A review of research on object detection based on deep learning

A review of research on object detection based on deep learning

Jun Deng[1,a], Xiaojing Xuan[2,b], Weifeng Wang[3], Zhao Li[4], Hanwen Yao[5], Zhiqiang Wang[6]

[1,3]Xi'an University of Science and Technology, School of Safety Science and Engineee, Xi'an, Shaan Xi

[2,6]Xi'an University of Science and Technology, School of Computer Science and Technology, Xi'an, Shaan Xi

[4]Xi'an University of Science and Technology, School of Energy and Power Engineering, Xi'an, Shaan Xi

[5]Xi'an University of Science and Technology, School of Electrical and Control Engineering, Xi'an, Shaan Xi

[a]dengj518@xust.edu.cn

[b]974734265@qq.com

Abstract—As one of the important tasks in computer vision, target detection has become an important research hotspot in the past 20 years and has been widely used. It aims to quickly and accurately identify and locate a large number of objects of predefined categories in a given image. According to the model training method, the algorithms can be divided into two types: single-stage detection algorithm and two-stage detection algorithm. In this paper, the representative algorithms of each stage are introduced in detail.Then the public and special datasets commonly used in target detection are introduced, and various representative algorithms are analyzed and compared in this field. Finally, the potential challenges for target detection are prospected.

1. INTRODUCTION

Object detection is a basic research direction in the fields of computer vision, deep learning, artificial intelligence, etc. It is an important prerequisite for more complex computer vision tasks, such as target tracking, event detection, behavior analysis, and scene semantic understanding. It aims to locate the target of interest from the image, accurately determine the category and give the bounding box of each target. It has been widely used in vehicle automatic driving, video and image retrieval, intelligent video surveillance[1], medical image analysis[2], industrial inspection[3] and other fields.

Traditional detection algorithms on manually extracting features mainly include six steps: pre-processing, window sliding, feature extraction, feature selection, feature classification and post-processing and generally for specific recognition tasks. Its disadvantages mainly include small data size, poor portability, no pertinence, high time complexity, window redundancy, no robustness for diversity changes, and good performance only in specific simple environments.

In 2012, AlexNet image classification model based on convolutional neural network (CNN) was proposed by Krizhevsjy[4]and others.In the image classification competition of the image dataset

ImageNet[5], they won the competition with a huge advantage of 11% accuracy over the second place using traditional algorithms. Many scholars have begun to apply deep convolutional neural networks to target detection tasks, and have proposed many excellent algorithms. It can be roughly divided into two categoriesa:the single-stage detection algorithm based on region proposal and the two-stage detection algorithm based on regression.

2. TWO-STAGE TARGET DETECTION FRAMEWORK

2.1 R-CNN

In 2014, the R-CNN[6] algorithm was proposed by Girshick, which is the first real target detection model based on convolutional neural networks. The improved R-CNN model achieves 66% mAP. As shown in figure 1, the model first uses the Selective Search to extract approximately 2000 region proposals of each image to be detected. Then the size of each extracted proposals is uniformly scaled to a fixed-length feature vector and these extracted image features are input into the SVM classifier for classification. Finally, a linear regression model is trained to perform the regression operation of the bounding box. Compared with the traditional detection method, the accuracy of the R-CNN does improve a lot, but the amount of calculation is very large, and the calculation efficiency is too low. Secondly, directly scaling the region proposal to a fixed-length feature vector may cause object distortion.

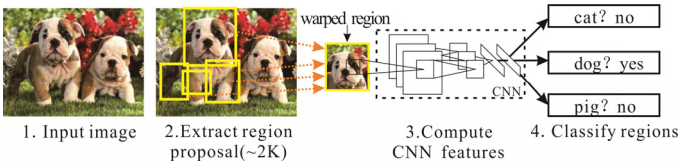

Figure 1. R-CNN architecture

2.2 SPP-Net

In 2015, the Spatial Pyramid Pooling (SPP) model proposed by He[7]. solves the problems of low detection efficiency and the need for fixed input size image blocks in R-CNN. This algorithm extracts the features of the regions proposal on the feature map after the original image has passed through the convolution layer, and all the convolution calculations are performed only once. At the same time, the spatial pyramid pooling layer is added after the last convolutional layer, and the feature of region proposal is passed through the spatial pyramid pooling layer to extract the feature vector of fixed size. Compared to the R-CNN , Spp-Net performs feature extraction on the entire image only once, avoiding repeated calculations. However, it still has the same shortcomings as R-CNN: 1) Multi-step training steps are complicated. 2) Separate SVM classifiers need to be trained and additional regressors are required.

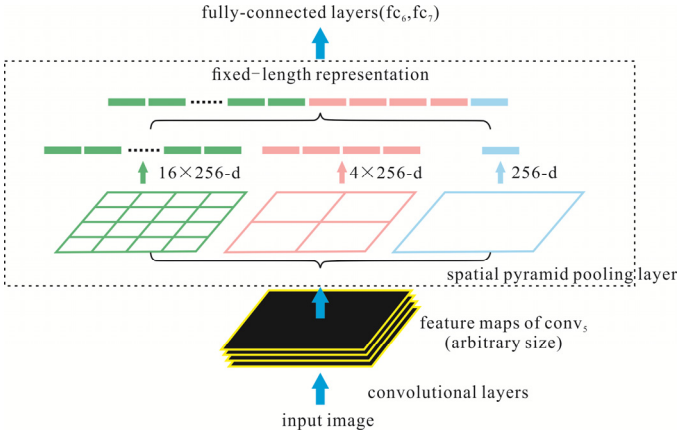

Figure 2. SPP-Net architecture

2.3 Fast R-CNN

In 2015, the Fast R-CNN[8] model was proposed by Girshick. In the joint dataset of VOC2007 and VOC2012[15], the mAP reaches 70.0%. Its structure is shown in figure 2. Compared with R-CNN, Fast R-CNN has made three changes. First, it replaced the SVM used in R-CNN with softmax function for classification. Secondly, the model draws on the pyramid pooling layer in SPP-Net, and uses the region of interest pooling layer to replace the last pooling layer in the convolutional layer, so as to transform the feature of the candidate box into a feature map with fixed size for access to the full connection layer. Finally, the last softmax classification layer of the CNN network is replaced by two parallel fully connected layers. However, it still cannot meet the needs of real-time detection.

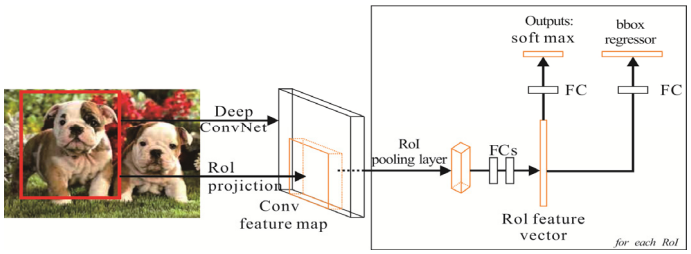

Figure 3. Fast R-CNN architecture

2.4 Faster R-CNN

The Faster R-CNN[9] model proposed by Ren uses region proposal networks to replace the previous Selective Search method to generate region proposal. The model is divided into two modules, one of which module is a fully convolutional neural network used to generate all region proposal, and the other is the Fast R-CNN detection algorithm. A set of convolutional layers is shared between these two modules. The input image is propagated forward through the CNN network to the final Shared convolutional layer. On the one hand, the feature map for the input of the RPN network is obtained; on the other hand, the image is propagated forward to the specific convolutional layer to produce a higher-dimensional feature map. Although Faster R-CNN is excellent in detection accuracy, it still cannot achieve real-time detection.

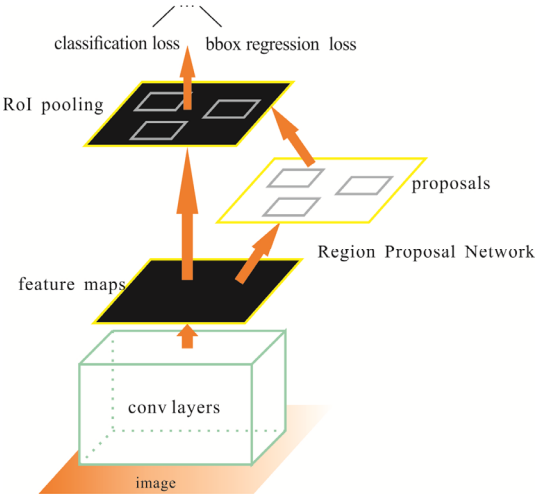

Figure 4. Faster R-CNN architecture

3. ONE-STAGE TARGET DETECTION ALGORITHM

3.1 YOLOv1

In 2016, the YOLOv1[10] object detection model was proposed by Joseph Redmon. YOLOv1 detection model does not require the extraction process of region proposal. The entire detection model is just a

simple CNN network structure. Its core idea is to use the entire graph as the input of the network and directly return the location and category of the bounding box at the output layer. First, an image is divided into an S*S grid, each grid cell predicts B bounding box and confidence scores for these boxes. That is, each cell predicts B*(4+1) values in total. On a single TitanX, its detection speed can reach 45fps per second, fully real-time detection. However, YOLO produces fewer background errors, but has poor recognition performance when dealing with objects in group form.

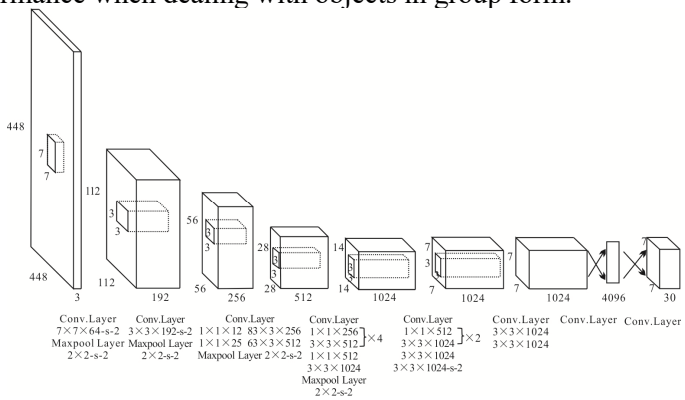

Figure 5. YOLOv1 architecture

3.2 YOLOv2

In 2016, Redmon proposed YOLOv2[11] model. The main goal is to improve the recall and localization while maintaining classification accuracy. YOLOv2 uses a new fully convolution feature extraction network Darknet-19, which contains a total of 19 convolutional layers and 5 maximum pooling layers. By adding a batch normalization layer to the convolutional layer and removing dropout, introducing anchor box mechanism, using k-means clustering on the training set bounding box, and multi-scale training, the recall and accuracy are significantly improved. However, the detection of targets with high overlap and small target still needs to be improved.

3.3 YOLOv3

YOLOv3[12] proposed by Redmon is the most balanced object detection model for detection speed and detection accuracy by far. In terms of category prediction, YOLOv3 is mainly to change the original single-label classification into multi-label classification, and replace the original softmax layer used for single-label multi-classification with a logistic regression layer for multi-label multi-classification. At the same time, the model uses a combination of multiple scales for prediction. It adopts the upsampling fusion method similar to FPN, and finally merges three scales, which improves the detection effect of small targets significantly. The network structure of this model adopts deeper feature extraction network Darknet-53. Although the YOLOv3 model further improves the detection speed and the detection effect of small targets has also been significantly improved, the detection accuracy has not been significantly improved, especially when IOU>0.5.

3.4 SSD

In 2016, the SSD[13] model was proposed by Liu. The model uses the regression idea used in the YOLO algorithm and draws on the concept of the anchor box proposed in the Faster R-CNN detection model. In order to improve the effect of multi-scale object detection, SSD model proposes to use both the bottom and high level feature maps for detection. The basic architecture is VGG and last two fully connected layers are replaced by convolutional layers. SSD draws on the anchor mechanism in the RPN network. SSD achieves 74.3% mAP on VOC2007 at 59 FPS on a Nvidia Titan X. However, the classification result of SSD for small targets is poor, and the feature maps of different scales are independent, leading to the simultaneous detection of the same object by boxes of different sizes.

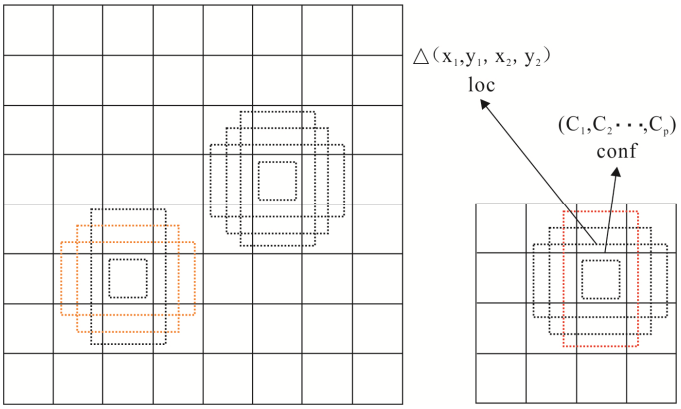

Figure 6. SSD architecture

3.5 YOLOv4

In 2020, the YOLOv4[14] was proposed by Alexey Bochkovskiy and it achieves a new benchmark with the best balance of speed and accuracy. In theory, YOLOv4 is not much innovative. It adds Weighted Residual Connection, Cross Stage Partial connection, Cross mini Batch Normalization, Self adversarial training, Mish activation, Mosaic data augmentation, DropBlock, CIou on the basis of the original YOLO detection framework. CSP Darknet53 is selected as the backbone network, and on this basis, SPP module was attached to increase the receptive field to separate the most important context features. Meanwhile, YOLOv4 uses PANet instead of FPN used in YOLOv3 as the path aggregation method, and follows the head structure of YOLOv3. Compared with the YOLOv3, the accuracy and speed of the YOLOv4 are improved by 10% and 20% respectively.

4. DATASETS AND PERFORMANCE COMPARISON OF VARIOUS ALGORITHMS

4.1 Dataset

As early as 1956, the concept of "artificial intelligence" was proposed. But it was not until 2012 that artificial intelligence began to usher in a peak. This is mainly due to the rising data volume, computing power and the emergence of machine learning algorithms. The development of detection systems is closely related to the explosion of data volume. This is because the performance test and algorithm evaluation need to be obtained through dataset, and dataset is also a powerful driving force to promote the research field of detection approaches. The parameters of common public data sets are shown in table 1.

TABLE I. PUBLIC DATA SET AND ITS PARAMETERS

Dataset	Amount	Sort	Size/Pixel	Year
Caltech101[18]	9145	101	300×200	2004
PASCAL VOC 2007	9963	20	375×500	2005
PASCAL VOC 2012	11540	20	470×380	2005
Tiny Images[19]	80 million	53464	32×32	2006
Scenes15	4485	15	256×256	2006
Caltech256	30607	256	300×200	2007
ImageNet	14197122	21841	500×400	2009
SUN[16]	131072	908	500×300	2010
MS COCO[17]	328000	91	640×480	2014
Places[20]	More than10 million	434	256×256	2014
Open Images	More than 9 million	More than 60 million	Different size	2017

4.2 Performance comparison of various algorithms

Table 2 makes statistics and comparisons of single-stage and two-stage detection algorithms.

TABLE II. COMPARISON OF OBJECT DETECTION ALGORITHMS

Method	Backbone	Size/Pixel	Test	mAP/%	fps
YOLOv1	VGG16	448×448	VOC 2007	66.4	45
SSD	VGG16	300×300	VOC 2007	77.2	46
YOLOv2	Darknet-19	544×544	VOC 2007	78.6	40
YOLOv3	Darknet-53	608×608	MS COCO	33	51
YOLOv4	CSP Darknet-53	608×608	MS COCO	43.5	65.7
R-CNN	VGG16	1000×600	VOC2007	66	0.5
SPP-Net	ZF-5	1000×600	VOC2007	54.2	-
Fast R-CNN	VGG16	1000×600	VOC2007	70.0	7
Faster R-CNN	ResNet-101	1000×600	VOC2007	76.4	5

5. CONCLUSION

As one of the most basic and challenging problems in computer vision, object detection has received great attention in recent years. Detection algorithms based on deep learning have been widely applied in many fields, but deep learning still has some problems to be explored:

1) Reduce the dependence on data.
2) To achieve efficient detection of small objects.
3) Realization of multi-category object detection.

REFERENCES

[1] Wu, R.B. Research on Application of Intelligent Video Surveillance and Face Recognition Technology in Prison Security. China Security Technology and Application. 2019,6: 16-19.
[2] Tian, J.X., Liu, G.C., Gu, S.S., Ju, Z.J., Liu, J.G., Gu, D.D. Research and Challenge of Deep Learning Methods for Medical Image Analysis. Acta Automatica Sinica,2018, 44: 401-424.
[3] Jiang, S.Z., Bai, X. Research status and development trend of industrial robot target recognition and intelligent detection technology. Guangxi Journal of Light Industry, 2020, 36: 65-66.
[4] Krizhevsky, A., Sutskever, I., Hinton, G. ImageNet Classification with Deep Convolutional Neural Networks. Advances in Neural Information Processing Systems,2012, 25: 1097-1105.
[5] Russakovsky, O., Deng, J., Su, H., et al. ImageNet Large Scale Visual Recognition Challenge. International Journal of Computer Vision,2015, 115: 211-252.
[6] Girshick, R., Donahue, J., Darrel, T.,Malik, J. Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation. In: Computer Vision and Pattern Recognition. Columbus.2014, pp. 580-587.
[7] He, K.M., Zhang, X.Y., Ren, S.Q., Sun, J. Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition. IEEE Transactions on Pattern Analysis & Machine Intelligence,2015, 37: 1904-1916.
[8] Girshick, R. Fast R-CNN.In: Proceedings of the IEEE international conference on computer vision. Santiago.2015, pp. 1440-1448.
[9] Ren, S.Q., He, K.M., Girshick, R., Sun, J. Faster R-CNN: towards real-time object detection with region proposal networks. In: Advances in neural information processing systems. Montreal.2016, pp. 91-99.
[10] Redmon, J., Divvala, S., Grishick, R., Farhadi, A. You Only Look Once: Unified, Real-Time Object Detection. In: Computer Vision and Pattern Recognition. Las Vegas.2016, pp. 779-788.

[11] Redmon, J., Farhadi, A. YOLO9000: better, faster, stronger. In: Computer Vision and Pattern Recognition. Hawaii.2017, pp. 7263-7271.
[12] Redmon, J., Farhadi, A. (2018) Yolov3: An incremental improvement. arXiv: Computer Vision and Pattern Recognition.
[13] Liu, W., Anguelov, D., Erhan, D., et al. SSD: Single Shot MultiBox Detector. European Conference on Computer Vision, 2016, pp. 21-37.
[14] Bochkovskiy, A., Wang, C.Y., Liao, H.Y.M. YOLOv4: Optimal Speed and Accuracy of Object Detection. arXiv: Computer Vision and Pattern Recognition, 2020.
[15] Everingham, M., Eslami, S.M.A., Van Gool, L. The Pascal Visual Object Classes Challenge: A Retrospective. International Journal of Computer Vision,2015, pp.98-136.
[16] Xiao, J.X., Ehinger, K.A., Hays, J.,Torralba, A.,Oliva, A. SUN Database: Exploring a Large Collection of Scene Categories. International Journal of Computer Vision, 2016,pp.3-22.
[17] Lin T Y , Maire M , Belongie S , et al. Microsoft COCO: Common Objects in Context. European Conference on Computer Vision, 2014, pp.740-755.
[18] Li, F.F., Rob, F., Pietro, P. Learning generative visual models from few training examples: An incremental Bayesian approach tested on 101 object categories. Computer Vision and Image Understanding,2007,pp. 59-70.
[19] Torralba, A., Fergus, R., Freeman, W.T. 80 Million Tiny Images: A Large Data Set for Nonparametric Object and Scene Recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence,2008, pp.1958-1970.
[20] Zhou, B., Lapedriza, A., Khosla, A., et al. Places: A 10 million Image Database for Scene Recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2016, pp.1452-1464.