Proceedings of the 3rd International Conference on Signal Processing and Machine Learning
DOI: 10.54254/2755-2721/5/20230656

Research advanced in object detection based on deep learning

Dongyang Li

Xi'an International University, Xi'an, 710077, China

190123020011@post.xaiu.edu.cn

Abstract. In the field of computer vision, identifying specific objects in images and predicting their location and class have long been hot research topics.Algorithms for early object detection rely on traditional handcrafted features, and the speed and precision of their detection cannot meet the needs of real-world applications. Rapid development in convolutional neural networks has accelerated the development of object identification systems based on deep learning.Existing deep learning-based object identification algorithms mostly use two-stage detection and single-stage detection, according to various detection frameworks. In this paper, around the above two types of frameworks, the latest research progress in the field of object detection is systematically introduced. Specifically, we first introduce representative object detection algorithms, including their design ideas, basic processes, and advantages and disadvantages. Second, using widely-used datasets, we objectively compare the effectiveness of several detection techniques. Finally, we summarize the unsolved problems in the detection of objects and talk about the route this topic will take going forward.

Keywords: Object Detection, Deep Learning, Multi-class, Lightweighting.

1. Introduction

In terms of computer vision, finding things of interest in images, precisely classifying each object, and providing the bounding box of each object have all been hot topics for research. Deep learning-based object detection has made strides in both detection accuracy and speed thanks to the quick development of convolutional neural networks and powerful GPU hardware. Deep learning-based object detection is currently also commonly employed. For example, in the field of unmanned driving, (1) Detection of lane lines, including various lane lines (dotted lines, solid lines, double lines, etc.). (2) Detection of static objects such as traffic lights, street signs, and so on.

The classical object detection period (before to 2014) and the deep learning-based detection period are the two major historical periods in which object detection has evolved (after 2014). The sliding window method is typically used by traditional object detection algorithms to choose region of interest in the detected image by sliding a window across the image one by one, respectively, for each window of the sliding feature extraction, after extraction of the extracted features using machine learning algorithms, and finally analyze whether the window contains a certain class of objects . However, since the objects in the picture vary in size and are not fixed in size, if a fixed window is used for sliding when the object is relatively large, it will appear that the object cannot be a completely framed pillar object but only a part of the object. However, when the object is small, the window will frame in features that do not belong to the small object will frame in some background, both of w hich will have an impact on the results. Though the design of various sized windows alleviates the problem to a certain extent, but also brings a decrease in detection speed. In summary, traditional object detection algorithms have the following three drawbacks: (1)The recognition effect is insufficient, and the

DOI: 10.54254/2755-2721/5/20230656

accuracy rate is low. (2) More computationally intensive and slower. (3) May produce multiple correctly identified results.

The R-CNN technique was put forth by Girshick et al. in 2014, object detection has entered the era of deep learning. The two main types of object detection approaches in the deep learning era are generally two-stage detection and one-stage detection, depending on the detection idea. (1) Two-stage detection consists of two steps. The first step is to extract the object region first; the next step is then to classify and recognize the region by CNN. It is characterized by high accuracy but slow speed. (2) The single-stage detection algorithm is a "one-step" process. object detection can be achieved by extracting features only once, so it is much faster than the two-stage one and suitable for industrial use, but the detection accuracy will be slightly lower.

Focusing on above two frameworks,this paper analyzes and summarizes the state-of-the-art of target detection algorithms based on deep learning, provides a detailed introduction to the general target detection data set, presents the experimental findings of various algorithms on the common data set, and projects the direction of object detection's future development.

2. Two-Stage detection algorithms

The R-CNN family is the most prevalent Two-Stage algorithm, which includes R-CNN, SPP-Net, Fast R-CNN, Faster R-CNN, and so on.

R-CNN s a region-based convolutional neural network algorithm that applies a strategy of region recommendation on convolutional neural networks to form a bottom-up target localization model.[1] It was the first algorithm to successfully apply deep learning to object detection.The input image is first searched for a part of the image that has a high probability of belonging to the object using the Selective Search method to generate approximately 2000 candidate regions. Compared with exhaustive enumeration, Selective Search has a significant speedup. Next, the candidate regions are fixed into images of the same size, and then the candidate regions that may contain objects are taken out and fed into the CNN network to obtain a 4096-dimensional feature vector, after which the feature vector is classified using SVM to predict the probability value of belonging to each class of the objects contained in the candidate regions. Finally, the output result is suppressed by a non-maximal value. But R-CNN is too computationally intensive, so it is very time-consuming.

Prior to this, all neural networks required a fixed-size image as input. In order to detect images of different sizes, we must first convert them into uniform-size images, which results in some visual information loss and distortion and lowers the effectiveness of recognition. SPP-Net improves on these existing drawbacks. SPPNet adds an ROI pooling layer to the normal CNN structure, allowing the input images to be of arbitrary size. In R-CNN, the candidate region must be fixed into a uniform size image before feature extraction; however, SPP-Net adds the RIO pooling layer, which can first convolve the image once to get the convolutional features of the whole image, and then only need to find the corresponding position of the convolutional features in the candidate region, and then use the RIO pooling layer to perform feature extraction on the corresponding candidate frame. The RIO pooling layer is then used to extract the features from the corresponding candidate frames. Therefore, SPP-Net is much faster than R-CNN with similar accuracy. However, its architecture is similar to R-CNN, so it also has the disadvantages of R-CNN.

Fast R-CNN is an enhancement to the R-CNN and SPPNet algorithms. Each candidate region is fed into the CNN model individually in R-CNN in order to calculate the feature vector, which is a very time-consuming task. For the candidate box region in the original image, Fast R-CNN does the same as in SPP-Net. They both map it to the corresponding region of the convolutional feature, i.e., the ROI projection in Fig.After that, SPPNet uses SVM to classify the features, while Fast R-CNN gives up using SVM and uses the fully connected layer directly, which has two outputs, one for classification, i.e., softmax in the figure 1, and the other for box regression, i.e., bbox regressor in the figure 1.

DOI: 10.54254/2755-2721/5/20230656

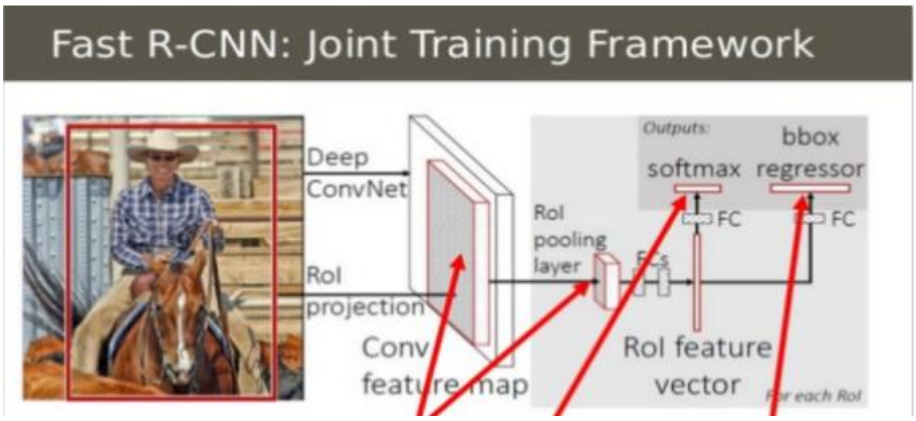

Figure 1. The framework of Fast-RCNN.

The Ren et al-proposed Faster R-CNN algorithm. Faster R-CNN includes a region proposal network (RPN) behind the convolutional layer in place of the selective search strategy used by Fast R-CNN because it is much faster.This significantly accelerates the creation of detection frames. The experimental results on the PASCAL VOC 2007 dataset demonstrate that with the identical VGG backbone network, the difference in mAP between Faster R-CNN and Fast R-CNN is not significant, however, both training and testing time are both significantly reduced.[2][3][4][5]

3. One-stage detection algorithms

The one-stage method allows for a single full training of shared features, considerably increases speed, and guarantees a specific accuracy rate. Typical algorithms are YOLO, SSD, YOLOv2, YOLOv3, etc. Redmon et al. were the first to propose the YOLO algorithm in 2015 as the one-stage object detection approach; The YOLO series algorithm has a wide range of practical applications and is a high-precision algorithm that can fulfill the needs of real-time detection (FPS>30). Treating object detection as a single regression problem is the central tenet of YOLO.It divides the image into SxS grids first.and if the center of the object falls in which grid. It is that grid that is responsible for monitoring that object. The advantage of this algorithm is that the method of dividing the grid, which avoids a lot of repeated calculations, allows the YOLO algorithm to achieve a fast detection speed of 45 frames per second with an mAP of 63.4 on the Pascal VOC detection challenge dataset. The disadvantages are poor detection for groups of small targets, multiple adjacent targets or targets with abnormal dimensions, relatively poor accuracy, and poor detection for very small objects.

SSD's base network is VGG-16-Atrous. and then tops it off with another convolutional layer to get more feature maps for detection.SSD uses multiple feature layers, and the feature layer sizes are 38*38, 19*19, 10*10, 5*5, 3*3, and 1*1, for a total of six different feature map sizes. Large-size feature maps, using shallow information, predict small targets; small-size feature maps, using deep information, predict large targets. The multi-scale detection approach allows for adequate detection and better detection of small targets.

YOLOv2 is much better than YOLOvl. In recognition of the type, accuracy, speed, and positioning accuracy have been greatly improved. Its recognition object is able to detect 9000 different objects, also called YOL09000. In 2018, Redmon J et al. proposed YOLOv3 based on YOLOv2. The improvements of YOLOv3 are: the network structure is adjusted to Darknet-53, object detection using multi-scale features, and object classification with logistic instead of softmax. The use of the residual network model Darknet-53 for feature extraction improves the overall performance of the algorithm.Using feature fusion at many scales to detect objects, the feature pyramid network (FPN) is able to obtain fine-grained features by using three different scales of feature maps to obtain more useful information about small targets and so increase the algorithm's accuracy for small object detection. Using Logistic, you can support multi-label objects (e.g., an apple can have two labels, "fruit" and "apple").YOLOv3 is very fast, and when using COCO mAP50 as an evaluation metric, YOLOv3 is 3-4 times faster than other models with comparable accuracy.[6][7]

DOI: 10.54254/2755-2721/5/20230656

YOLOv4 has done some improvement operations on the input side, mainly including SAT self-adversarial training, cmBN, Mosaic data enhancement, and YOLOv4 algorithm Yolov4 adds SPP module and also references FPN+PAN structure.[8] YOLOv5 is an improvement on YOLOv4, using the PyTorch framework instead of the Darknet framework of the YOLOv4 version, with no major changes in the network layer and the fastest detection speed of 140 FPS. yOLOv5 is small. yOLOv5 has a weight file of 27 megabytes. yOLOv4 (with the Darknet architecture) has a weight file of 244 MB. Since YOLOv5 is small and easy to deploy, its portability to mobile devices is also a highlight.[9]

4. Experiments and performance analysis

4.1. Common datasets

Current popular datasets for generic object detection tasks are PASCAL VOC2007, PASCAL VOC2012, MS COCO, ImageNet, Open Images, LIVS, etc.

The PASCAL VOC dataset is mainly used for tasks using computer vision such as image classification and object detection, and the VOC2007 and VOC2012 datasets are commonly used. They contain 20 common categories, and each image has a corresponding XML file labeled with the location and category of each target to be detected.

MS COCO is a huge dataset containing object detection, segmentation, and captioning. Used for tasks for example human key point detection, semantic segmentation, object detection, and recognition in scenarios, the targets in this dataset are mainly intercepted from complex everyday scenes, so it is one of the most challenging datasets. The dataset uses annotation files in JSON format to give segmentation information at the target pixel level in each image, and the dataset contains a total of 80 object classes of targets to be detected, with a large variation in scale between targets and a great number of small target objects.

The ImageNet dataset is used for tasks include object detection, image classification, and scene classification and consists of around 14.2 million photos in over 20,000 categories, and in at least one million images, as well as pictures with annotations for object positions and specific category annotations, making it the largest database for picture recognition in the world today. The object detection task is a significant dataset with 200 object classes, and the annotations for each image are saved in XML files in PASCAL VOC data format.

Open Image is a dataset released by the Google team image segmentation, object detection, image categorization, visual relationship detection, and image description. In 2016, Google launched Open Images, and Open Image contains about 9 million annotated images. Updated in 2018, Open Images V4 includes 600 object categories and 15.4 million bounding boxes. Google 2020 released Open Images V6 to add a large number of new visual relationship annotations and human action annotations. Google hopes that this large and diverse training set will stimulate research into more advanced instance segmentation models.

4.2. Evaluation metrics

mAP (mean Average Precision), which is the most important one in the object detection algorithm, is mAP = the sum of the average precision of all categories divided by all categories. The mAP must be of size in the interval [0,1], and the higher, the better. IOU is the intersection and merges ratio, the ratio of the area of the intersection of the predicted box and the real box to the area of the merged area. @IoU=0.5 denotes that when IoU is set to 0.5, the AP of all categories of images is calculated, and then the average of all categories is found, that is, mAP. @IoU=0.5-0.95 means that IoU is tested every 0.05 map between 0.5-0.95. and then averaged at the end.

4.3. Comparison of the effect of different algorithms

From the Table 1, we can understand that with the development of deep learning, the speed of Two-Stage target detection algorithm is also increasing but the speed of YOLO series calculation is still far ahead. Two-Stage target detection approach is unsuitable for detecting scenarios with low real-time requirements, such as river monitoring, crop monitoring, etc. One-Stage target detection approach is

DOI: 10.54254/2755-2721/5/20230656

appropriate for real-time monitoring, such as automatic driving, online monitoring of workers wearing helmets.

5. Summary and Prospects

There is a wide range of promising applications for object detection widely used in pedestrian detection, intelligent video surveillance, unmanned vehicles, industrial inspection, robot navigation, and other fields. And in the past few years, object detection has made tremendous progress. Compared with traditional methods, the performance of today's object detection models has been significantly improved, and even some specific areas of algorithms have now reached a high level of achievement. However, there are still some challenging challenges that need to be addressed.

(1) Small object detection: In the subject of object detection, this is a challenging problem, such as traffic sign and signals recognition, face recognition, and other fields related to small object detection.

(2) Dataset scenario and quality issues: The quality of the dataset is not high enough, so it leads to a decrease in algorithm detection accuracy, and there are also no complete datasets established in many areas.

(3) Weakly supervised detection: Because there are huge amounts of unlabeled images in real life, it is very important to study how to use weakly supervised learning for object detection algorithms.[10]

(4) Lightweight object detection: improve the detection speed by reducing the amount of computation while ensuring that the difference in accuracy is not too large, such as in the application of autonomous driving for judging objects outside the vehicle. At present, more high-end hardware has to be used to improve the detection speed. If it the lightweight, then will not be very high requirements for hardware.

Table 1. Performance comparison of representative algorithms.

Algorithms	Basic Networks	Speed/ (frame-s)	VOC2007 (mAP@IoU=0.5)	VOC2012 (mAP@IoU=0.5)	COCO (mAP@ IoU =0.5:0.95)
SPP-Net	-	0.43	60.9%	59.1%	-
R-CNN	-	0.02	58.5%	53.3%	-
Fast R-CNN	-	3	70.0%	68.4%	19.7%
Faster R-CNN	VGG16	5	78.8%	75.9%	21.9%
M2Det	vGG16	12	-	-	44.2%
D2Det	ResNet-101	6	-	-	45.4%
SSD512	VGG16	22	76.8%	75.9%	26.8%
YOLO v1	—	45	63.4%	57.9%	-
YoLov2	Darknet-19	40	78.6%	73.4%	21.6%
YoLov3	Darknct-53	20	-	-	33.0%
YoLO v4	CSPDarknet-53	33	-	-	43.5%
CornerNet	Hourglass-104	4	-	-	42.2%
CenterNet	Hourglass-104	3.7	-	-	41.6%
FSAF	ResNeXt-101	2.8	-	-	44.6%
FCOS	ResNeXt-64x4 d-101-FPN	10	-	-	44.7%
FoveaBox	RcsNeXt-101	7	-	-	43.9%

6. Conclusion

In this essay, according to different algorithm frameworks in the realm of target detection, the mainstream target detection algorithms are introduced in detail from the two directions of two-stage detection and single-stage detection, including the design idea, basic process, advantages and disadvantages of representative algorithms. At the same time, we compare and analyze the experimental results of general data sets, and related algorithms on mainstream data sets, and analyze

DOI: 10.54254/2755-2721/5/20230656

and prospect the hot research directions in this realm, such as how to achieve small target detection, weak supervision detection and lightweight detection.

References

[1] L. Kalake, W. Wan and L. Hou, "Analysis Based on Recent Deep Learning Approaches Applied in Real-Time Multi-Object Tracking: A Review," in IEEE Access, vol. 9, pp. 32650-32671, 2021.

[2] B. Liu, W. Zhao and Q. Sun, "Study of object detection based on Faster R-CNN," 2017 Chinese Automation Congress (CAC), Jinan, China, 2017, pp. 6233-6236.

[3] Y. Liu, "An Improved Faster R-CNN for Object Detection," 2018 11th International Symposium on Computational Intelligence and Design (ISCID), Hangzhou, China, 2018, pp. 119-123.

[4] X. Xiao and X. Tian, "Research on Reference Target Detection of Deep Learning Framework Faster-RCNN," 2021 5th Annual International Conference on Data Science and Business Analytics (ICDSBA), Changsha, China, 2021, pp. 41-44.

[5] S. Widiyanto, D. T. Wardani and S. Wisnu Pranata, "Image-Based Tomato Maturity Classification and Detection Using Faster R-CNN Method," 2021 5th International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT), Ankara, Turkey, 2021, pp. 130-134.

[6] L. Li and Y. Liang, "Deep Learning Target Vehicle Detection Method Based on YOLOv3-tiny," 2021 IEEE 4th Advanced Information Management, Communicates, Electronic and Automation Control Conference (IMCEC), Chongqing, China, 2021, pp. 1575-1579.

[7] J. Fan, J. Lee, I. Jung and Y. Lee, "Improvement of Object Detection Based on Faster R-CNN and YOLO," 2021 36th International Technical Conference on Circuits/Systems, Computers and Communications (ITC-CSCC), Jeju, Korea (South), 2021, pp. 1-4.

[8] M. Zhang, T. Wang, W. Zhao, X. Chen and J. Wan, "Research on Target Detection of Excavator in Aerial Photography Environment based on YOLOv4," 2020 International Conference on Robots & Intelligent System (ICRIS), Sanya, China, 2020, pp. 711-714.

[9] L. Xiaomeng, F. Jun and C. Peng, "Vehicle Detection in Traffic Monitoring Scenes Based on Improved YOLOV5s," 2022 International Conference on Computer Engineering and Artificial Intelligence (ICCEAI), Shijiazhuang, China, 2022, pp. 467-471.

[10] S. Bouraya and A. Belangour, "Approaches to Video Real time Multi-Object Tracking and Object Detection: A survey," 2021 12th International Symposium on Image and Signal Processing and Analysis (ISPA), Zagreb, Croatia, 2021, pp. 145-151.