Object detection based on SSD-ResNet

Xin Lu[1], Xin Kang[2], Shun Nishide[2], Fuji Ren[2]

[1]Tokushima University, Japan
[2]Faculty of Engineering, Tokushima University, Japan
[1]c501737063@tokushima-u.ac.jp, [2] {kang-xin,nishide,ren}@is.tokushima-u.ac.jp

Abstract: Nowadays, with the abundance and diversity of the data sets of the detected objects, detection and recognition technology has achieved excellent performance in learning effect. However, because the target objects are usually very small in many real-world applications while the background environment seldom varies, manually annotating these objects is extremely costly in time and manpower. These problems have challenged the learning effect of standard neural networks. In this paper, we propose a novel method to replace the original network structure and to extend the number of layers for detecting many kinds of dangerous goods among different background environments. Specifically, we employ SSD as the basic network structure and replace the inside VGG16 with a ResNet101 network. The experimental results show that the ResNet network is effective in detecting many kinds of dangerous objects in small data sets. The proposed model outperforms other neural networks in learning efficiency and accuracy.

Keywords: SSD; ResNet; CNN; Object detection and recognition

1 Introduction

The problem of computer detection and classification in different fields is the most concerned hotspot at present. Limited by the environment, variety, size and other problems of the detected objects, there are obvious deviations in computer learning of object features. With the popularization of face recognition technology and the rise of unmanned vehicle technology, recognition technology has been applied in all fields around us. From the perspective of computers, the final result of recognition is only to do with the probability. It still does not have the human understanding mode and the humanized processing mode. So once the computer misjudges, it will often cause huge economic losses or traffic accidents to the society and individuals. Therefore, people must have strict high standards and high requirements for recognition technology, especially in accuracy.

Object detection technology mainly depends on detectors from traditional DPM [1], OverFeat [2], SPPNet [3], Rcnn [4], Fast-Rcnn [5], Faster-Rcnn [6], YOLO [7], YOLO2 [8], YOLO3 [9] to SSD [10]. By gradually optimizing the detector, the accuracy and efficiency in the field of recognition are continuously improved. The earliest DPM had a huge amount of computation. As for OverFeat, Alex-net was improved and CNN [11] was applied to the sliding window on the image pyramid, and an image location method was proposed to realize the synchronization of classification, location and detection. SPPNet through spatial pyramid pooling makes the features of CNN no longer a single scale. After that, Rcnn uses region proposal to get a certain number of detected objects, gets the feature of the region through CNN, and then judges whether it belongs to a certain kind of object or background by classifier. YOLO3 treats objects as regression problems, so bounding box and category probability can be carried out directly through CNN input images. SSD combines with YOLO3 to make up for its shortcomings. Through recognition and detection under feature map of different levels, more diverse of object detection can be realized.

YoLo3 model has advantages of excellent precision and speed in object detection, but the effect of target detection for small objects is not ideal. At the same time, it also has the problem of low recall rate. SSD model uses conv4-3 as a very low feature to detect small objects, resulting in fewer layers of winder and insufficient feature extraction. If the basic network of VGG16 is deepened. There are also some problems such as network degradation and gradient dispersion. For the detection of objects in specific areas (such as pistols, cutters), it is often observed from different angles, and there is a huge difference among the detected objects. Moreover, the volume of the detected object is very large, but because of the distance and the problem of being covered, there are obvious problems in the learning and recognition of the model. Therefore, the general model has obvious problems in multi-object detection with different sizes in specific areas.

In this paper, based on ImageNet image database, we first apply excellent object detection models You Only Look Once (YOLO3) and Single Shot MultiBox Detector (SSD) to detect objects. By comparing the recognition results of YOLO3 and SSD models, this paper uses SSD model which is more suitable for solving multi-classification problems. The method of deepening the network is used to solve the problem of poor learning effect caused by insufficient feature extraction. On the basis of the SSD model, the original VGG16 model is replaced by the ResNet101 [12] network structure. Compared with the previous network structure, it not only deepens the structure, but also avoids it. The problem of network degradation caused by VGG network deepening is overcome. The accuracy of VGG network exceeds the original SSD model and YOLO3 model.

978-1-7281-3863-3/19/$31.00 ©2019 IEEE

The contributions of this article are:

1.Train and apply state-of-the-art detection techniques for the detection and identification of hazardous objects.

2. It improve SSD model and adds ResNet to make it more effective in multi-classification and small data set learning.

The rest of this paper is arranged as follows: Section 2 describes the related work. Section 3 mainly makes a detailed description of the improved model based on SSD-ResNet.Section 5 shows the experiments and draws the conclusion.Finally, Section 6 is a summary of this article and an overview of the issues to be solved.

2 Related Work

The research of computer vision can be divided into four areas. First, Image classification, which is given a group of images labeled as a single category, and then predicting the category of a new set of test images. Second, object detection. The task of recognizing objects in images usually involves outputting boundary boxes and labels for each object. Third, target tracking, the process of tracking one or more specific objects of interest. Fourth, semantic segmentation, which divides the whole image into groups of pixels, and then marks and classifies them. The purpose of this paper is to detect and identify dangerous goods. It belongs to the task of object detection which aims at detecting objects quickly and accurately.

In recent years, the convolution neural network will recognize every object in the image as an object or background. So it is necessary to use the convolution neural network in a large number of locations and scales. The result is carrying a huge amount of computation in the research of object detection. In order to solve this problem, people introduce the concept of region. By searching the "speckle" image region of possible objects, we can reduce unnecessary searching region. The most classical model is the R-CNN family. Faster R-CNN is a typical case based on deep learning object detection. By inserting the Regional Proposal Network (RPN), recommendations from features are predicted. RPN decides where to check. It reduces the computational complexity of the entire network reasoning process. It also means that a fast neural network is used to replace a selective search algorithm with slow operation speed. But compared with Faster R-CNN two-stage processing, YOLO and SSD transform the detection problem into a regression problem. By regression, both coordinates and probabilities of each kind are generated, which greatly accelerates the recognition speed. SSD can be divided into grids on different feature maps and then uses RPN-like regression to take all sizes of objects into account the scope of detection, and effectively applies these detection on the corresponding output feature map. This paper presents a method of replacing VGG16 and resnet, the basic network of SSD, to detect dangerous goods. This method solves the problem of insufficient recognition effect when detecting multi-objects of different sizes.

3 SSD-ResNet method

SSD model is selected in this paper to detect dangerous goods in various complex environments. In order to improve the learning effect of features and improve the precision, a similar framework solution based on SSD model is proposed. The most obvious difference is that the VGG16 network framework of SSD is replaced by ResNet101 network framework. Through the characteristics of ResNet network, which is feasible to build a stackable layer network, the problem of network degradation is solved. Also, the efficiency of detection can be improved by adopting the network overlaying method.

Degradation problem always exists in deep neural networks. This change is not caused by over-fitting. As a result, when the network is deepened through the hierarchy, the performance will be saturated, which leads to the decrease in both the accuracy and the test accuracy, resulting in degradation problem. ResNet uses this leaping structure as the basic structure of the network.

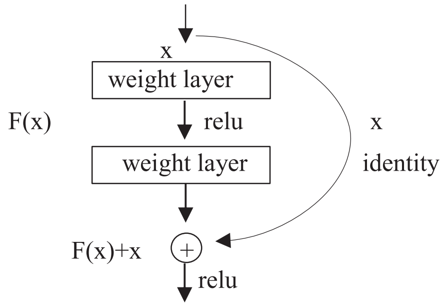

Figure 1 ResNet: a buiding block.

This kind of structure has two layers, the following expression, where σ represents the non-linear function ReLU [13].

$$\mathrm{F} = \mathrm{w}_{2\sigma}(\mathrm{w}_{1^{x}}) \tag{1}$$

Then through a short cut and the second ReLU, the output y is obtained.

$$\mathrm{y} = \mathrm{F}(\mathrm{x}, \{\mathrm{w}_{\mathrm{i}}\}) + \mathrm{x} \tag{2}$$

In terms of a stack or a structure formed by several stacks, when the input is x, the feature it learns is H (x), hoping that it can learn the residual F (x) = H (x) - X. Its original learning feature is F (x) +x. When the residual is 0, only identity mapping is carried out on the stack, at least the network performance will not decline. In reality, the residual will not be 0. It enables the stack to learn new features on the basis of the input feature, so as to have better performance. It’s expressed by formula as follows:

$$\mathrm{y} = \mathrm{F}_{(\mathrm{x},\{\mathrm{w}_{\mathrm{i}}\})} + \mathrm{w}_{\mathrm{s}}\mathrm{x} \tag{3}$$

The SSD model adopts VGG16 network structure because the entire network uses the same convolution core size (3*3) and maximum pooling size (2*2), which makes its structure very simple. Initially, we tried to increase the depth of the network, through multi-layer feature extraction, to learn more features in complex environments, thus improving the precision and accuracy. However, the author found through the experiment that the deep-seated network has the degradation problem: when the depth of VGG16 network increases, the accuracy of the network has a problem of saturation and decline. Therefore, aiming at such problems, the author replaces VGG16 network with ResNet101 network structure in this paper. ResNet is adopted because when the size of feature map is halved, its number doubles over the same period, thus maintaining the complexity of the network. In this way, the problem of saturation and decline of accuracy can be solved when the network depth is increasing. To further deepen the depth of the network and improve the accuracy, in this paper, the network framework of SSD model from VGG16 network to conv8-2 is replaced by ResNet network framework. The last three layers use the original convolution layer of SSD to get the feature map for detection, as is shown in Figure 2:

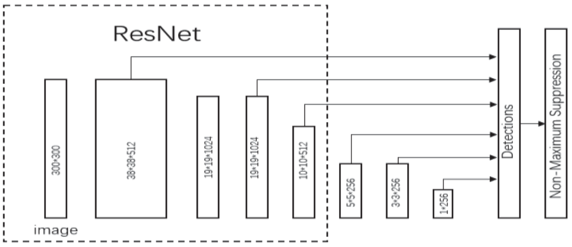

Figure 2 Model architecture

Network structure: ResNet101 is adopted as the basic model. Conv3_block4_out layer of resnet is used as the first feature map for detection. The original Conv7 of SSD is used as the second feature map by conv5_block3_out layer of resnet. Conv8_2 is filled by the boundaries of conv5_block3_out layer through the reel machine. Conv6_2 layer is adopted as the third feature map, followed by conv7_2, conv8_2 and conv9_2 in order after that. Six feature maps were extracted from the original size of SSD, which were (38,38), (19,19), (10,10), (5,5), (3,3), (1,1). After the feature maps were obtained, it was necessary to convolute the feature maps to get the detection results. The detection values consisted of two parts: class confidence and boundary frame location, each completed by a 3*3 convolution. Since each priori frame predicts a boundary frame, SSD300 can predict 8732 boundary frames, which is essentially a method of intensive sampling.

4 Experimental

The purpose of the experiment is to compare the SSD network and SSD-ResNet, and the difference between the deepened network structure and the original SSD on the dangerous goods data set. As shown in Figure 1, the data is selected from the ImagNet2011 [14] data set. As a training data, a total of 2767 pictures were selected for training. The VOC [15] structure method was used to preprocess the pictures. Mainly through Mean. Average Precision, or mAP., is the main evaluation indicator.

Figure 3 Classification sample

First, we compared the two models of YOLO and SSD in the inspection of dangerous goods. As shown in Table 1, YOLO is about 53.40% and SSD is about 55.00%. So it proves that SSD is better than YOLO in the accuracy of multi-class object detection. After the network, we compare the original resnet network with the original SSD on the ImagNet2011 dataset. As shown in Table 2. The SSD uses the original VGG16 network. The accuracy is about 55.00%, and the lowest is the identification of the rifle. At 41.80%, the highest is the identification of the long whip, which is about 76%. After using the resnet network, the recognition accuracy of the map is improved from 72.40%. The recognition rate of the previous rifle and long whip is increased to 69.10% and 81.00% respectively.

Table 1 The objective evaluation of different methods

ImagNet2011	
MODEL	mAP
YOLO3	0.530
SSD	0.550
SSD-ResNet	0.724

Table 2 Model improved experimental comparison results

Category	SSD	SSD-ResNet
Knife	0.445	0.696
Revolver	0.736	0.868
Long whip	0.757	0.810
Rifle	0.418	0.691
Bullet	0.531	0.662
Missile	0.477	0.695
Bow	0.486	0.647

By comparing the SSD model with YOLO3 and SSD-ResNet, compared to the original SSD model, the number of SSD-Res network layers increased to 85, and the accuracy increased by 17%. This experiment used a CPU of 3.40GHz, 6 CPU cores, 32G RAM, GeForce GTX TITAN X, 12G memory, and 64-bit centos. The obvious difference is mainly reflected in the fact that the ResNet network can extract more representative features when extracting scenes with small objects or higher complex environments.

5 Conclusions

In recent years, object detection has gradually entered the bottleneck, and more and more high-precision detection methods are two-stage detection models. First, based on the image, extract the area of the object that may be included. This was followed by running the best performing classification network on the proposed areas to obtain the categories of objects in each area. However, for complex environments, when the size of objects is different, and the amount of data is insufficient, the existing models are not satisfactory. This paper proposes a new scheme based on SSD model and ResNet network architecture to compare the classification of dangerous goods in VOC. Experiments show that ResNet-based SSD can improve the accuracy of about 17.40% in a deepening network. Compared with the original VGG16 network architecture, SSD-ResNet is significantly better than the original model in accuracy, but the amount of calculation will increase. This is also a problem that needs to be resolved and improved.

Acknowledgements

This work was supported by JSPS KAKENHI Grant Number 16H05877.

References

[1] Felzenszwalb P F, Girshick R B, McAllester D, et al. Object detection with discriminatively trained part-based models[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2010. 1, 4, 32(9):1627–1645.

[2] Sermanet P, Eigen D, Zhang X, et al. Overfeat: Integrated recognition, localization and detection using convolutional networks[J]. arXiv preprint arXiv,:2013,1312.6229.

[3] K. He, X. Zhang, S. Ren, and J. Sun. Spatial pyramid pooling in deep convolutional networks for visual recognition. IEEE transactions on pattern analysis and machine intelligence, 2015, 37(9): 1904-1916.

[4] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In CVPR, 2014: 580-587.

[5] R. Girshick. Fast R-CNN. In ICCV, 2015: 1440-1448.

[6] S. Ren, K. He, R. Girshick, and J. Sun. Faster R-CNN: Towards realtime object detection with region proposal networks. In NIPS, 2015: 91-99.

[7] J. Redmon, S. Divvala, R. Girshick, A. Farhadi. You only look once: Unified, real-time object detection. Proceedings of the IEEE conference on computer vision and pattern recognition.,2016: 779-788.

[8] Joseph Redmon and Ali Farhadi. 2016. YOLO9000: Better, Faster, Stronger. Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 7263-7271.

[9] Redmon, Joseph, and Ali Farhadi. Yolov3: An incremental improvement. arXiv preprint arXiv:2018,1804.02767.

[10] W. Liu, D. Anguelov, D. Erhan, S. Christian, S. Reed, C.-Y. Fu, and A. C. Berg. SSD: single shot multibox detector. European conference on computer vision. Springer, Cham, 2016: 21-37.

[11] Krizhevsky, A., Sutskever, I., Hinton, G.E. ImageNet classification with deep convolutional neural networks. Advances in neural information processing systems. 2012: 1097-1105.

[12] K. He, X. Zhang, S. Ren, and J. Sun. Deep Residual Learning for Image Recognition. Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 770-778.

[13] V. Nair and G. E. Hinton. Rectified linear units improve restricted boltzmann machines. Proceedings of the 27th international conference on machine learning (ICML-10),2010, 807-814.

[14] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, et al. Imagenet large scale visual recognition challenge, arXiv: 2014, 1409.0575.

[15] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman. The Pascal Visual Object Classes (VOC) Challenge. IJCV, 2010, pages 303–338.