Object Detection Algorithms Based on Convolutional Neural Networks

Zihan Deng [1,*], Ang Li [2]

[1] Suzhou High School Affiliated to NUAA, Soochow, China

[2] Dulwich International High School Suzhou, Soochow, China

* Corresponding Author Email: leo.li25@stu.dulwich.org

Abstract. Object detection is a fundamental and challenging task in computer vision, and it has attracted much attention from researchers worldwide. In recent years, deep learning technology has made remarkable progress and enabled new possibilities for object detection. Convolutional neural networks (CNNs), which are powerful tools for feature extraction and representation learning, have become the dominant approach for object detection, surpassing the traditional methods. This article reviews the development history of CNNs and their applications to object detection. It also introduces and compares two main branches of CNN-based object detection algorithms: region-based methods, which use a two-stage pipeline to first generate candidate regions and then classify them, and regression-based methods, which directly predict the bounding boxes and labels of objects in a single stage. Finally, it summarizes the current state-of-the-art and discusses the future directions of object detection research.

Keywords: object detection; convolutional neural network; deep learning; computer vision.

1. Introduction

Since Professor Yann LeCun put forward "Lenet5" in 1988, which laid the groundwork for Convolutional Neural Network, it has been applied to many fields including vehicle identification and human identification. Today's major work falls into two types: Candidate Target Detecting Method, also called Two-step Detecting Method, and Regression Based Target Detecting Method, also called One-Step Method. It is easy to tell from their names that their major distinction lies in the existence of regional suggestions.

Two-stage detecting method is separated into two stages. In this method, a specific area proposal is chosen by means of a recommendation box in the input picture. Then, by means of the convolution neural network, we can get the result of classification. This algorithm has high precision but is slow. Classic representatives include: SPP-Net [1], R-CNN [2], Faster R-CNN [3], Fast R-CNN [4], etc.

The steps of the one-stage algorithm are more concise, it does not need to generate region proposals, but directly completes feature extraction and target classification. This algorithm increases the calculation rate but is not good in accuracy. Classic representatives include YOLO [5] series, SSD [6] series and RetinaNet [7], etc.

This article introduces the development of CNN, IoU and a few classic representatives of two-stage detection algorithm and one-stage algorithm, including their basic operation principle，strengths and weaknesses. At the end, there is a summary and outlook.

2. Development of Convolutional Neural Networks and IoU

2.1. Development of Convolutional Neural Networks

In 1988, Professor Yann LeCun introduced the Lenet5 Convolutional Neural Network, which is considered to be one of the pioneering works in solving handwritten digit recognition, and it is also a milestone in the field of deep learning. It is also provided inspiration and guidance to subsequent related research. However, the network struggled with the task of processing larger, and more realistic

datasets. Until 2012, Alex Krizhevsky proposed AlexNet, which has a more complex and deep structure and has made great achievements in over-fitting.

With an effective improvement, the algorithm defeated other machine learning methods of the same period with an absolute advantage in the ImageNet competition and successfully won the championship. In 2014, ZFNet, an improved version of AlexNet, came out. It adjusted the convolution kernel and step size of the first layer, accordingly, improving the feature extraction ability. In the same year, in the ILSVRC competition, VGGNet and GoogLeNet won the championship of the positioning project and the classification project respectively. Among them, GoogLeNet reduces the amount of calculation by cleverly using the 1*1 convolutional layer, which is effective in avoiding the issue that the number of parameters will skyrocket due to an increase in the depth of the net.At the same time, the biggest highlight of VGGNet is the use of multiple 3*3 small convolution kernels in series, and many subsequent convolutional neural network structures were also borrowed from this design. Kaiming, He proposed ResNet in 2015. The net has solved the short circuit link efficiently and has prevented the network deterioration. This makes it possible to significantly increase the network depth to 152 levels, which is of epoch-making significance for the progress of neural networks. Subsequently, two upgraded versions of ResNeXt and ResNet-D were proposed in 2017 and 2018 respectively; ResNeXt separates the input channels, simultaneously performs convolution on multiple paths individually, and finally merges them by summing the outputs of all different paths. In this way, ResNeXt can learn more diverse and powerful features than ResNet, while keeping the same number of parameters and computation costs.

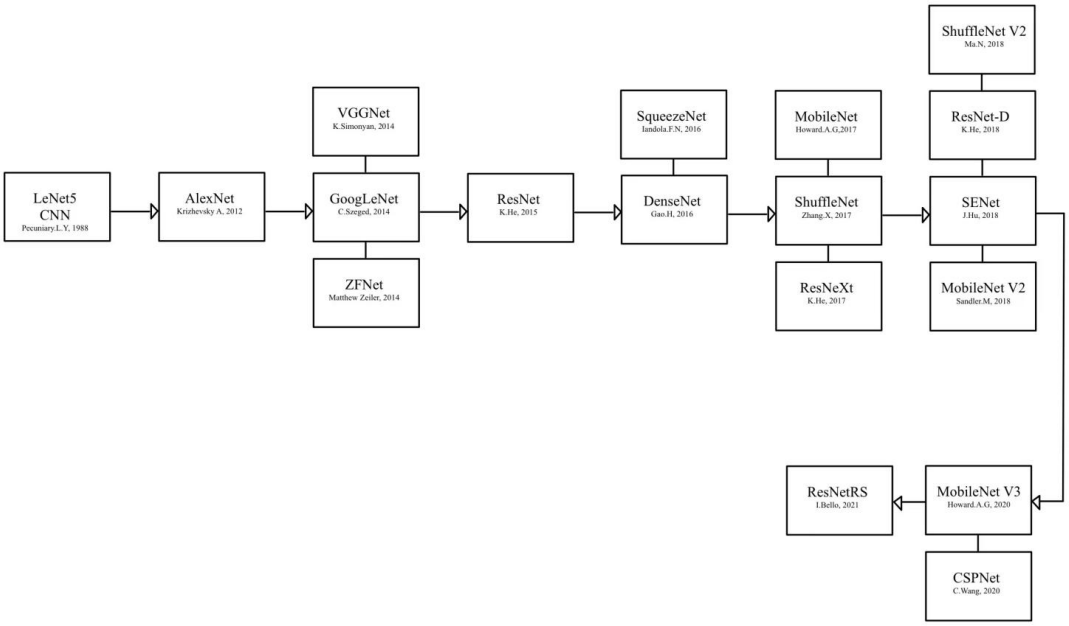

Fig. 1 Development of Convolutional Neural Networks (original)

2.2. IoU

IoU (Intersection over Union) is the intersection over union. It is an index to measure the overlap between the predicted borders and the real borders in target detection. By comparing the overlapping area between them, it can be used to evaluate the positioning accuracy of the object detection algorithm. The calculation method is shown in Figure 2, which is equal to the intersection area of two borders divided by their union area. The closer the value of the intersection over union ratio is to 1, the more the two borders match, and vice versa. It is generally agreed that when the intersection and union ratio is equal to or greater than 0.5, the detection will be considered correct.

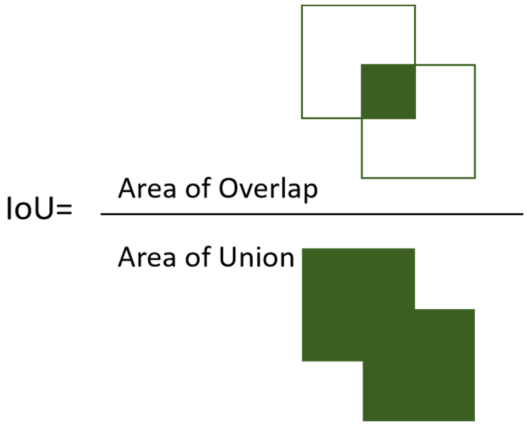

Fig. 2 IoU (original)

3. Convolutional Neural Network

3.1. Two-stage detection algorithm

A new method of detecting target based on area suggestion is also known as two stage detecting method. In this method, we select a specific region from a reference box in an input image, and then use a CNN to obtain the classification result. Typical examples include SPP-Net [1], R-CNN [2], Faster R-CNN [3], Fast R-CNN [4] and so on. Following is a brief introduction of the major models of the proposed algorithm, and a summary of their improved performance and results.

3.1.1 R-CNN

R-CNN [2] is a region-based CNN, which can be applied to detect targets. The idea of CNN is to generate approximately 2000 area recommendations from an image, then compress it into a pre-trained convolutional neural network model, and then apply SVM to sort the features. Lastly, the edge of the proposed area is modified by linear regression, and the overlay area is eliminated. R-CNN achieves 53.3% mAP on the PASCAL VOC 2012 data set, which is more than 30% higher than the traditional DPM method. R-CNN has the merit of taking advantage of the strong performance of deep learning, which increases the precision and velocity of object detection. However, R-CNN also has some shortcomings: (1) The detection efficiency is low.Due to the fact that every candidate frame is extracted individually, it leads to a lot of repetitive computation, so it takes much time to detect. (2) The detection accuracy is damaged. Due to the scaling operation of candidate box, the shape and proportion of original image and have changed, which affects the feature extraction recognition. (3) The training process is complicated. The generation of candidate boxes, feature extraction, classification, and regression are all performed separately, requiring multiple manual interventions, and the storage of intermediate data also consumes additional space.

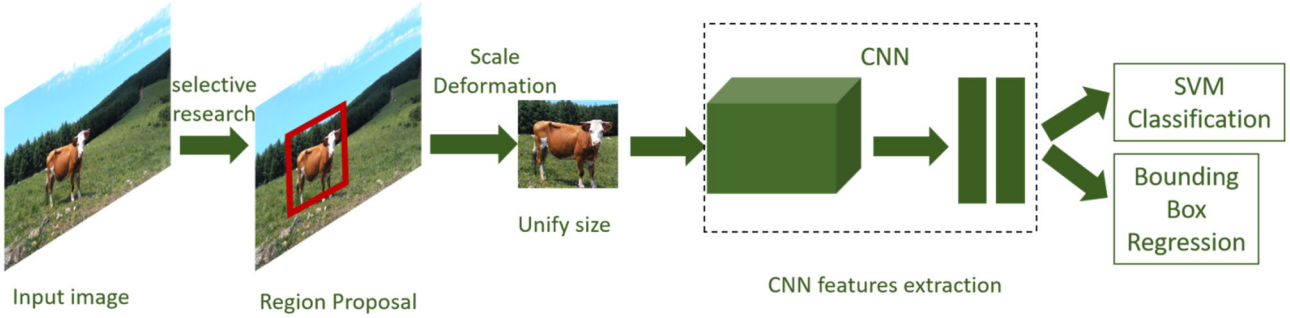

Fig. 3 R-CNN flow chart (original)

3.1.2 SPP-Net

SPP-Net [1], the primary benefit of this method is that the final convolution layer is supplemented with a Space Pyramid Pooling Layer (SPP), so that the input image size does not affect the extraction

and representation of features.In this workflow, it firstly produces about 2,000 candidate boxes by selecting search algorithm, then it carries out convolutions on the whole picture to get the characteristic map. Then, it finds the region related to each candidate frame in the feature map, and then integrates the SPP layer to attain the fixed-length feature vector, which could be applied to complete link. Finally, classifiers and regressors are used for object detection. The SPP layer effectively improves the size robustness. On the other side, because the algorithm only uses convolution once, the speed is greatly improved. On the Pascal VOC 2007 dataset, it achieves 59.2% accuracy and is 24 to 102 times quicker than R-CNN. However, SPP-Net also has some disadvantages: (1) Classification and regression training are still separated. This means that SPP-Net needs to train two different models separately, one for object classification and one for position regression, which increases the complexity and time of training. (2) Feature storage is still a problem.Since SPP-Net requires the acquisition of a fixed length characteristic vector for every candidate frame, it takes up a large number of spatial resources and decreases the rate of reading and writing. (3) SPP-Net still uses SVM as a classifier. As a result, SPP-Net cannot fully exploit the merits of convolution neural networks, and SVM has a complicated training procedure and a very long training period.

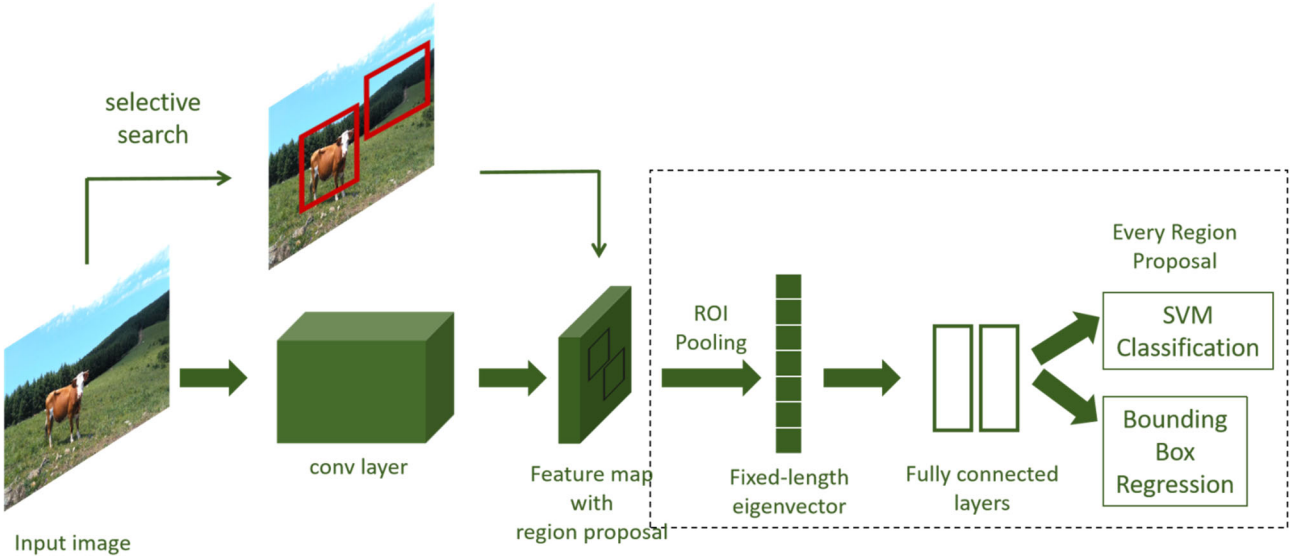

Figure. 4 SPP-Net flow charts (original)

3.1.3 Fast R-CNN

Fast R-CNN [4] is an enhanced SPP-Net model that utilizes ROI pooling [4] to transform the candidate boxes into uniform characteristic maps and characteristic vectors. The key step is to make a convolution of the input picture, then create a sequence of candidate boxes with Selective Search, and then map them into the convolution characteristic map. Then ROI pooling is carried out on every area proposal to get a new characteristic map with a constant size, and lastly, a full connection level for regression and categorization. Compared with the former one, Fast R-CNN is more innovative. Firstly, it uses multitask loss to train both classes and regressions at the same time. Secondly, it can make use of all convolutions in the procedure of training and testing, so as to make the network more efficient. With VGG-16 as baseline, Fast R-CNN has a speed of nine times that of R-CNN, three times of SPP-Net, 213 times of R-CNN, and ten times of SPP-Net. But Fast R-CNN is still based on Selective Search, so it takes too much time and restricts its real time capability. Furthermore, Fast R-CNN is not an end-to-end system because all of the suggestions are produced prior to convolution.

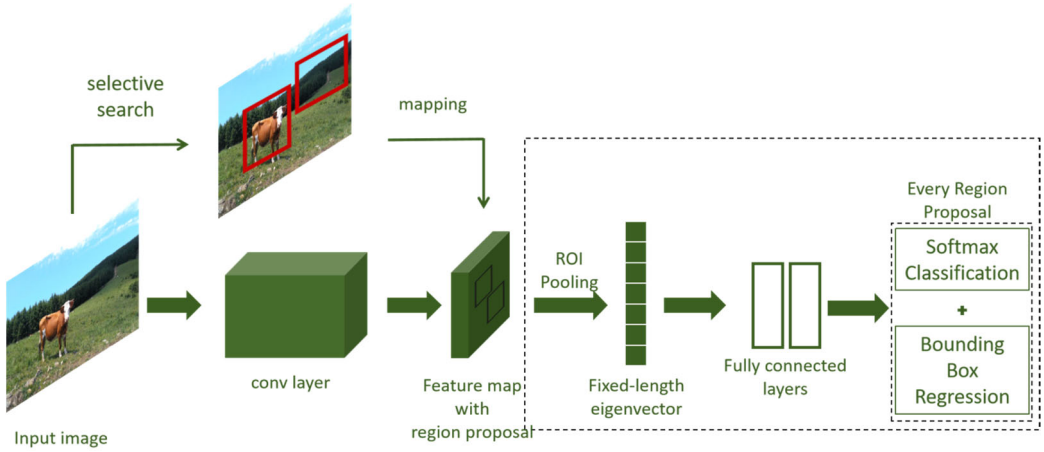

Figure. 5 Fast R-CNN flow charts (original)

3.1.4 Faster R-CNN

Faster R-CNN [3], which is based on Fast R-CNN, proposes a regional proposal network [3] (RPN) instead of the conventional area proposal generating approach. RPN is a region recommendation network based on convolutional feature maps. It performs foreground and background classification and position regression on the anchor boxes in each window by sliding windows on the feature map, thereby quickly generating high-quality region proposals. Faster R-CNN mainly made up of convoluting the input picture to get the characteristic map, and then the RPN net is used to output a sequence of area suggestions, which is then transmitted to the ROI pool level to get the new characteristic map of constant size. Finally, target classification and position regression tasks are completed through the fully connected layer. Compared with Faster R-CNN, Fast R-CNN has two significant advantages: one is to use the RPN network to generate region proposals, which avoids relying on external algorithms and improves detection speed and accuracy; the second is to share the convolutional feature map between the Fast R-CNN network and the RPN network, which reduces the amount of calculation and memory consumption. When using VGG-16 as the basic network, Faster R-CNN achieved 73.2% mAP on the VOC2007 data set, and the detection speed was increased from 3 f/s of Fast R-CNN to 7 f/s. However, Faster R-CNN also has some shortcomings: (1) The ROI pooling level is used in this paper, which results in the loss of characteristic graph dimension and insensibility of location, which influences the precision of object location. (2) Multiple fully connected layers are connected behind each region, which increases the network parameters and calculation amount, and reduces the detection speed. (3) Based on the characteristic map, a fixed anchor frame is adopted, and it can be found that the object cannot be detected well by multiple downsampling.

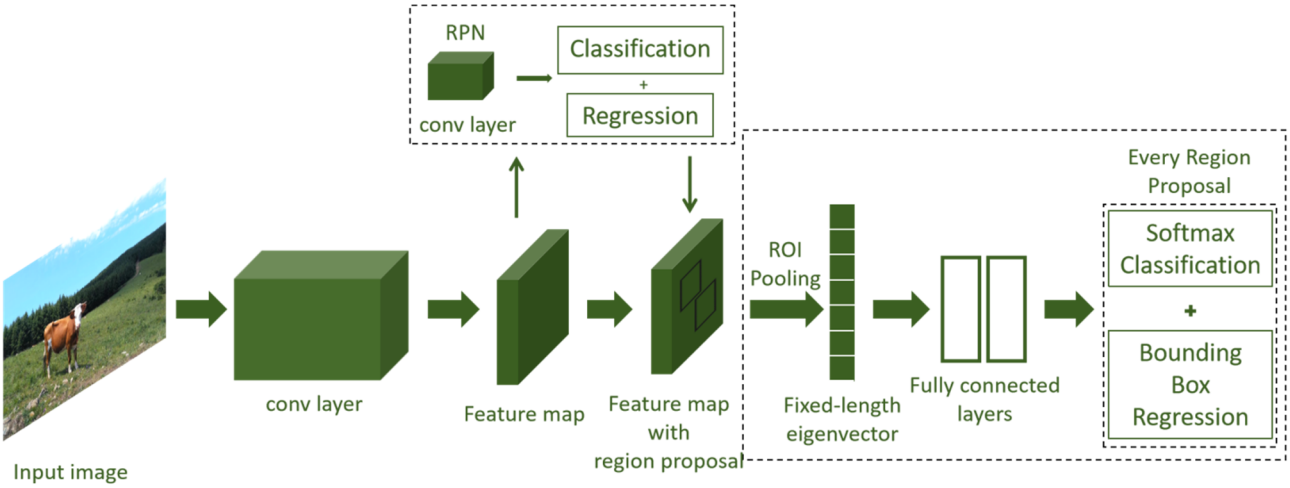

Figure. 6 Faster R-CNN flow charts (original)

3.1.5 R-FCN

R-FCN [8] is also a region-based fully convolutional network. Based on Faster R-CNN, ROI Pooling is enhanced, location sensitivity graph is added, location information is added, and the difference is resolved. The conflict between translation invariance and translation variability.The key point of this algorithm is that it can make a convolution of the input picture and get the characteristic map. Next, it will transmit a series of local recommendations through RPN, and then transmit them to location sensitive score pooling level to get a new feature map with rooted size. Finally, the target classification and position regression tasks are completed through the full convolutional layer, as shown in Figure 7. Compared with Faster R-CNN, R-FCN has three main advantages: one is to use ResNet as a feature extraction network, which improves the feature expression ability and classification effect; the second is to get rid of the full connection layer and use a fully convolutional network. Decrease the number of parameters and computation. the third is the use of position-sensitive score maps, which increases position information and improves the accuracy of target positioning. In the case where ResNet101 is a base network, R-FCN has reached 79.5% mAP in VOC2007 data set, and the detecting rate has been raised from Faster R-CNN's 7f/s to 10f/s. However, R-FCN also has some problems:(1) Because the position sensitivity graph is introduced, the complexity of the network and the training time are increased. (2) Since the position sensitivity graph is produced on the characteristic map, and the characteristic map is obtained by downsampling several times, which corresponds to the bigger region of the raw image, R-FCN is not effective for the small object.

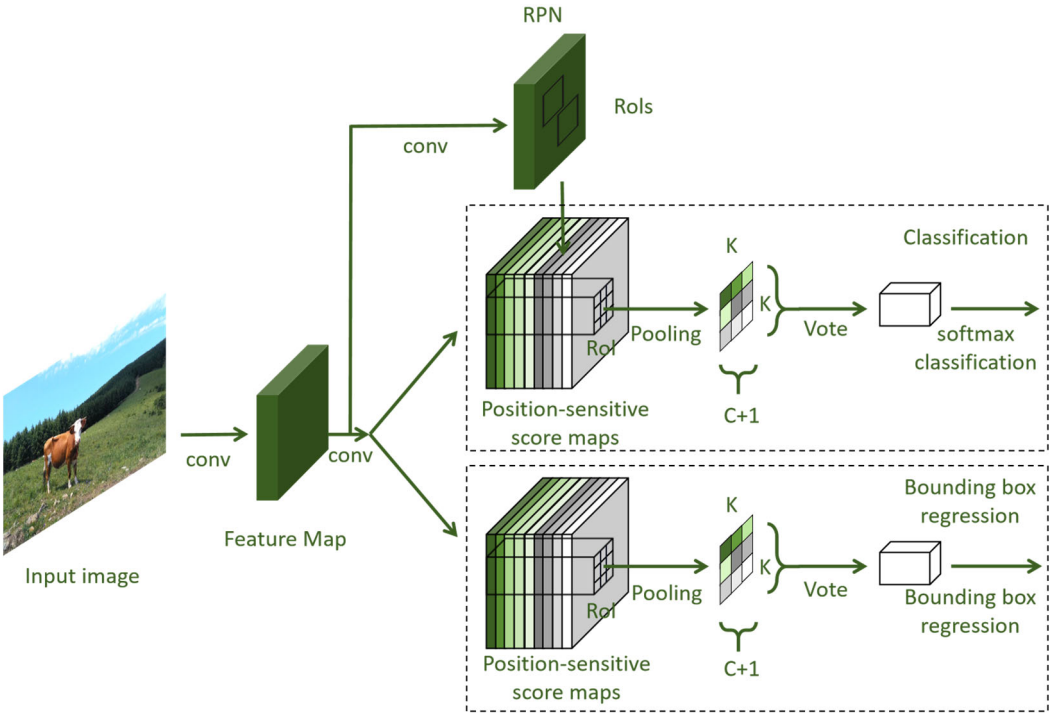

Figure. 7 R-FCN flow charts (original)

3.2. One-stage algorithm

Object detection based on regression is also called one-phase. Instead of producing area suggestions, it can extract the characteristic and classify the object directly. One stage target detector has the merit of quick detecting, but it has poor detecting precision. Typical examples include YOLO [5] family, SSD [6] family and RetinaNet [7]. Next, these models will be introduced in turn, and their performance improvements and achievements will be summarized.

3.2.1 YOLO v1

YOLO [5] v1 is a single-step method for detecting objects in an image. The kernel of this method is the convolution neural network, which can transform the input picture into 7*7 * 30-character map. In this feature map, each 7*7 grid represents a region of the picture, and the feature vector of each region contains the probability of 20 categories, the coordinates of 2 borders, and the confidence of 2 borders. The algorithm performs classification and regression according to these feature vectors to obtain the final detection results. The benefit of YOLO v1 is that it is fast, and because it is trained end-to-end, it has a low false false detection rate for the background. However, YOLO v1 also has some disadvantages, such as it cannot handle small objects and overlapping objects, because it is limited to only detect one object per region. In addition, it is less accurate than R-CNN.

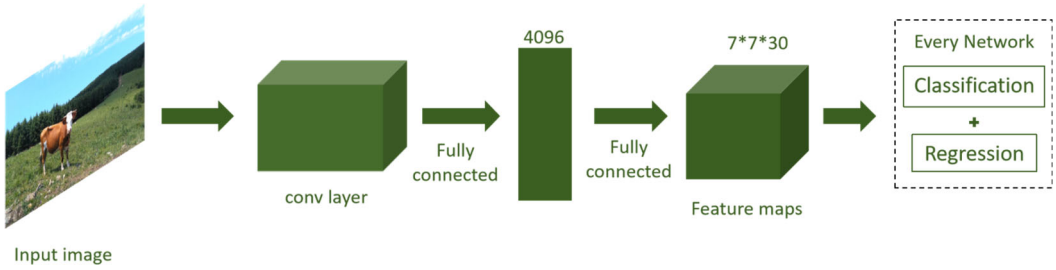

Figure. 8 YOLO v1 flow chart (original)

3.2.2 YOLO v2-v3

YOLO v2 [9] is an improved one-stage detection algorithm that introduces some new techniques based on YOLO v1. The most important of these is the use of anchor boxes, which are predefined shapes and positions of object boxes, which can be clustered according to the distribution of objects in the training set to better accommodate objects of different sizes and shapes. YOLO v2 improves the recall rate of detection by increasing the number of anchor boxes,and also it enhances the detection ability of small objects and overlapping objects, because each anchor box can independently predict the category and coordinates of the object. YOLO v2 also uses some other improvement methods, such as adding batch normalization and using Darknet-19 as a feature extraction network. YOLO v3 further optimizes the detection accuracy, mainly by fusing multi-scale feature maps and using the residual network structure.

3.2.3 SSD

SSD [6] (Single Shot Multi-Box Detector) is an efficient one-stage detection algorithm. The VGGNet architecture consists of four sub-convolution layers, which are followed by a further four convolution layers. This method has a number of default fields at every location of the characteristic map, which is analogous to the Faster R-CNN anchor box. Based on the feature map, a sliding window is constructed, and the classification and border position of the target are estimated at each position. SSD has many merits, such as multi-scale characteristic map, which can not only increase the precision of detection, but also resolve the problem of small object detection, which is much better than YOLO.

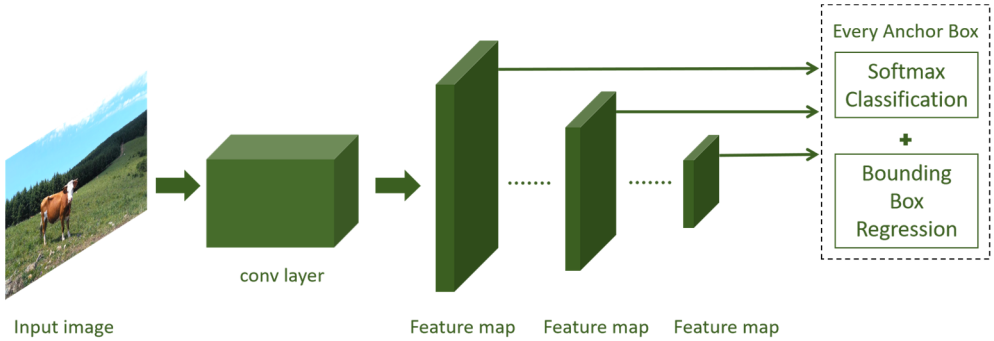

Figure. 9 SDD flow chart (original)

3.2.4 RetinaNet

The RetinaNet [7] is one of the most efficient methods to detect class disequilibrium. The most innovative feature of this paper is the introduction of a Focal Loss function that improves on the cross-entropy loss function. This method can decrease the effect of losing function on a single sample by adding an adjusting coefficient, but also enhance the effect of hard samples. RetinaNet's network structure is based on ResNet and FPN, which performs target detection on feature maps of different scales, while using the concepts of anchor boxes and default boxes. RetinaNet achieved good results on the COCO data set, with an AP value of 40.8%, comparable to two-stage detection algorithms. However, RetinaNet is slower than YOLO and SSD.

4. Conclusion

In this paper, we give an overview of the progress of CNN and focus on two kinds of CNN: Candidate Detecting Method, or Two-Step Detecting Method, and Regression Based Target Detecting Method.

Generally speaking, there is a screening procedure for detecting target based on area proposals. Therefore, with the development of technology, the precision is becoming greater and greater, and the process is getting simpler and simpler over the years.But in general, the detecting rate is very slow and cannot satisfy the requirements of the application. The regression-based object detection network is directly converted into a regression problem, so the detection speed is improving. But it has the problem of unbalanced negative and positive samples, so the accuracy will be affected. However, with the evolution and development of the framework in recent years, its accuracy has also been continuously improved, and it has advantages in practical application.

In recent years, due to recent advances in the field of computer vision, deep neural networks have been widely used, the object detection algorithm based upon CNN has surpassed the traditional algorithm, but there are still many difficulties to be overcome [10]. For example, how to train the object detection network under a limited data set, how to achieve sufficient accuracy requirements in the actual application field, etc., are all the focuses of future research in the domain of object detection.

Over the past few years, as the Internet of Things continues to evolve, we have continued to redefine and refine the definition of connectivity. On the one hand, we are evolving towards a more general IoT architecture. On the other hand, we are constantly expanding the range of connections from low power to high reliability. At the same time, we continue to strive to improve the quality and performance of our connections. We have made a major breakthrough in the definition of connectivity, connecting devices to the most important devices in the network, rather than to the most important sensors. This solves many problems: First, as the number and type of sensors increases, we need a simple and efficient way to extend network connectivity. Second, we need to ensure that every device is fully utilized and managed. Third, we need to implement only one device and one network communication within the same network. Fourth, in order to solve these problems, we need to research and develop new technologies and methods. Second, we are working to expand the scope of the connection. We have made a major breakthrough in the definition of connectivity. The final point is to address these issues: evolving and updating network connectivity technologies for a more general evolution of the IOT architecture and more efficient management and optimization of connections to meet user needs. That's all there is to know about the definition and evolution of connectivity on the Internet of Things.

Authors Contribution

The contributions of all the writers were equal, and their names were listed alphabetically.

References

[1] He K, Zhang X, Ren S, et al. Spatial pyramid pooling in deep convolutional networks for visual recognition[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2015, 37(9): 1-14.

[2] Girshick R, Donahue J, Darrell T, et al. Rich feature hierarchies for accurate object detection and semantic segmentation[C]. Conference on Computer Vision and Pattern Recognition, 2014: 580-587.

[3] Ren S, He K, Girshick R, et al. Faster R-CNN: towards real-time object detection with region proposal networks[C]. Conference on Neural Information Processing Systems, 2015: 91-99.

[4] Ross Girshick. Fast R-CNN[J]. IEEE International Conference on Computer Vision, 2015:1440-1448.

[5] Redmon J, Divvala S, Girshick R, et al. You only look once: unified, real- time object detection[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition,2016: 779-788.

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

[7] Lin T Y, Goyal P, Girshick R, et al. Focal loss for dense object detection[C]//Proceedings of the IEEE international conference on computer vision. 2017: 2980-2988.

[8] Dai J, Li Y, He K, et al. R-FCN: Object detection via region-based fully convolutional networks[J]. Advances in neural information processing systems, 2016, 29.

[9] Liu Hongjiang, Wang Mao, Liu Lihua, Wu Jibing, Huang Hongbin. Review of Small Object Detection Based on Deep Learning [J]. Computer Engineering and Science, 2021, 43 (08): 1429-1442.

[10] Yin Hongpeng, Chen Bo, Chai Yi, et al Overview of Visual Based Object Detection and Tracking [J] Journal of Automation, 2016, 42 (10): 1466-1489