Journal of Physics: Conference Series

PAPER • OPEN ACCESS

Summary of Target Detection Algorithms

To cite this article: Wei Li *et al* 2021 *J. Phys.: Conf. Ser.* **1757** 012003

Summary of Target Detection Algorithms

Li Wei, Feng Xiangsheng*, Zha Kaiwen, Li Shuning, Zhu Huasheng

*School of Information Engineering, Nanchang Institute of Technology, No.289, Tianxiang Avenue, NanChang 330099

*fengxiangsheng@nit.edu.cn

Abstract. In recent years, the soaring development of CNN has facilitated the maturity of the Computer Vision Algorithm. This paper will briefly introduce some representative Target Detection Algorithm, and systematically analyze the underlying problems, the modified methods, and the prospective direction of the algorithm in accordance with its merits and demerits. It is generally divided into a single-stage detection model and a double-stage detection model in terms of whether candidate areas need to be extracted during target detection for further tasks. Featured with scale in the double-stage detection model, the algorithm is divided into single-scale detection and multi-scale detection based on whether it can appropriately integrate with a network structure, which enhances the accuracy of the network model towards small targets. Meanwhile, it can also be divided into anchor-base and anchor-free in a single-stage detection model on the basis of the anchor bolt. A predictable development of the Target Detection Algorithm will show to us in the future.

Keywords: Computer vision, Target Detection, Convolutional Neural Network.

1. Introduction

Computer vision technology has been integrated into all aspects of life in the continuous development of today’s society. Target detection is a very basic but very important task in computer vision technology. Target detection is used in social security management, traffic vehicle monitoring, environmental pollution detection, and forest disasters. There are very outstanding application results in the fields of early warning and national defense security. The task of target detection mainly includes the recognition and location of single or multiple targets of interest in digital images. People process the training images containing the target to extract stable and unique features or specific abstract semantic information features, and then match these distinguishable features or use classification algorithms to give confidence in each category Degree to classify.

Target detection algorithms have been studied for many years. In the 1990s, many effective traditional target detection algorithms appeared. They mainly used traditional feature extraction algorithms to extract features and then combined with template matching algorithms or classifiers for target recognition. However, traditional algorithms have encountered a bottleneck in their development due to the lack of strong semantic information and complex calculations. In 2014, Ross Girshick proposed a convolutional network-based target detection model RCNN[1] with high detection accuracy and strong specific robustness and generalization ability, making people pay more attention to the use of convolutional neural networks to extract High-level semantic information of images and many excellent detection models of convolutional neural networks have been proposed.

2. Traditional algorithm Models of Target Detection

The traditional target detection training model can be roughly divided into two steps ：

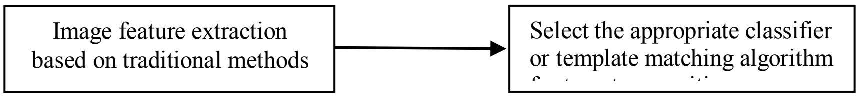
Image feature extraction based on traditional methods
Select the appropriate classifier or template matching algorithm

The different modes based on image feature acquisition can be divided into two categories:

Feature algorithm model based on feature regions. (Haar, LBP, HOG[2] features, etc.), the feature vector that is easy to distinguish is mainly obtained by selecting the appropriate detection frame or feature template.

Feature algorithm model based on feature points. (SIFT[3], SURF[4], ORB[5] features, etc.), mainly to locate some stable and unique feature points in complex scenes such as extreme points, dark points in bright areas, and bright spots in dark areas. Then use the feature point descriptor with higher distinguishability to distinguish different feature points.

There are many classification matching methods used in traditional methods, which can be divided into Similarity model (K-Nearest Neighbor[6], Rocchio) probability model (Bayes[7]) linear model (SVM[8]) nonlinear model (decision tree[9]) ensemble classifier (Adaboost[10]).

3. Convolutional Neural Network

In 2014, the proposal of the RCNN convolutional network opened a new stage in the development of target detection. Its accuracy and stability greatly exceed the traditional target detection algorithm, so it is quickly accepted by people. The detection model of the convolutional neural network is mainly divided into single-stage and two-stage detection models. The difference is that the two-stage detection model needs to train a region proposal network (ROI), but it increases the computational complexity and the model is difficult to achieve real-time detection. The single-stage model discards this link and converts the target detection problem into a regression problem. Although the accuracy is sacrificed, the calculation speed of the model is greatly improved, and the model can be detected in real-time.

3.1. The process of Convolutional neural network detection

Convolutional neural network target detection also has certain similarities with traditional detection, which can be regarded as feature extraction and use of features to identify targets. Both backbone feature extraction network and "detection head"

Convolutional neural networks use convolutional networks to extract high-level semantic features of images, which are generally the backbone of the network, and then process the feature maps, such as connecting a fully connected network and softmax or svm to form a classification head to complete the classification task, or use a small volume The product core is processed into feature dimensions and a position loss function is used to form target positioning.

The network judges the error through the loss function and updates the network weight parameters by the reverse gradient propagation of the network, thereby continuously reducing the value of the loss function to improve the detection accuracy. The detection network calculates multiple times through a large amount of training data and can learn a set of optimal weight values from the set of data to predict the detection target.

3.2. Backbone network of Convolutional Neural Network model

Currently, the more famous backbone networks are: (1998) LeNet-5 (2012) AlexNet (2013) ZFNet (2014) GoogLeNet (2014) VGGNet (2014) VGGNet (2015) ResNet (2016) ResNet v2 (2017)

ReNetXt (2018) DenseNet (2019) VoVNet and (2020) VoVNet-v2 basically multiply and accumulate the image of the image by a variety of convolution kernels according to a certain spatial position to obtain the next-level feature image.

4. Development of the backbone network of the convolutional neural network

4.1. LeNet-5

The earliest classic convolution feature backbone network was LeNet-5[11] proposed by Lecun et al. in 1995. Although the network model at that time was relatively simple, it already included the most basic convolution pool in convolutional neural networks. The transformation and fully connected layers play a guiding role in the development of convolutional neural networks, which are mainly used for handwriting recognition.

4.2. AlexNet

Convolutional neural networks became popular in 2012. Among them, Alex Krizhevsky proposed that the structure of the AlexNet network[12] is similar to LeNet on the whole, which is convolution first and then fully connected. But the network is more complicated, using a five-layer convolution, a three-layer fully connected network, and the final output layer is a softmax of 1000 channels. AlexNet uses two GPUs for calculations, which greatly improves the computing efficiency. In the ILSVRC-2012 competition, it obtained a top-5 error rate of 15.3%. In order to obtain a larger receptive field, the first layer of the network uses an 11*11 convolution kernel and adds LRN local response normalization to each convolution layer to improve accuracy, but in 2015 Very Deep Convolutional Networks for Large-Scale Image Recognition. It is basically useless to mention LRN.

4.3. ZFNet

Because people want to understand the working principle of convolutional neural networks, ZFNet[13] was proposed in 2013, which provides a visual network to understand the various layers of convolutional networks. The main improvement to AlexNet is to use a smaller convolution to reduce the time complexity while using deconvnet and visual feature maps to visualize it, and at the same time won the ILSVRC championship. By visualizing the neural network, it can be seen that the low-level network has extracted the edge texture features of the image, and the high-level network has extracted the abstract features of the image. This feature has translation and scale invariance but does not have rotation invariance.

4.4. GoogLeNet

In order to further improve the performance of the neural network, the most direct way is to increase the depth and width of the network, but it will cause too many parameters to increase the amount of calculation, and the limited training set will cause problems such as gradient dispersion or overfitting. For example, the 22-layer AlexNet has about 60 million parameters. GoogLeNet[14] proposed in 2014 has only 5 million parameters in the same situation. It mainly uses the convolution solution. GoogLeNe v1 decomposes a 5*5 convolution operation into two 3*3 convolution operations. When they obtain the same receptive field, the parameters are reduced by 2.78 times. GoogLeNe v2[15] divides a 3 The *3 convolution operation is decomposed into 1*3 and 3*1 convolution operations. GoogLeNe V3[16] decomposes the 7*7 convolution kernel into 7*1 and 1*7 convolution kernels, which deepens the depth of the network and reduces the parameters. GoogLeNe v4[17] is based on GoogLeNe v3 The addition of the residual network greatly increases the depth.

4.5. VGGNet

The VGGNet [18] proposed by Karen Simonyan et al. in 2014, which is equivalent to the network deepening version of AlexNet, consists of two parts: a convolutional layer and a fully connected layer. All activation layers use relu, and the pooling layer uses maximum pooling. The simple structure and strong feature extraction ability made it second place in the classification project of ILSVRC2014 and first place in the positioning project. VGGNet, used in the test, used a 1*1 convolution kernel to improve the fully connected layer, becoming a fully connected layer with convolution. This overcomes the disadvantage that the traditional fully connected layer requires a fixed input dimension. Therefore, multi-scale training is adopted, and the training image scale is randomly selected from the range of [256, 512] side length. This scale jitter method can enhance the training set.

4.6. ResNet series

As the network depth increases, the acquired features become richer but the optimization effect becomes poor. The accuracy of detection decreases (gradient explosion and disappearance). For a shallower network, the input data of each layer can be normalized to make the network converge. But the deep network still has optimization problems. So in 2015, He Kaiming proposed ResNet [19] to break this bottleneck, mainly using the jump connection structure. On the basis of ResNet v2 [20] proposed in 2016, by changing the order of the normalization layer, the pooling layer, and the convolutional layer, a set of jump structures with the best performance (Figure 22) was found, and it was proposed in 2017 ReNetXt [21] draws on the idea of GoogLeNet to add 1*1 convolution on both sides of the convolution layer, reducing the number of control cores and reducing the parameters by about two-thirds.

4.7. DenseNet

The previous convolutional network is either as wide as GoogLeNet or as deep as ResNet. The author of DenseNet [22] published in 2018 found two characteristics of two neural networks through experiments: 1. After removing the middle layer, the next layer is directly connected to the upper layer, that is, the neural network is not a progressive hierarchical structure, and it is not necessary to connect adjacent layers. You can rely on the previous layer for feature learning. 2. Many layers of ResNet are randomly removed during training, and it will not affect the convergence and prediction results of the network. It proves that ResNet has obvious redundancy, and each layer in the network only extracts a few features (the so-called residual). Compared with ResNet, DenseNet has obvious advantages, improved performance, and reduced parameters.

4.8. SENet

SENet [23], published in 2019, is aimed at the detection task and proposes a channel weight combined with the idea of the attention to suppressing features that are not useful for the current task. The SE module is mainly used for weight distribution for the convolutional layer. This sub-module form makes it compatible with other networks. The article is mainly used in the ResNet network. The SE module is embedded in ResNeXt, BN-Inception, Inception-ResNet-v2, and has achieved a lot of gains. It can be seen from this that the gain effect of SE is not only limited to some special network structures, it has strong generalization.

4.9. EfficientNet

Considering that the previous network mainly improves the accuracy of the network through a single scaling of the width (WideResNet and MobileNets), depth, and resolution of the network model. The EfficientNet network model [24] quantifies the relationship between these three dimensions and uses a constant ratio to simply increase to balance the three dimensions of the network at the same time.

4.10. VoVNet series

The VoVNet network [25] proposed in 2019 has completely surpassed ResNet and can be used as a backbone network for real-time target detection. Considering the factors of energy consumption and model inference speed, optimizing memory access costs (the highest efficiency when the number of input and output channels is the same) and GPU computing efficiency (GPU processing large tensors is strong, CPU processing small tensors is strong) is more critical. The most important thing is to put forward the OSA module, which optimizes the problem of densely connected DenseNet modules. At the same time, it is improved in the 2020 CenterMask article, and the residual block and eSE module are added (in the original SE module improvement, one FC is used to replace the original two FC to reduce information loss) to greatly increase its performance and form VoVNet – v2 structure. Compared with ResNet, the VoVNet network has a stronger ability to extract small targets, and speed and accuracy are better.

5. Network Models of Target Detection

The network model is generally to detect sub-regions in the image. Because the traversal detection based on the sliding frame requires a large amount of calculation, the candidate frame is used to initially locate the area of interest, and then each candidate area is detected to greatly reduce the calculation of the network Complexity, this kind of algorithm that extracts candidate regions and then detects and locates the target is called a two-stage detection algorithm. The accuracy of the two-stage detection algorithm is high, but the amount of calculation is still large, so it is difficult to achieve real-time detection.

In view of the practicability of the two-stage detection, the single-stage detection algorithm does not need to extract candidate regions but performs regression prediction on each feature map, which greatly reduces the time complexity of the network algorithm. The accuracy of the single-stage detection algorithm in recent years has been close to that of the two-stage detection algorithm while maintaining a high detection speed, making its development attract more people's attention.

5.1. Network models of Two-stage target detection

The development of the two-stage detection model started from the initial RCNN, and there are many improved models around the RCNN model, such as SPP-net, Fast-rcnn, Faster-rcnn, R-FCN, etc. These models all improve the RCNN network on a single scale feature to greatly improve accuracy and speed.

There are also some improvements to the RCNN network that combine the idea of multi-scale feature fusion. Such as ION, FPN, MASK-RCNN, etc., this multi-scale feature fusion improves the ability of the network model to detect small targets.

5.1.1. Based on the single-scale feature model

The R-CNN [26] process proposed by Ross Girshick et al. in 2014 is relatively simple. First, select more than 2000 candidate frames randomly by using (selective search) on the input image, and then zoom to 227*227, and then use AlexNet CNN Extract the features to obtain a 2000*4096 matrix, and then use the svm algorithm to classify, that is, multiply the feature matrix by the matrix 4096*20 (representing 20 classes). The class with a score greater than a certain threshold is judged as this class. The accuracy of the R-CNN model on VOC 2010 reached 53.7mAP.

The SPP-net [27] proposed by Kaiming He et al. in 2015 solved two pure problems of the RCNN network at that time: 1. Candidate frames are randomly selected from the original image, and each candidate frame will be subjected to a featured network. The extraction of this kind of repeated convolution calculation greatly increases the computational burden. 2. A fixed-size input image is required, so the original image needs to be cropped or scaled. These operations may cause the target information to be lost and affect the accuracy. For the first problem, a shared feature convolution layer is used, and the final convolution layer The selection of candidate regions is performed to reduce the amount of calculation. For the second problem, the root cause is that the fully connected layer needs to input a fixed-dimensional feature vector. In order to solve this problem, SPP-net added a pyramid pooling layer to the last convolutional layer (feature map) of the feature network for ordered output. The accuracy of the SPP-net(ZF5) model on VOC 2007 reached 59.2mAP.

The Fast-rcnn[28] proposed by Ross Girshick et al. in 2015 draws on the method of sharing convolutional image feature layers in spp-net. All prediction boxes share a convolutional network by mapping to the feature map layer, and at the same time in the last layer of the feature map proposes a simplified version of SPP-net's feature pooling layer to output fixed-dimensional feature vectors, and then connect to the fully connected layer for subsequent operations. Combining the idea of spp-net to optimize the rcnn network, greatly reducing the time complexity of the network (but the use of selective search for candidate boxes is still very time-consuming). It laid the foundation for the development of Faster-rcnn in the future. The accuracy of the SPP-net model on VOC 2007 reached 66.9mAP.

Shaoqing Ren et al. 2016 proposed Faster-rcnn [29] to solve two problems in Fast-rcnn: 1. The suggestion box uses a selective search algorithm, which greatly increases the number of network calculations. 2. The objective loss function of the positioning frame is unstable at the optimal solution point using L1 distance. The Faster-rcnn training network is an end-to-end network, which realizes the sharing of most of the calculations, and has high detection accuracy and anti-interference. Although the real-time performance is not high, its unique regional suggestion network RPN is the goal of the whole stage Detection opens up new ideas. The accuracy of the Faster-rcnn(VGG-16) model on VOC 2007 reached 69.9mAP.

The R-FCN [30] proposed by Jifeng Dai et al. in 2016 solves the problem of the Faster-rcnn network model because each suggestion box in the pooled ROI layer needs to be separately connected to the fully connected layer for classification and positioning. Each feature point will generate 9 suggestion boxes and consume a lot of computing resources. R-FCN shares calculations with all the suggestion boxes after passing through the ROI. The backbone feature network uses a deeper residual network, but due to the increase in depth, the feature map is further reduced. When the object on the original image is shifted, the perception ability on the feature map becomes weak after the convolutional network, and the translation variability of the network changes. difference. (Classification tasks need better translation invariance, positioning tasks need better translation variability) So R-FCN adds a position-sensitive score map to solve this problem. The accuracy of the R-FCN(ResNet-101) model on VOC 2007 and VOC 2012 reached 79.5mAP.

5.1.2. Based on multi-scale feature fusion model

The ION[31] model proposed by Sean Bell et al. in 2015 has two problems with the target detection model at the time: 1. At that time, fast-rcnn or spp-net detected the proposed suggestion box, which lacked the suggestion box. Contextual information outside (around the target). 2. Both only use the feature map of the last layer, only use the high-level semantic features, and lack the use of the low-level detailed features. For problem 1, the ION network model uses the idea of recurrent neural network to extract context information by connecting two IRNN units, and for problem 2, multi-scale feature fusion is used for detection. The ION network model uses contextual information to obtain relatively broad image feature information combined with image detail information obtained by multi-channel fusion to achieve better prediction results. The accuracy of detecting small objects is improved, and the accuracy of detecting occluded objects is mentioned. The accuracy of the ION model on coco reached 33.1AP.

The FPN proposed by Tsung-Yi Lin et al. in 2017 [32] utilizes the high-level semantic information fusion of the underlying network structure to increase the resolution of the feature map, and predicting on a larger feature map is conducive to obtaining more small targets. The feature information, makes the small target prediction effect significantly improved. The accuracy of the FPN model on coco reached 59.1AP.

The Mask-rcnn [33] proposed by Kaiming He et al. in 2018 is similar in structure to Faster-crnn. It is a flexible multi-task detection framework that can complete: target detection, target instance segmentation, and target keypoint detection. Simply put, a "detection head" (segmentation task layer) is added to the Faster-crnn framework structure. Due to the introduction of the mask layer, the network can handle segmentation tasks and key point tasks. ROIAlign avoids the two quantization of Faster-rcnn and improves detection accuracy. The accuracy of the Mask-rcnn model on coco reached 36.4AP.

5.2. Network models of Single-stage target detection

The earliest single-stage detection model is YOLO v1. One type of improvement is to use anchor-base on the feature map obtained by the feature extraction network to detect the target point by point according to the preset anchor frame, such as SSD, YOLO V2, RetinaNet, YOLO V3, YOLO V4, EfficientDet, etc.

At the same time, another type of improvement is to use the anchor-free idea to directly point the two corner points and the center point of the target through the network, and use these key points to achieve the return positioning task of the target. Such as CornerNet, CenterNet, CornerNet-Lite, FCOS, CenterMask, etc. The anchor-free model overcomes the following five shortcomings of the anchor-base model: 1. The detection performance is very sensitive to the size, aspect ratio, and the number of the anchor frame, so the hyperparameters related to the anchor frame need to be carefully adjusted. 2. The size and aspect ratio of the anchor frame is fixed. Therefore, it is difficult for the detector to process candidate objects with large deformation, especially for small targets. 3. The pre-defined anchor boxes also limit the generalization ability of the detector, because they need to be designed for different object sizes or aspect ratios. 4. In order to improve the recall rate, dense anchor frames need to be placed on the image. Most of these anchor boxes belong to negative samples, which causes an imbalance between positive and negative samples. 5. A large number of anchor boxes increase the amount of calculation and memory usage when calculating the intersection ratio.

5.2.1. Based on the anchor-base detection model

The SSD[34] network model proposed by Wei Liu et al. in 2016 proposes two improvements to the problems of Yolo v1 target detection frame inaccurate positioning and poor detection of small targets. 1. SSD uses multi-scale fusion to enhance detection accuracy (that is when it contains Predict small target objects on large-scale feature maps with rich spatial detail information, and predict larger target objects on high-level feature maps containing highly abstract semantic information). 2. SSD uses Anchors (Candidate frames with different aspect ratios) similar to those in Faster rcnn, to a certain extent, overcoming the inaccurate positioning of the YOLO v1 algorithm and the difficulty of locating small targets. The accuracy of the SSD（512） model on coco reached 26.8AP.

YOLO V2 [35] proposed by Joseph RedmonR et al. in 2016 improves the difficulty of small target detection and inaccurate target frame positioning in the YOLO V1 model. YOLO V2 first used the Darknet-19 feature extraction network to replace GoogleNet of YOLO V1. The use of higher resolution feature maps for prediction and the use of multi-label models to combine data sets make the flat network structure simplified into a structure tree. At the same time, the joint training classification and detection data mechanism are used to expand the training data set and improve the detection accuracy, and its accuracy exceeds the two-stage faster-rcnn. The accuracy of the YOLO V2 model on coco reached 21.6AP.

The RetinaNet[36] model proposed by Tsung-Yi Lin et al. in 2018 usually has a much larger number of negative samples than positive samples during training. This type of imbalance (often causes the final calculated training loss to be an absolute majority but contains The negative samples with a small amount of information are dominated by the negative samples, but the key information provided by the few positive samples cannot play a normal role in the generally used training loss so that it is impossible to draw a loss that can provide correct guidance for model training) will be trained in loss A large accuracy error occurs when So RetinaNet divides the samples into hard examples: difficult to distinguish samples (0.4<IOU<0.5) easy examples: easy to distinguish samples. And use Focal Loss (eliminate category imbalance + mining difficult samples) to improve accuracy. The accuracy of the RetinaNet（ResNeXt-101-FPN） model on coco reached 40.8AP.

YOLO v3 [37] proposed by Joseph Redmon et al. in 2018 is a further improvement of YOLO V3 on the basis of YOLO V2, its detection is more accurate, and the speed is still very fast. The main change is to use Darknet-53 to replace the Darknet-19 backbone feature extraction network (YOLO v3 uses the previous 52 layers of darknet-53 (no fully connected layer)). With the addition of multi-scale fusion detection, three outputs of y1, y2, and y3 are obtained in different layers (there are only three a priori boxes on the feature map point of each prediction scale). And modified the loss. The accuracy of the YOLO v3（Darknet-53） model on coco reached 33.0AP.

YOLO v4 [38] proposed by Alexey Bochkovskiy and others in 2020 adopts the best optimization strategy in the CNN field in recent years on the framework of the traditional YOLO series, from data processing, backbone network, network training, activation function, loss function, etc. Various aspects have been optimized to varying degrees. It balances the detection speed and accuracy, which is greatly improved compared to YOLO V3. The accuracy of the YOLO v4（CSPDarknet-53） model on coco reached 43.5AP.

The EfficientDet[39] proposed by Mingxing Tan et al. in 2020 achieves high accuracy while maintaining a low amount of floating-point operations. The EfficientDet model size increases from D0 to D7 according to different accuracy requirements. . EfficientDet mainly uses the FPN network and makes a BiFPN Layer structure of multi-layer feature fusion of the feature maps of different layers, and follows the idea of EfficientNet feature extraction network, using a simple parameter ϕ to realize its backbone network, feature fusion network BiFPN, Box/Class predicts the composite scaling of the network scale, making the network more efficient. The accuracy of the EfficientDet-D0（512） model on coco reached 34.6AP, and the accuracy of the EfficientDet-D7x（1536）model on coco reached 55.1AP.

5.2.2. Based on the anchor-free detection model

Joseph Redmon et al. first proposed a more classic single-stage YOLO V1 [40] model in 2016. In order to improve the detection speed, the single-stage detection removes the two-stage region proposal network and directly determines the target category and target frame in the output layer. The positioning of the target, taking the entire picture as input, turns the target detection task into a regression task. The earlier Yolo v1 algorithm has a concise structure and can well reflect the characteristics of the single-stage detection network model. The accuracy of the YOLO V1model on coco reached 57.9mAP.

The CornerNet [41] proposed by Hei Law et al. in 2018 uses a single convolutional network to change the target boundary to predict a pair of key points (that is, the upper left corner and the lower right corner of the target box). This design can eliminate the commonly used single-stage detection Forecast the demand for anchors. At the same time, the pooling layer is improved, and corner pooling can be used to locate the corners of the bounding box (Figure 46). A 42.1% AP is achieved on the MS COCO data set, which is better than all single-stage detectors at the time, and comparable to the detection performance of two-stage detectors. The accuracy of the CornerNet511(single scale, Hourglass-104)model on coco reached 40.6AP, and the accuracy of the CornerNet511(multi scale, Hourglass-104)model on coco reached 42.2AP.

The CenterNet [42] proposed by Kaiwen Duan et al. in 2019 is also a single-stage anchorless frame network model. For CornerNet, the target is determined only by detecting the upper left and lower right corners of the target. It does not make full use of the internal feature information of the object, so many For the phenomenon of false detection frames, an improved cascade corner pooling with richer semantic information and a center pooling for detecting the center point characteristics of objects are proposed. Using such key point triples to detect objects greatly improves the detection accuracy and becomes the best single-stage detection model at the time, with a speed of about 270ms (52-layer feature network) and 340ms (104-layer feature network). The accuracy of the CenterNet 511(single scale,Hourglass-104)model on coco reached 44.9AP, and the accuracy of the CenterNet 511(multi scale, Hourglass-104)model on coco reached 47.0AP.

The CornerNet-Lite [43] proposed by Hei Law et al. in 2019 optimizes its backbone network on the basis of CornerNet to form CornerNet-Squeeze. The CornerNet-Saccade, which uses the attention mechanism for cropping, removes the redundant image part of the network detection target (Similar to two-stage detection, first cut out the approximate area of the target for detection). This method has made a good breakthrough in speed and accuracy, reaching the highest accuracy (47.0%) of the single-stage detector at that time. The accuracy of the CornerNet-Saccade model on coco reached 43.2AP.

The FCOS[44] network model proposed by Zhi Tian et al. in 2019 is roughly composed of the FPN feature pyramid and three branch detection heads. FCOS discards the traditional anchor box and directly performs regression operations on each point on the feature map. And the use of FPN's multi-scale hierarchical detection greatly reduces the fuzzy samples generated in multiple BBs (detection frames) in one location. Center-ness weighting combined with NMS (non-maximum suppression) is a good way to suppress the distance from the target center The low-quality BB. Compared with some of the most mainstream first-order and second-order detectors, FCOS is superior to the classic algorithms of Faster R-CNN, YOLO, and SSD in terms of detection efficiency. FCOS lacks speed in order to improve accuracy, but it is better than RetinaNet in terms of accuracy and speed. The accuracy of the FCOS(ResNeXt-64x4d-101-FPN) model on coco reached 44.7AP.

The CenterMask [45] proposed by Youngwan Lee et al. in 2020 is based on FCOS and adds the SAG-Mask instance segmentation module integrated into the attention mechanism and replaces its feature extraction backbone network (VoVNet-V2). Using the ResNet101-FPN backbone network can reach 38.3% mask AP surpasses all previous networks, but the speed is only 13.9FPS. The lightweight CenterMask-Lite can reach 33.4% mask AP and 38% box AP. The speed can reach 35FPS, so it can meet real-time requirements. The accuracy of the CenterMask (V-39-FPN) model on coco reached 36.3APmask.

6. The future development direction and summary of the target detection algorithm

In order to pursue faster and more accurate target detection algorithm models, the algorithm model will incorporate more other advanced model algorithms, and single-stage and two-stage methods will gradually merge. For example, the target position estimation proposed by the single-stage CornerNet-Lite model is pseudo The two-stage model adopts the idea of two-stage target detection.

With the diversification of detection task requirements, the target detection model is no longer a single task model, which adds instance segmentation (similar to multi-target detection, but uses edge contours instead of bounding boxes (target boxes)) and some are also added Panoramic segmentation (it is a combination of semantic segmentation and instance segmentation: semantic segmentation refers to assigning a category to each pixel on the image (can be distinguished by color) but does not distinguish between individuals). After panoramic segmentation, we can know which individual in which category each pixel on the image belongs to, which is a more refined classification task. At the same time, there is also key point detection for detecting the human body posture (that is, the joints of the human body are replaced by points and connected by adjacent line segments, which abstractly represent the human body posture actions).

References

[1] Girshick R, Donahue J, Darrell T and Malik J 2014 Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation *CVPR. IEEE*

[2] Yuan W, Huanhuan L and Kefeng W 2015 Fusion with layered features of LBP and HOG for face recognition *J.Jisuanji Fuzhu Sheji Yu Tuxingxue Xuebao/Journal of Computer-Aided Design and Computer Graphics* pp640-650

[3] Guomei W and Xiaowei C 2007 Research on SIFT Feature Matching Algorithm *J.Journal of Yancheng Institute of Technology* pp1-5

[4] Hongrong L， Xiaoming L. Research on feature detection methods based on SIFT PCA-SIFT and SURF *J. Journal of Taiyuan Normal University* pp74-76

[5] Xiaohong L, Chengming X and Yizhen J 2013 Fast target detection algorithm based on ORB feature *J.Journal of Electronic Measurement and Instrument* pp455-460

[6] Yu Z 2008 Improvement and realization of K-nearest neighbor algorithm *J.Computer development and application* pp18-20

[7] Zhongqiang Y 2013 *Research on Naive Bayes Algorithm Based on Attribute Weighting and*

Reduction （Nanning: Guangxi University）

[8] Xiaoyun W 2005 SVM algorithm analysis and research *Journal of Western Chongqing University*

[9] Wei L and Cong W 2007 Optimization and Comparison of Decision Tree Algorithms *J.Computer Engineerin*g pp189-190

[10] Ying C, Qiguang M and Jiachen L 2013 Research progress and prospects of AdaBoost algorithm *J.Acta Automatica Sinica* pp745-758

[11] LeCun Y. Le Net-5 convolutional neural networks

[12] Krizhevsky A, Sutskever I and Hinton G 2012 ImageNet Classification with Deep Convolutional Neural Networks *J.Advances in neural information processing systems*

[13] Zeiler M D and Fergus R 2013 Visualizing and Understanding Convolutional Networks *European Conference on Computer Vision*. *Springer International Publishing*

[14] Szegedy C, Wei L and Yangqing J 2014 Going Deeper with Convolutions *IEEE*

[15] Ioffe S, Szegedy C 2015 Batch Normalization Accelerating Deep Network Training by Reducing Internal Covariate Shift *J.arXiv*

[16] Sun M, Farhadi A and Seitz S 2014 Ranking Domain-Specific Highlights by Analyzing Edited Videos *European Conference on Computer Vision. Springer International Publishing*

[17] Szegedy C, Ioffe S and Vanhoucke V 2016 Inception-v4,Inception-ResNet and the Impact of Residual Connections on Learning *J.arXiv*

[18] Simonyan K, Zisserman A 2014 Very Deep Convolutional Networks for Large-Scale Image Recognition *J.Computer ence*

[19] Kaiming H, Xiaoyu Z and Shaoqing R 2016 Deep Residual Learning for Image Recognition *IEEE Conference on Computer Vision*

[20] Kaiming H, Xiaoyu Z and Shaoqing R 2016 Identity Mappings in Deep Residual Networks *IEEE Conference on Computer Vision*

[21] Saining X, Girshick R and Dollár, Piotr 2017 Aggregated Residual Transformations for Deep Neural Networks *CVPR.IEEE*

[22] Gao H, Zhuang L 2016 Densely Connected Convolutional Networks *IEEE*

[23] Mingxing T, Quoc V 2019 EfficientNet Rethinking Model Scaling for Convolutional Neural Networks *J.arXiv*

[24] Jie H, Li S and Albanie S 2017 Squeeze-and-Excitation Networks *IEEE.J.Transactions on Pattern Analysis and Machine Intelligence* p99

[25] Youngwan L, Joong-won H, Sangrok L, Yuseok B and Jongyoul P 2019 An Energy and GPU-Computation Efficient Backbone Network for Real-Time Object Detection *CVPR*

[26] Girshick R, Donahue J，Darrell T and Malik J 2014 Rich feature hierarchies for accurate object detection and semantic segmentation *CVPR*

[27] Kaiming He, Xiaoyu Z and Shaoqing R 2014 Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition *IEEE.J.Transactions on Pattern Analysis* p16

[28] Girshick R 2015 Fast r-cnn *ICCV.IEEE*

[29] Shaoqing R, Kaiming He and Girshick R 2016 Faster R-CNN Towards Real-Time Object Detection with Region Proposal Networks *J.arXiv*

[30] Jifeng D, Yi L, Kaiming H and Shaoqing R 2016 R-FCN Object Detection via Region-based Fully Convolutional Networks *J.arXiv*

[31] Bell S, Zitnick C L and Bala K 2016 Inside-Outside Net Detecting Objects in Context with Skip Pooling and Recurrent Neural Networks *J.arXiv*

[32] Lin, T Y，Dollár P，Girshick R，Kaiming H and Hariharan B 2017 Feature Pyramid Networks for Object Detection *CVPR*

[33] Kaiming H, Georgia G and Piotr D Mask R-CNN 2017 *IEEE.J.Transactions on Pattern Analysis*

[34] Wei L, Anguelov D and Erhan D 2016 SSD Single Shot MultiBox Detector *J.arXiv*

[35] Redmon J, Farhadi A 2017 YOLO9000 Better Faster Stronger *IEEE.Conference on Computer Vision* pp6517-6525

[36] Tsung Yi L, Goyal P and Girshick R 2017 Focal loss for dense object detection *IEEE J.Transactions on Pattern Analysis* pp2999-3007

[37] Redmon J, Farhadi A 2018 YOLOv3 An Incremental Improvement *J.arXiv*

[38] Bochkovskiy A, Chien Yao W and Hongyuan L 2020 YOLOv4 Optimal Speed and Accuracy of Object Detection *J.arXiv*

[39] Tan M, Pang R and Quoc V 2020 EfficientDet Scalable and Efficient Object Detection *CVPR.IEEE*

[40] Redmon J, Divvala S and Girshick R 2015 You Only Look Once Unified, Real-Time Object Detection *CVPR*

[41] Law H, Jia D 2018 CornerNet Detecting Objects as Paired Keypoints *J.International Journal of Computer Vision J.arXiv*

[42] Kaiwen D, Song B and Lingxi X 2019 CenterNet Keypoint Triplets for Object Detection *J.arXiv*

[43] Law H, Yun T and Russakovsky O 2019 CornerNet-Lite: Efficient Keypoint Based Object Detection *J.arXiv*

[44] Tian Z, Shen C, Chen H and Tong He 2020 FCOS Fully Convolutional One-Stage Object Detection *ICCV*

[45] Lee Y, Park J 2019 CenterMask Real-Time Anchor-Free Instance Segmentation *J.arXiv*