

Research on Daily Objects Detection Based on Deep Neural Network

Sheng Ding^{1,*}, Kun Zhao²

¹Key Laboratory of Fiber Optic Sensing Technology and Information Processing,
Ministry of Education, Wuhan University of Technology, Wuhan, China

²Fiberhome Telecommunication Technologies Co.,Ltd, Wuhan, China

*Corresponding author e-mail: dingsheng@whut.edu.cn

Abstract. With the rapid development of deep learning, great breakthroughs have been made in the field of object detection. In this article, the deep learning algorithm is applied to the detection of daily objects, and some progress has been made in this direction. Compared with traditional object detection methods, the daily objects detection method based on deep learning is faster and more accurate. The main research work of this article: 1. collect a small data set of daily objects; 2. in the TensorFlow framework to build different models of object detection, and use this data set training model; 3. the training process and effect of the model are improved by fine-tuning the model parameters.

1. Introduction

Object detection is a very popular research direction in the vision field. Launched in 70s, object detection began to be on track until 90s when computers became powerful and application plentiful. It is easy for us as human to recognize objects in the images, however, things become difficult for computers. Adding the different posture of objects and the complex environment around, object detection is more ambiguity.

As we know, the evolution of detection algorithm is divided into two stages. Stage one is based on the traditional features of the solution, and the second stage is the deep learning algorithm. Before 2013, most of the researches was based on the traditional feature optimization detection method. After that, both academia and industry turned to deep learning algorithm.

With the increasing amount of detection data, the traditional detection method performance will become saturated. The detection performance will gradually improve, yet the improvement decreases after a certain amount of data. However, the method of deep learning is different. While the data of the scene distribution accumulates, the detection performance promote continuously.

In this article, a set of data of daily supplies is collected, and then different training object detection models are applied on the data. And by comparing the direct training and parameter adjustment model training, it will be proved that the convergence speed and accuracy of object detection are improved by adjusting the parameters.

2. Literature Survey

In ILSVRC 2014, deep learning increases the average object detection rate to 43.933%.

R-CNN proposed by Ross Girshick introduced CNN method into target detection field for the first time. It introduces CNN method into target detection field. Selective Search[5] window extraction



algorithm is proposed to replace the traditional sliding window extraction method. Girshick proposed Fast R-CNN model, which integrates feature extraction and classification into a classification framework. In the R-CNN model, the deep convolution network for feature extraction and support vector machines for classification are trained separately. The training time of Fast R-CNN model is 9 times faster than that of R-CNN model. The region proposal extraction and part of Fast-RCNN are put into a network model RPN (region proposal net) in Faster R-CNN. The detection stage is very convenient and the accuracy is similar with Fast R-CNN. YOLO (You Only Look Once) [6] proposed by Joseph Redmon et al., is a one-time convolutional neural network predict multiple candidate frame position and classification, target detection and recognition can achieve end to end. It solves object detection as a regression problem. Based on a single end-to-end network, the output from the original image to the position and category of the object is completed. R-FCN [7] is an accurate and effective method for object detection. Compared with the previous region detection, R-FCN is based on the whole image convolution calculation. In order to achieve this goal, the model uses a position-sensitive score maps to balance the shift invariance in image classification and the translational transformation in object detection.

3. Proposed Method

Object detection algorithm usually contains three parts. the first is the design of features, the second is the choice of detection window, and the third is the design of classifier. Feature design methods include artificial feature design and neural network feature extraction. The selection of detection window mainly includes: Exhaustive Search [4], Selective Search [5], and RPN method based on deep learning. This article adopts deep convolutional neural network (CNN) image feature extraction, using the most advanced RPN as the detection window selection method, the bounding box regression analysis, using softmax classification processing, and output the detection result. The model structure is shown with the help of blocks in Figure 1.



Fig.1 Object Detection Model

3.1. CNN Feature Extraction

At present, the image feature extraction using CNN mainly includes three steps: convolution, activation and pooling.

Convolution is the core of feature extraction, and the feature of the image is obtained by different convolution kernels. Convolution kernel is equivalent to a filter, and different filters extract different features. Convolution is used to cross the convolution of the image with the size of 3*3 and the number of 2ⁿ. The activation function generally uses ReLU to perform non-linear operation on the output of the stacked layer, and has the function of accelerating convergence. Formula is as follows:

$$y = f\left(\sum_{j=0}^{j-1} \sum_{i=0}^{i-1} x_{ij} w_{ij} + b\right) \quad (1)$$

where x represents the input vector, w represents the parameters of a convolution kernel, b represents the bias term, f represents the activation function, and y represents the output.

A pooling layer is placed behind each roll layer to reduce dimension. Generally, the output matrix size of the original convolution layer is changed to half of the original one, which is convenient for the operation at the back. In addition, the pooling layer increases the robustness of the system, turns the original accurate description into a rough description, and avoids overfitting to some extent.

3.2. Region Proposal Networks

Region Proposal Networks (RPN) take the feature map extracted from the upper CNN as the input of this layer, maps the midpoint of the feature map back to the original image, and designs these different

fixed scale windows in the original design. According to the window and the ground truth Intersection-over-Union (IoU) value to its positive and negative labels, let it learn whether there are objects inside, so training a Region Proposal Network.

Only the approximate place need to be found, because that the precise positioning position and size can be accomplished by following works. As the consequence the anchors can be fixed in three aspects: fixed scale changes (three scales), fixed length and width ratio changes (three ratio), fixed sampling method, only in the eigenvalues of each point in the original map of the corresponding Region of Interest(RoI) on the sampling, in the back of the work can be adjusted. This can reduce the complexity of the task.

After extracting the proposal on the feature map, the convolution calculation can be shared in front of the network. The result of this network is that each point of the convolution layer has an output about the k anchor boxes, including whether it is an object or not, adjusting the corresponding position of the box. The RPN's overall Loss function can be defined as:

$$L(\{p_i\}\{t_i\}) = \frac{1}{N_{cls}} \sum_i L_{cls}(p_i, p_i^*) + \lambda \frac{1}{N_{reg}} \sum_i p_i^* L_{reg}(t_i, t_i^*) \quad (2)$$

where the i denotes the i-th anchor, $p_i^* = 1$ when the anchor is positive, and $p_i^* = 0$ when the anchor is negative. t_i^* represents a ground true box coordinate associated with the positive sample anchor (each positive anchor may correspond to only one ground true box: a positive sample anchor corresponds to a ground true box, then the anchor with the ground true box is either the largest IoU of all anchor, or greater than 0.7).

3.3. RoI Pooling

For traditional CNN, when the network is trained, the input image size must be a fixed value, and the network output is also a fixed size vector or matrix. If the input image size is uncertain, the problem becomes more cumbersome. There are two kinds of solutions: 1. Cut from the image part of the incoming network; 2. Zoom the image into the desired size and then into the network. But the clipping will destroy the complete structure of the image, and the zoom will destroy the original shape information of the image. So Faster R-CNN proposed RoI Pooling to solve this problem.

First, the region proposal is mapped back to the original feature graph scale, and then each proposal level and vertical are divided into k copies, each of which is max pooling processing. After this processing, even if the size of the proposal, the output is k * k size, to achieve a fixed-length output.

3.4. Classification and regression

Using the region proposal feature map that has been obtained, the probability vector of each proposal is calculated by the full connection layer and the softmax layer. At the same time, the bounding box regression is used to obtain the position offset of each proposal again for the more accurate target detection box.

4. Experimental Results

4.1. Data Set

There are many data sets in the field of object detection, such as the Oxford-IIIT Pets data set, the PASCAL VOC data set, the Microsoft COCO data set. This article collects a small daily object dataset, including the selection of 20 categories of items that are frequently touched in everyday life: backpacks, handbags, umbrellas, beds, cups, chairs, sofas, potted plants, laptops, cell phones, books, Clocks, tables, toothbrushes, bottles, bananas, apples, bikes, cars and trucks.

we collect pictures of the 20 categories on the web, collect 100 pictures for each category, a total of 2000 images, and use the marking tool to mark. In 100 pictures, 80 pictures as a training set, 20 pictures as a validation set. This data set will be used in the next experiment.

4.2. Experimental Environment

This experiment is based on the TensorFlow framework. TensorFlow is a deep learning library supported by Google. The workflow is relatively easy, the API is stable and the compatibility is good. One of its highlights is that it supports distributed computing of heterogeneous devices. It can be used on various platforms Automatically run models, from mobile phones, single CPU / GPU to hundreds of thousands of GPU cards consisting of distributed systems. This experiment runs on google cloud machine learning platform, using 8 NVIDIA Tesla K80 GPU for parallel computing, greatly accelerating the model convergence speed.

4.3. Experiment

In this article, the daily object detection data set is trained with different object detection models, and the results of object detection under different models are obtained. The network structure is shown in the following figure:

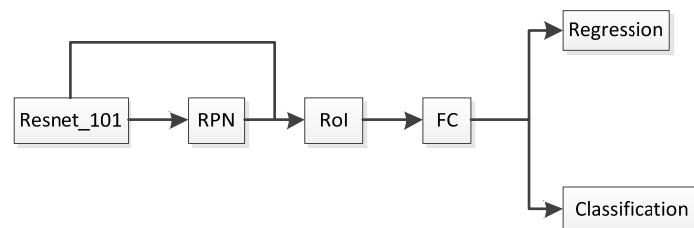


Fig.2 Faster_rcnn_resnet_101 Model

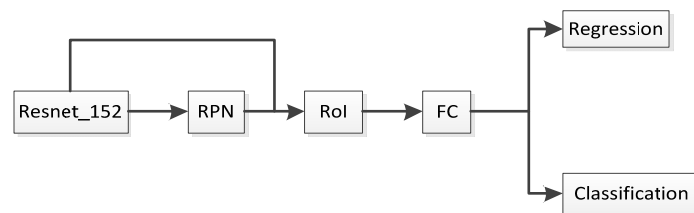


Fig.3 Faster_rcnn_resnet_152 Model

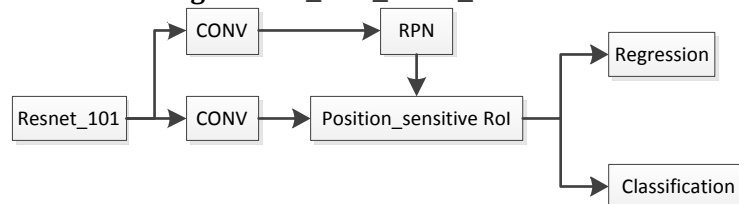


Fig.4 R-FCN_resnet_101 Model

The network structure of each model is shown in Fig.2~Fig.4. Resnet_101[8] or Resnet_152[8] represents the use of this convolutional network structure for feature extraction. RPN represents the area to generate the network, CONV represents the convolution layer, FC represents the full connection layer, and Position_sensitive RoI [7] represents the location sensitive pooling operation.

Table.1 Mean Average Precision Table

model	mAP(%)
Faster_rcnn_resnet_101	69.5
Faster_rcnn_resnet_152	72.3
R-FCN_resnet_101	69.8
Finetune Faster_rcnn_resnet_101	70.1

Table 1: Through the above experiment, we use the same data set to train different models, and get the corresponding model of the mean average precision. It can be seen that the model trained by the daily item identification data set is highly accurate in the test. But also can find the deep of the neural

network, structure, parameter settings affect the accuracy of object detection. In the case of parameter setting, the number of convolution kernel sizes, the learning rate setting, the regularization mode and the gradient descent algorithm type play a key role in the final convergence rate and effect of the model, and it is necessary to try to get good results, By repeated experiments, fine-tuning parameters obtained by the model, compared with the original model accuracy is improved.

5. Conclusion

This article mainly established a small daily items detection data set. The data set is then trained on different object detection models and has achieved good results in daily object detection. These well-trained models can be used in the mobile platform, Nao robot platform or other intelligent devices, the daily items to achieve accurate detection. In the future, we can get a better model by increasing the capacity of the data set, the optimization of the model structure and the fine tuning of the parameters.

References

- [1] Girshick R, Donahue J, Darrell T, et al. Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation[J]. 2013:580-587.
- [2] Girshick R. Fast R-CNN[J]. Computer Science, 2015.
- [3] Ren S, He K, Girshick R, et al. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks.[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2015, 39(6):1137.
- [4] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ra-manan. Object detection with discriminatively trained part based models. TPAMI, 32:1627–1645, 2010.
- [5] Sande KEAVD, Uijlings JRR, Gevers T, et al. Segmentation as selective search for object recognition[C]// IEEE International Conference on Computer Vision. IEEE, 2012:1879-1886.
- [6] Redmon J, Divvala S, Girshick R, et al. You only look once: Unified, real-time object detection[C]. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016: 779 -788.
- [7] Dai J, Li Y, He K, et al. R-FCN: Object Detection via Region-based Fully Convolutional Networks[J]. 2016.
- [8] He K, Zhang X, Ren S, et al. Deep Residual Learning for Image Recognition[C]// Computer Vision and Pattern Recognition. IEEE, 2016:770-778