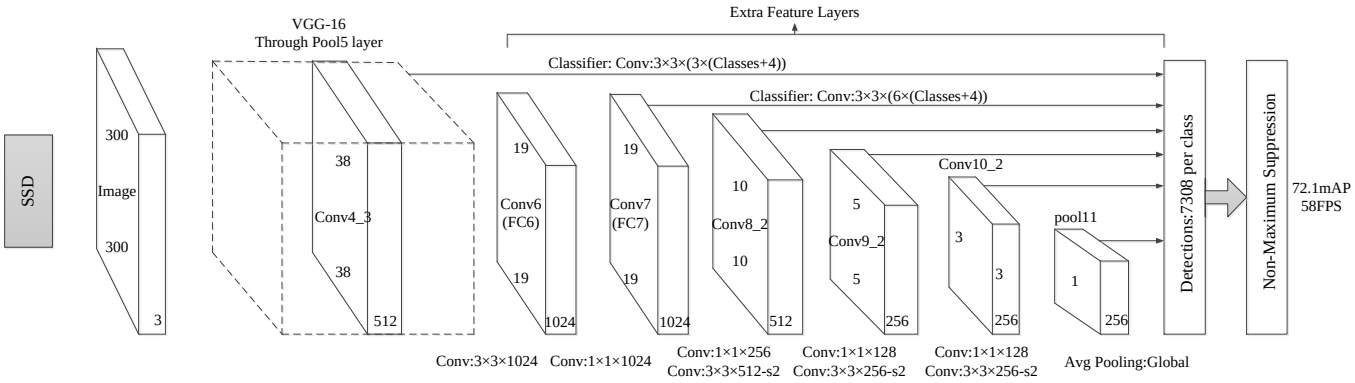


Fig.3 SSD model

Each layer of the VGG16 structure has multiple convolutional layers. Compared to AlexNet's filter7*7, it uses a smaller 3*3 as the size of the filter layer, analogous to the vector x of the n-dimensional space, Orthogonal decomposition of x:

$$x = x_1(1,0,0,\cdots) + x_2(0,1,0,\cdots) + \cdots + x_n(0,0,0,\cdots,1)$$

SSD uses multi-scale feature maps to predict objects, uses high-level feature information with large receptive fields to predict large objects, and low-level feature information with small receptive fields to predict small objects. This brings a problem: when using the low-level network feature information to predict small objects, because of the lack of high-level semantic features, the detection effect of SSD for small objects is poor. The solution to this problem is to integrate high-level semantic information and low-level detail information. The filter of each layer of each group is analogized to the base of n-dimensional Euclidean space. If a group of VGG16 network structures contains a 3-layer 3x3 filter, we assume that a 7x7 filter can be decomposed into 3 "orthogonal" 3x3 filters. While reducing the size of the filter, increase the depth to improve accuracy. Is it better to increase the number of layers and the better the network will be? Of course not, direct mapping is difficult to learn, and the paper [18] proposes a correction method: Instead of learning the basic mapping from x to $H(x)$, learn the difference between the two, ie the residual (residual). As shown in Figure 5.

Assuming the residual $F(x) = H(x) - x$, we want to get $H(x)$ an| operations at the element level. If the size of input and output is different, then you can use zero filling or projection to get the size of the match. The convolution feature layer is then added to the end of the truncated basic network, and the size of these layers decreases gradually, resulting in predictive values for multiple scale detections. The detected convolutional model is different for each feature layer. For the SSD, we modify its original VGG16 network, so that not only the projection of the target box is well-created, but also a very good classification recognition effect can be obtained. The detection accuracy and time are both an improvement.

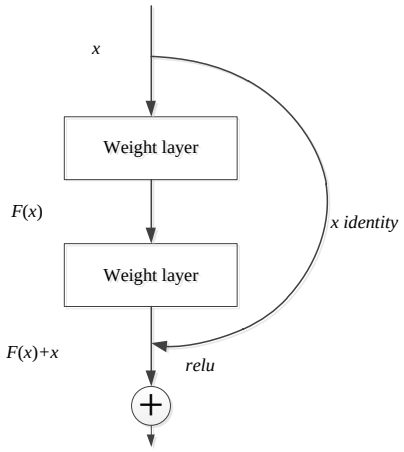


Fig.5 ResNet residual network

3.3 Migration Learning

One of the problems that sometimes may bother you is the calibration of training data. This will cost a lot of manpower and material resources. In addition, machine learning assumes that training data and test data obey the same data distribution. However, in many cases, the same distribution hypothesis is not satisfied. Usually, if the training data is out of date, the hard-to-calibrate data will be discarded, and there will be a lot of new data to be re-calibrated. The goal of transfer learning is to use the knowledge learned from one environment to assist learning tasks in the new environment. In other words, there is currently only a small amount of new marked data, but there is a large amount of old marked data (or even other categories of valid data), then by selecting valid data from these old data, adding to the current training data, training the new model [19]. Moreover, compared with traditional methods, the other advantage of transfer learning is that it can do multi-task learning. Traditional models face different types of tasks and need to train multiple different models. With transfer learning, we can realize simple tasks first, and apply the knowledge obtained from simple tasks to more difficult problems, so as to solve the problems of lack of annotation data and inaccurate annotation.

The essence of the migration learning is the transfer of weight. First, the neural network is trained by the datasets of the source task. Then the full connection layer is redesigned according to the requirement of the target task. Finally, the parameters of the frozen coiling layer are frozen, and the parameters of the full connection layer are retrained with the datasets of the target task. In view of the image recognition task, even though the difference in the content of

the different images is huge, it is all made up of the edges, texture and color in the underlying representation of the neural network. For this kind of task, the feature abstraction ability of the model is a common [20].

In this paper, according to the target detection task, modifies the target of the output layer neuron, and the weights of the random initial all connected layer neurons. At the same time, the parameters of the remaining network layer are initialized using the parameters obtained by the VOC2012 datasets pre-training, and finally the entire network is trained using the target task training set. To get the final model. The migration learning process is shown in Figure 6:

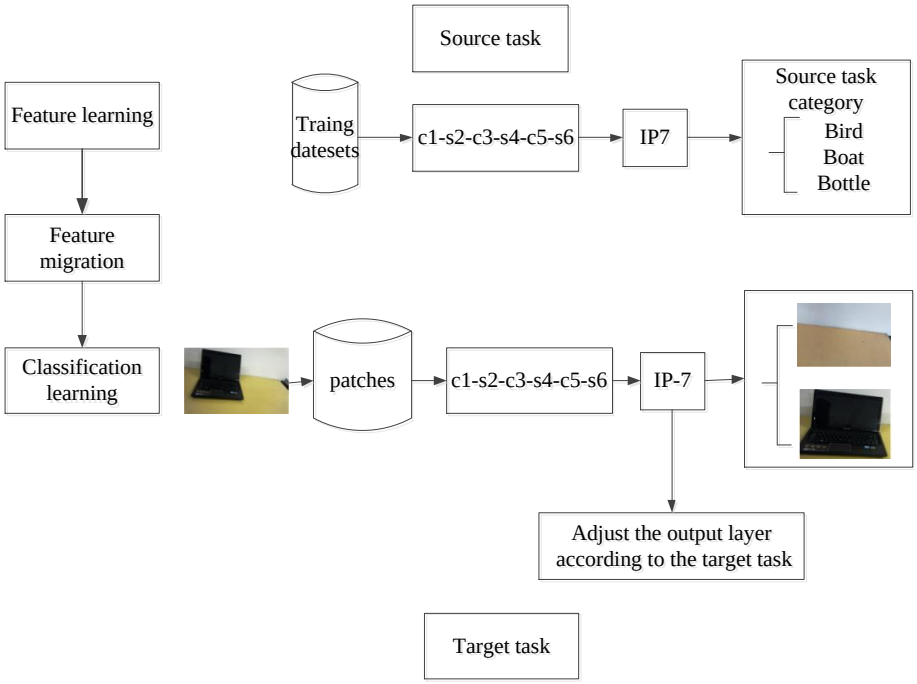


Fig.6 Schematic diagram of the migration learning process

3.4 Object Detection

In order to deal with objects of different scales, some models synthesize the results by processing images of different sizes. In fact, using the same network and feature map on different levels can achieve the same effect. The image segmentation algorithm FCN shows that the low-level feature map can improve the segmentation effect, because the low-level retains more details of the image. Therefore, in the SSD model, anchors mechanism is used to enhance the robustness of default frames to objects by using different aspect ratios for default frames on the same feature layer.

In the SSD model, the multi-scale method is adopted. The convolution receptive field at different scales is different from the convolution receptive field, and the formula for calculating the receptive field of the convolution layer is as follows[21]:

$$S_{RF}(t)=\left(S_{RF}(t-1)-1\right)N_s+S_f \tag{1}$$

$S_{RF}(t)$ the sense field of the convolution layer on the t th layer,

N_s :step length,

S_f : filter size.

At the same time, the correspondence between the coordinates of the default frame on the feature map and the coordinates of the object on the original image can be expressed by formula (2):

$$\begin{cases} x_{\min}=\dfrac{c_x+\dfrac{w_b}{2}}{w_{feature}}=\left(\dfrac{a+0.5}{|f_k|}-\dfrac{w_k}{2}\right)w_{img} \\ y_{\min}=\dfrac{c_y+\dfrac{h_b}{2}}{h_{feature}}=\left(\dfrac{b+0.5}{|f_k|}-\dfrac{w_k}{2}\right)h_{img} \\ x_{\max}=\dfrac{c_x+\dfrac{w_b}{2}}{w_{feature}}=\left(\dfrac{a+0.5}{|f_k|}+\dfrac{w_k}{2}\right)w_{img} \\ y_{\max}=\dfrac{c_y+\dfrac{h_b}{2}}{h_{feature}}=\left(\dfrac{b+0.5}{|f_k|}+\dfrac{w_k}{2}\right)h_{img} \end{cases} \tag{2}$$

The letters in the formula are expressed as:

$(x_{\min},y_{\min},x_{\max},y_{\max})$: The coordinates of the default frame are mapped onto the original image,

(c_x,c_y):The coordinates of the default box center on the feature layer,

w_b : Width of default box,

h_b : Height of default box,

w_{img} : Width of original graph,

h_{img} : Height of original graph,

$w_{feature}$: Width of feature graph,

$h_{feature}$:Height of feature graph,

$\left(\frac{a+0.5}{|f_k|},\frac{a+0.5}{|f_k|}\right)$: The center on the K level characteristic map.

In the high-level convolution layer, especially after adopting ResNet network, its feature extraction content is more abstract, and the more abstract feature extraction, the less corresponding detailed information. In addition, when the default frame is calculated according to the feature map, the proportion of the occupied image is used, and the mapping relationship between the two is mapped to the image. Therefore, when we target the image after training to get the object detection model, a sub-region operation is performed on the input image, which is equivalent to the eyeball imaging principle when the human eye observes objects at close range. In the process of detection, dividing each area for testing will result in a set of test results, and finally multiple sets of target test results. After completing this step, it is necessary to reintegrate the structures detected in these areas to obtain the final structure to be detected..

4. Experimental results

The object detection performed in the experiment, the accuracy of the classification and the accuracy of the target positioning are the standards for measuring the experiment. The correctness of the classification is measured by the confidence of the prediction frame. The accuracy of the positioning is measured by the coordinate information of the predictor. From the experiment to the object detection, we choose the confidence threshold of 0.4, and the execution of the detected objects is over 0.4. The Ubuntu16.04 system workstation is adopted in this experiment.

4.1 Candidates in the experiment

In this experiment, the output of multiple convolution layers is used to classify and position regression. If the size and position of the ground truth of each convolution layer are similar, it is possible that the selected region of the candidate box is the same bbox. In this case, the target can not be identified, so in the structure Restart the default box of the Faster R-CNN model to select the largest one, and use the default box at multiple levels for regression.

For better compatibility and understanding, the experiment uniformly describes the position of the box as [top_X, top_Y, width, height]. In the SSD paper, set the default box $width$ as shown in Equation (3), $height$ as shown in Equation (3):

$$\begin{cases} width = sqrt(scale_0 \times aspect\_ratio) \\ height = sqrt(scale_0 / aspect\_ratio) \\ ratio = 1 \end{cases} \quad (3)$$

In original paper, $scale = sqrt(scale_0 * scale_1)$, That is 1.414, and close to $scale_4 = 1.5$, is not conducive to distinguish the default box.

It is, therefore, necessary for us to take the middle of the sum directly. Furthermore, the generation of the default_box_scale can use $np.linspace$ to generate an equal-difference array, consistent with the original effect. Finally, in the experiment, the box scale was changed from [0.2, 0.9] to [0.1, 0.9], so that a minimum box area of 0.1 would be helpful for identifying objects with smaller areas.

4.2 Training of network structure

There are three main training stages: the first stage is to train a primitive SSD model; the second stage is to train deconvolution branch, freeze all the layers in the primitive SSD model.

The training objective, like the SSD model, is derived from the MultiBox object function, but this article extends it to handle multiple target categories to see if the default box matches the ground truth box. According to the above matching strategy, the total objective loss function is obtained by the weighted sum of localization loss (loc) and confidence loss (conf).

During the training, the ground truth boxes and the default boxes need to be paired. First, look for the default box that has the greatest cardoverlap with every ground truth box. In this way, every ground truth box and only one default box can be guaranteed. Correspondence. The SSD then tries to pair the remaining unread default box with any ground truth box. If the jaccard overlap between the two is greater than the threshold, the pairing is considered successful. During training, each training image is randomly selected as follows: using the original image, randomly sampling multiple patches (Crop Image), and the smallest Jaccard overlap between the object is: 0.1, 0.3, 0.5, 0.7 and 0.9, the sampling pattern is the original image size ratio is [0.3, 1.0], aspect ratio is 0.5. Or 2. When the center of the ground truth box is in the sampled patch and the ground truth box area is greater than 0 in the sampled patch, we retain the Crop Image. After these sampling steps, the patches for each sample are resized to a fixed size and flipped at a random level with a probability of 0.5.

After a series of predictions are generated, there are many predictions boxes that conform to the ground truth box, but at the same time, there are many non-ground truth boxes, and this negative box is much more than positive boxes. This will cause imbalance between negative boxes and positive boxes. It is difficult to converge during training.

Therefore, in the experiment, the predictions (default boxes) corresponding to the position of each object are sorted according to the size of the confidence of the default boxes. Choose the highest, and make sure that the ratio of negatives to positives is 3:1, so that the training results will be more stable.

4.3 Candidates in the experiment comparison of test results on datasets

In the data set of this experiment, the image is divided into two categories for experimental analysis. One is a simple background, and the object is a single example, Figure 7 shows the results; two are more complex images, objects are not single, more details of the object scene, the experimental results are shown in Figure 8.

In Figure 7, figure 7(a)，figure 7(c) ,figure 7(e) and figure 7(g) are the experimental results obtained by the classical SSD algorithm. It not only has high accuracy for small objects, but also has a large number of omissions in the existence of multiple objects in the graph. Figure 7(b), figure 7(d), figure 7(f) and figure 7(h) are the algorithm used in this experiment. First, it can be seen that the use of the ResNet network model and Inception after the classification of objects has greatly improved, and second we can see that for similar objects, the classical SSD network model is still very poor in the recognition of objects and the recognition of many small objects.

The reason is that the size of the calibrated object in the picture is too small, so Figure 7(b), figure 7(d), figure 7(f) and figure 7(h) are obtained at the multi-view angle combined with the segmented picture area. The object detection is more perfect. Moreover, the direct regression of the object shape and classification object category, rather than using two decoupling operations, makes the model can better locate the object and reduce the positioning error.

Fig.8 Experiments on complex scenes show that objects in the scene are no longer single. When we use this experiment to detect objects, we can not only detect more obvious objects, but also detect small objects in detail. Fig. 8 (i) is two kinds of food, can be seen in Fig. 8 (k) and Fig. 8 (m) are the scene of the athletes, in this case, we need to detect not only the athletes in the image, but also the items of the athletes, according to the tools or sports equipment they use to distinguish, relative and transmission. For the traditional target detection SSD algorithm, this experiment can detect the players, but also can detect the equipment used by the players, so that we can judge the type of athletes described, the players in Figure 8(l) are tennis players, the players in Figure 8(n) are baseball players. In Figure 8(o) and Figure 8(p), the objects in the scene are more complex, but it can also be seen that compared with the original SSD algorithm, this experiment can be seen through confidence in the detection of the object accuracy has been improved, in the target location is more accurate.

In summary, the experimental method has the following advantages compared with the traditional SSD target detection method.

1、In the same scenario, more objects can be detected;

2、For identifying the same object, the confidence level is higher;

3、The accuracy of the object similar to the scene is higher, and the false detection rate is less.

4.4 Comparison of object detection accuracy

First, analysis of object retrieval ability.For each graph, the target-based quantitative analysis of the retrieval ability of each graph to evaluate the detection ability of the entire system. The retrieval ability is generally expressed by F value, which is the weighted average of precision and recall rate. The expression of F value is.

$$F = \frac{\left(\gamma^2 + 1\right) \times P \times R}{\gamma^2 \times \left(P + R\right)} \tag{4}$$

Formula letter meaning:

P : Precision,

R : Recall rate,

γ : Weight: in general, $\gamma = 1$.

Through the analysis of the image detection results, most of the target retrieval results are better. If the original SSD algorithm cannot detect the image sequence, or cannot meet the IoU, the result of this experiment is 0, and meets the recall rate and accuracy requirements, so it has a good detection ability. Next, the accuracy of object detection is analyzed. According to the results of the experiment, the target retrieval of each image is achieved to a certain extent, and most of the target retrieval is higher than the original SSD model.

Using the object detection method of this experiment, for the 100 images that need to be verified, the confidence thresholds chosen are all 0.4. By calculating the mAP on the classical SSD algorithm and the target detection algorithm in this paper, we can draw the following table 1.

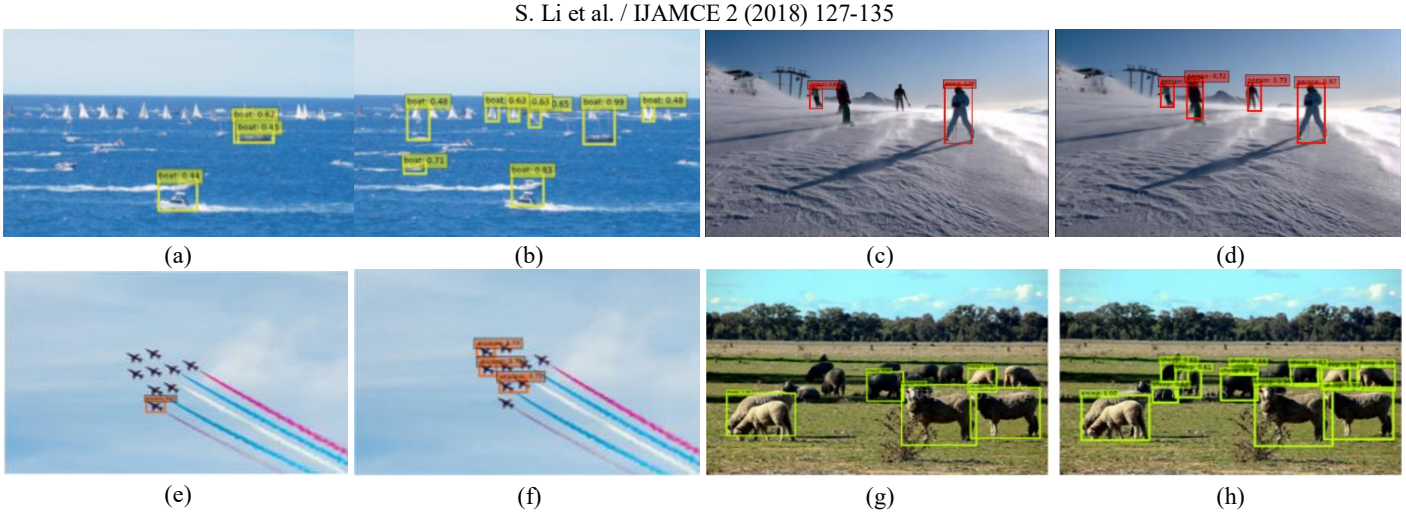
It can be seen that the detection of this experiment is 0.131 higher than that of the SSD algorithm. Therefore, the optimized algorithm has improved the retrieval ability and accuracy of the target detection.

Table 1 object detect result

method	mAP	Bird	boat	bottle	car	chair	computer	headset	person	plant	mouse
SSD	0.513	0.612	0.620	0.182	0.721	0.447	0.359	0.502	0.536	0.288	0.532
NSSD	0.644	0.662	0.729	0.228	0.851	0.601	0.543	0.545	0.592	0.452	0.567

5. Conclusions

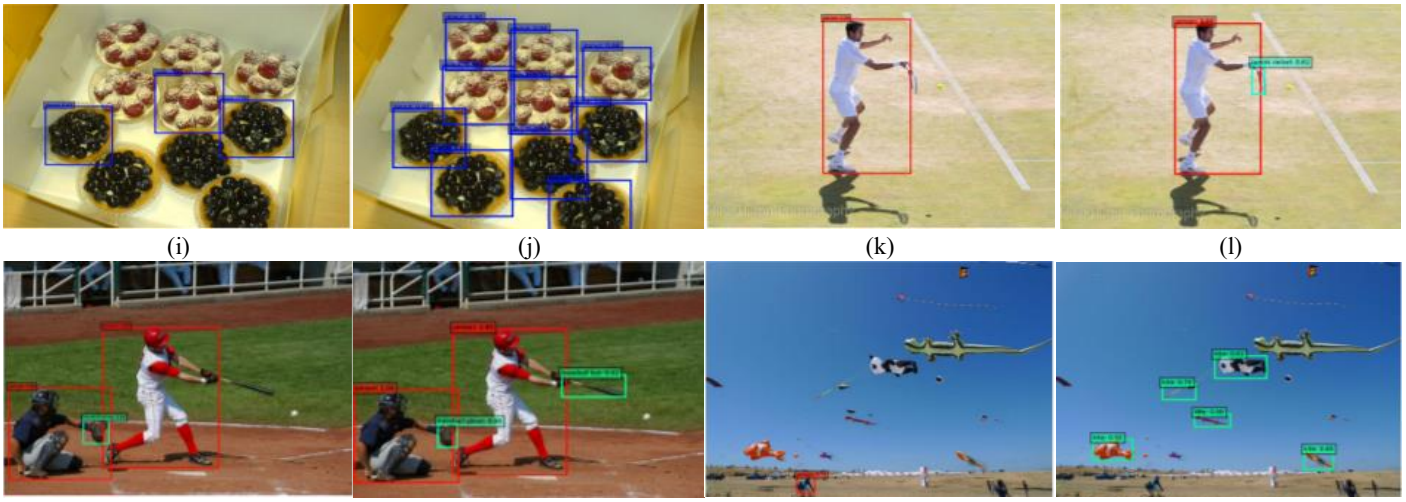
This article uses deep learning to detect objects and optimizes SSD algorithms to obtain experimental results. First of all, in order to achieve a higher classification effect, the network structure in the original SSD model is replaced. The ResNet and Inception structures are the parameters with higher utilization. This model with less parameter training not only improves the accuracy of classification, but also effectively reduces the over fitting phenomenon. Secondly, in order to solve the phenomenon of lack of sample datasets, migration learning is introduced, which can reduce the resource consumption and improve the training speed. It can also be seen that the pre-training model has powerful generalization ability and transplantation ability. Finally, we segmented the image by a part of the input image by segmenting the input image, and then combined the concepts of the convolutional receptive field and the mapping between the default frame of the feature layer and the input image. Ultimate, the results of these regional detections were merged to obtain the results of government image inspections. Through this experiment, not only has the problem of insufficient data set been solved, but also the accuracy of detection has been improved, and the problem of missed detection on the way has been greatly reduced. In the following work, the model will be further improved, its sharing mechanism will be enhanced, and its timeliness will be improved. At the same time, image global information will be further improved to improve the performance of the algorithm under the conditions of multiple image divisions.



(a) (b) (c) (d)

(e) (f) (g) (h)

Fig. 7 Comparison of detected objects, (a) (c) (e) (g) is a classical SSD algorithm detection, and (b) (d) (f) (h) is an experimental algorithm detection.



(i) (j) (k) (l)

(m) (n) (o) (p)

Fig. 8 Comparison of detected objects in complex scenes, (i) (k) (m) (o) is a classical SSD algorithm detection, and (j) (l) (n) (p) is an experimental algorithm detection

References

Smith, L. A., & Bull, J. M. (2001). A parallel java grande benchmark suite. Supercomputing, ACM/IEEE 2001 Conference (pp.8-8). IEEE.

Li Xudong, Ye Mao, Li Tao. Research Review of Target Detection Based on Convolutional Neural Network[J]. Journal of Computer Applications, 2017, 34(10):2881-2886+2891.

Wang Zhen, Gao Maoting. Design and implementation of image recognition algorithm based on convolutional neural network [J]. Modern Computer (Professional Edition), 2015 (20): 61-66.

Huang Lin. Research on traffic sign recognition based on deep neural network [D]. Jiangsu University of Science and Technology, 2015.

Wang Ruochen. Research on target detection and segmentation algorithm based on deep learning [D]. Beijing University of Technology, 2016.

Ma Dezhi, Li Bajin, Dong Zhixue. Research on moving object detection method based on Gaussian mixture model[J]. Electronic Measurement Technology, 2013, 36(10):47-50.

ZHOU Jianying, WU Xiaopei, ZHANG Chao, Lü Miao. Moving object detection method based on sliding window for Gaussian mixture model[J]. Journal of Electronics and Information Technology, 2013, 35(07): 1650-1656.

Tian Penghui, Zhai Lichun, Yan Sha. A survey of detection methods for small objects in infrared motion[J]. Journal of Detection and Control, 2013, 35(02): 76-80.

Hu Changyu. Research on target detection algorithm based on convolutional neural network [D]. Harbin University of Science and Technology, 2017.

Chen Tuo. Stereo Matching Technology Based on Convolutional Neural Network [D]. Zhejiang University, 2017.

Cao Linlin, Li Haitao, Han Yanshun, Yu Fan, Gu Haiyan. Application of convolutional neural network in high-resolution remote sensing image classification[J]. Surveying Science, 2016, 41(09): 170-17

Dalal N, Triggs B. Histograms of oriented gradients for human detection [C]//Computer Vision and Pattern Recognition, IEEE Computer Society Conference on, IEEE, 2005, 1: 886-893.

Fu Ruonan. Research on target detection based on deep learning[D]. Beijing Jiaotong University, 2017.

Girshick, R., Donahue, J., Darrell, T., Malik, J.: Rich feature hierarchies for accurate objectdetection and semantic segmentation. In: CVPR. (2014)

Redmon, J., Divvala, S., Girshick, R., Farhadi, A.: You only look once: Unified, real-timeobject detection. In: CVPR. (2016)

Faster R-CNN: Towards real-time object detection with region proposal networks. In: NIPS. (2015)

Liu W, Anguelov D, Erhan D, et al. SSD: Single shot multibox detector [C]//European Conference on Computer Vision, 2016: 21-37ImgL

Pan Rong, Sun Wei.Deep learning target detection based on pre-segmentation and regression[J].Optical Precision Engineering, 2017, 25 (10s): 221-227.

Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition[C]//ICLR, 2015.abel

K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. arXiv preprint arXiv:1512.03385,2015.

Jiang Tao. Target detection algorithm for convolutional neural networks based on migration learning [A]. Committee of Control Theory of China Association of Automation. Proceedings of the 36th China Conference on Control (G) [C]. Control Theory, China Institute of Automation. Committee: 2017:6.

Huang Jiabiao. Pedestrian detection method based on migration learning [D]. Nanjing University of Science and Technology, 2015.

Hao Yelin, Luo Bing, Yang Rui, et al. Improvement of human target detection algorithm in complex scene images [J]. Journal of Wuyi University (Natural Science Edition), 2018 (1).

Tang Cong, Ling Yongshun, Zheng Kedong, Yang Xing, Zheng Chao, Yang Hua, Jin Wei. Multi-window SSD target detection method based on deep learning[J]. Infrared and Laser Engineering, 2018, 47(01): 302-310 .

Li Shengwang, born in 1963, master's degree, professor, master tutor. He successively presided over and completed more than fifty items of scientific research and new products. Created a greater social and economic benefits, of which, five achievements through the provincial, ministerial, municipal identification, one by the Hebei Province Science and Technology Progress Award, one won the national patent. He published more than twenty papers and published academic papers in core journals. Four papers were indexed by three major indexes. The third-prize-winning "Pulverized Coal Pre-homogenization Process Intelligent Control System" project was widely used in Pakistan and in China. The main research direction: computer monitoring and control. It relates to hardware, application software, database, pattern recognition, artificial intelligence, information fusion technology, operation platform, data communication, network, virtual instrument, embedded system, electrical control, measurement and other functional hardware of the computer measurement and control system.

Chen Li is currently studying for a master's degree in computer technology at Hebei University of Science and Technology , Shijiazhuang, China. The main research direction is pattern recognition and image processing.

Hou Yifan is currently studying for master's degree in computer technology at Hebei University Of Science and Technology, Shijiazhuang, China. The main research directions are image processing and 3D reconstruction.