ISSN: 2414-1895 DOI: 10.6919/ICJE.202209_8(9).0056

Review of Target Detection Algorithms

Yingzhe Shao[1], Lin Tang[2,*], Xinyi Liu[3], Hongli Wang[1], Ruiyu Yang[1]

[1] Chongqing University of Posts and Telecommunications, Chongqing 400065, China

[2] University of Electronic Science and Technology of China, Chengdu 611700, China

[3] Henan Polytechnic University, Henan 454003, China

Abstract

The task of object detection is to find all the objects of interest in the image and determine their categories and positions, which is one of the core problems in the field of computer vision. Target detection is divided into two series -- RCNN series and YOLO series. RCNN series is a representative algorithm based on region detection. RCNN series algorithms are mainly used in target detection. The classical target detection algorithm uses the sliding window method to judge all possible regions in turn. Selective Search method is used in RCNN to extract a series of candidate regions which are more likely to be objects in advance, and then only features are extracted from these candidate regions (using CNN) for judgment.

Keywords

Computer Vision; Deep Learning; Object Detection.

1. Introduction

Since 2011, continued in a ferment of artificial intelligence, and target detection[1] is a computer vision and digital image processing is a popular direction, widely used in robot navigation, intelligent video surveillance, industrial testing, aerospace and other fields, through computer vision to reduce the consumption of the human capital, has important practical significance. Therefore, target detection has become a research focus in theory and application in recent years. It is an important branch of image processing and computer vision, and the core part of intelligent monitoring system. At the same time, target detection is also a basic algorithm in the field of pan-identity recognition. It plays an important role in face recognition, gait recognition, crowd counting, instance segmentation and other tasks.

2. R-CNN Series Algorithm Process

2.1 R-CNN

2.1.1 Sketch

R-CNN is Region-CNN, which is based on convolutional neural network (CNN), linear regression, support vector machine (SVM) and other algorithms to achieve target detection technology. It is the first algorithm that successfully applies deep learning to target detection. R-CNN follows the traditional idea of target detection and also adopts the extraction frame to carry out target detection through four steps of feature extraction, image classification and non-maximum suppression for each frame. However, in the feature extraction step, traditional features (such as SIFT and HOG feature, etc.) are replaced with features extracted by deep convolutional network.

ISSN: 2414-1895 DOI: 10.6919/ICJE.202209_8(9).0056

2.2 SPP-Net

2.2.1 Sketch

The speed and accuracy of R-CNN are improved. While the accuracy of R-CNN is reached, its speed is 24-102 times that of R-CNN, which has a great improvement in computing speed, but it still can not meet the real-time requirements.

2.2.2 Innovation Point

The last pooling layer is replaced by the SPP layer so that feature maps of different sizes are mapped to a fixed size. A spatial pyramid is introduced to accommodate feature maps of different sizes. The feature extraction of the whole image is carried out to accelerate the operation speed, and the strategy of weight sharing is proposed.

2.3 Fast-RCNN

2.3.1 Sketch

Compared with SPP-NET, FAST-RCNN does not use SVM classification, but neural network classification, directly using the full connection layer, the full connection layer has two outputs, one is responsible for classification, the other is responsible for regression. The same network can complete feature extraction, category judgment and box regression.

2.3.2 Innovation Point

Replace the SPP layer with the Rol pooling layer, allowing feature maps of different sizes to be mapped to a fixed size. Training can update all networks. Functional caching does not require disk storage. R-CNN to Fast R-CNN[2] greatly simplifies the training process of target detection framework from 4 independent training processes to 2 independent training processes. SVD is used to reduce the number of parameters of the fully connected layer by replacing a single large fully connected layer with two adjacent fully connected layers without nonlinear activation in between. Fast R-CNN is composed of classification loss and positioning loss by loss function calculation, so as to optimize these two parts at the same time. In addition, bb REGRESsor losses were changed to Smooth-L1 paradigm with stronger robustness.

2.4 YOLOv2

2.4.1 Sketch

In 2017, the author proposed YOLOv2 algorithm to solve the shortcomings of inaccurate positioning and low recall rate compared with two-stage method.

2.4.2 Innovation Point

Batch Normalization(BN) layer is added after each convolutional layer and the dropout layer is removed. Batch Normalization layer plays a regularization role to improve model convergence speed and prevent model overfitting. YOLO V2 improves the mAP by 2% by using the BN layer. CNN feature extracer is first pre-trained on ImageNet classification data set. Since YOLOv1 uses 224*224 input to pre-train 160 epoches, the input is adjusted to 448*448 and the training continues. Anchor boxes mechanism is adopted to predict boundary boxes, and a pooling layer is removed to make the convolution layer output a higher resolution. In the detection model, the input size is no longer 448*448 in the pre-training, but 416*416. Since the total step size of sampling under YOLOv2 is 32, a feature map of 13*13 will be obtained for images with the size of 416*416. K-means method is adopted to cluster the bounding boxes of the training set and try to find the appropriate Anchor box. It should be noted here that, due to the different sizes of boxes, the standard Euclidean distance is not adopted, but the IOU between box and the cluster center box is used as the distance indicator. A Passthrough layer is added. The function of this layer is similar to ResNet, connecting the feature map of 26*26 of the previous layer with the feature map of 13*13 of the current layer, so as to better detect small objects. Multi-scale Training is introduced, that is, during Training, every 10 epoches are trained, the network will choose another type of size input, which increases by multiple of 32.

ISSN: 2414-1895 DOI: 10.6919/ICJE.202209_8(9).0056

2.5 YOLOv3

2.5.1 Sketch

YOLOv3 was proposed in 2018, including the use of residual model DarkNET-53 and the adoption of FPN architecture for multi-scale detection.

2.5.2 Innovation Point

FPN architecture is adopted. Three feature graphs are output in total. The first one is downsampled 32 times, the second 16 times and the third 8 times. Proposed the ignore sample, which is divided into three parts: positive example, negative example, ignore sample. Take a GT and all 4032 enclosures to calculate the IOU value. Take the largest box as a positive example. If the IOU of all GT is smaller than the threshold except the positive example, set it as the negative example. Finally, the sample is ignored except for collation. There is a GT and its IOU greater than the threshold.

2.6 YOLOv4

2.6.1 Sketch

The core idea of YOLOv4 is basically the same as before, but a lot of improvements have been made to the sub-structure from data processing, backbone network, network training, activation function, loss function and other aspects.

2.6.2 Innovation Point

The CSP structure is integrated into Darknet53 and a new backbone network CSPDarknet53 is generated; SPP spatial pyramid pooling was used to expand the receptive field; SPP structure was introduced to increase the receptive field, and the maximum pooling of 1*1, 5*5, 9*9 and 13*13 was adopted to carry out multi-degree fusion, and the output was concat fusion according to the channel; PAN structure is introduced into the Neck part, that is, the form of FPN+PAN; Mish activation functions are introduced; Mosaic data enhancement was introduced; CIOU_loss is used in training and DIOU_nms is used in prediction.

2.7 YOLOv5

2.7.1 Sketch

YOLOv5[3] version of ultralytics lc is a minor tinker with YOLOv4.

2.7.2 Innovation Point

Extend the CSP structure of the v4 backbone network to the NECK structure. The FOCUS operation was added, but this operation was removed in the subsequent 6.1 version and replaced with a 6x6 convolution. SPPF structure is used instead of SPP.

2.8 YOLOv6

2.8.1 Sketch

YOLO V6 was introduced by Meituan. the main work was to bring the 2021 RepVGG architecture to YOLO in order to be more suitable for GPU devices. The algorithm of YOLOv6 is similar to that of YOLOv5 (Backbone + Neck) +YOLOX (head).

2.8.2 Innovation Point

The backbone network was changed from CSPDarknet to EfficientRep; Neck constructs rep-PAN based on Rep and PAN; The detection head part imitates YOLOX and performs decoupling operation with a little optimization; Introduce RepVGG; According to RepVGG's idea, a 1x1 convolution branch and an identity mapping branch are added parallel to each 3x3 convolution, and then fused into a 3x3 structure during reasoning, which is more friendly to computation-intensive hardware devices. YOLOv6 further optimizes the time consumption to further improve the performance of YOLO detection algorithm.

ISSN: 2414-1895 DOI: 10.6919/ICJE.202209_8(9).0056

3. Conclusion

Target detection is an important part in the field of machine vision, target detection based on neural network learning depth still exist some difficulties and challenges, such as target detection in terms of pedestrian detection, there is a small target (pedestrians occupy the pixel in the image is very less, such as target accounts for 5% of the whole image), pedestrian density and pedestrian block problem; In face detection, there are many problems such as face has different expressions, target is blocked by another target, target to be detected has various scales, etc. In the aspect of text detection, there are many problems, such as font and language diversity, target text damage and ambiguity, dense text arrangement and so on.

Nowadays, video detection has a very wide range of applications. For example, real-time target detection and tracking in high-definition cameras is of great significance to video surveillance, automatic driving and obstacle detection in vehicle video. In the future, video detection can be studied in the direction of generalization ability, such as making the model more adaptable to the detection requirements of real scenes. It can also be combined with weakly supervised learning to study how to achieve high-precision video detection under the condition of few or zero samples.

References

[1] Fu Miaomiao, Deng Miaolei, Zhang Dexian. A review of deep neural network image target detection algorithms [J]. Computer System Applications, 2022, 31(07): 35-45. DOI: 10.15888/j.cnki.csa.008595.

[2] Ji Chunsheng. Weld defect map recognition based on improved Faster R-CNN [J]. China Chemical Equipment, 2022, 24(02): 26-32+36.

[3] Fu Huijin, Shi Tianyun, Wang Rui, Xu Chengwei, Zhang Wanpeng, Li Wentao. Research on the detection method of intruders in the perimeter of high-speed railway based on improved YOLOv5 [J/OL]. Railway Standard Design: 1-8[2022-07-22] .DOI: 10.13238/j.issn.1004-2954.202203110002.

[4] Li Aijuan, Gong Chunpeng, Huang Xin, Cao Jiaping, Liu Gang. A review of object detection methods for autonomous vehicles [J]. Journal of Shandong Jiaotong University, 2022, 30(03): 20-29.

[5] Li Weiqiang, Wang Dong, Ning Zhengtong, Lu Mingliang, Qin Pengfei. A review of fruit target detection algorithms under computer vision [J]. Computer and Modernization, 2022(06): 87-95.

[6] Ji Chaoqun. Research on Lightweight Target Detection Algorithm for Road Scenes Based on Deep Learning [D]. Changchun University of Technology, 2022. DOI: 10.27805/d.cnki.gccgy.2022.000012.

[7] Tao Libo. Research and implementation of sensor fusion target detection algorithm based on driverless formula racing car [D]. Zhejiang University of Science and Technology, 2021. DOI: 10.27840/d.cnki. gzjkj.2021.000169.