Object Detection with Convolutional Neural Networks

Chapter · October 2020

DOI: 10.1007/978-981-15-7106-0_52

2 authors, including:

Sanskruti Patel
Charotar University of Science and Technology
32 PUBLICATIONS **104** CITATIONS

SEE PROFILE

Object Detection with Convolutional Neural Networks

Sanskruti Patel and Atul Patel

Abstract During the last years, a noticeable growth is observed in the field of computer vision research. In computer vision, object detection is a task of classifying and localizing the objects in order to detect the same. The widely used object detection applications are human–computer interaction, video surveillance, satellite imagery, transport system, and activity recognition. In the wider family of deep learning architectures, convolutional neural network (CNN) made up with set of neural network layers is used for visual imagery. Deep CNN architectures exhibit impressive results for detection of objects in digital image. This paper represents a comprehensive review of the recent development in object detection using convolutional neural networks. It explains the types of object detection models, benchmark datasets available, and research work carried out of applying object detection models for various applications.

Keywords CNN · Single-stage object detection · Two-stage object detection

1 Introduction

During the last years, a noticeable growth is observed in the field of computer vision research. Employing machine learning methods provides robust solution to solve computer vision tasks. In computer vision, object detection deals with detecting instances of objects from a particular class in a digital image or video [1]. It is a task of classifying and localizing the objects in order to detect the same. It determines the location where the object is presented in the image and scales one or more objects [2].

S. Patel (✉) · A. Patel
Faculty of Computer Science and Applications, Charotar University of Science and Technology, Changa, India
e-mail: sanskrutipatel.mca@charusat.ac.in

A. Patel
e-mail: atulpatel.mca@charusat.ac.in

A. Joshi et al. (eds.), *Machine Learning for Predictive Analysis*,
Lecture Notes in Networks and Systems 141,
https://doi.org/10.1007/978-981-15-7106-0_52

Object detection pertains to identify all objects presented to an image irrespective of that location, size, rendering, etc. Further information like class of an object, recognition of an object, and object tracking is gained once the object is detected accurately.

Object detection mainly comprises of two tasks: object localization and classification. Object localization determines the location and scale of one or more than one object instances by drawing a bounding box around it. Classification refers to a process to assign a class label to that object. For detection, object detection systems construct a model from a set of training data and for generalization, it is required to provide huge set of training data [3, 4]. In last decade, artificial intelligence made an impact in every field of human life and deep learning is a field of artificial intelligence that uses artificial neural network for representation learning. In the wider family of deep learning architectures, convolutional neural network (CNN) made up with set of neural network layers is used for image processing and computer vision [5]. It is having an input layer, a set of hidden layers, and an output layer. CNN takes an image as an input, processes it, and classifies it under certain category. The application of CNN for object detection was applied on years where an arbitrary number of hidden layers used for face detection [6]. With the availability of large datasets, increased processing capabilities with availability of graphical processing unit (GPU), deep CNN architectures exhibits impressive results in the field of image classification, recognition, and detection of objects in digital image [7]. The object detection models briefly perform following operations: (a) informative region selection (b) feature extraction (c) classification. The paper represented the application of various deep learning techniques based on convolutional neural network (CNN) for object detection.

2 Object Detection Models

With the increase number of usage of face detection systems, video surveillance, vehicle tracking and autonomous vehicle driving, fast and accurate object detection systems are heavily required. The output of object detection normally has a bounding box around an object with the determined value of confidence. Object detection can be single-class object detection, where there is only one object found in particular image. In multi-class object detection, more than one object pertaining to different classes is to be found [8]. Object detection systems mostly rely on large set of training examples as they construct a model for detection an object class. The available frameworks for object detection can be categorized in two types, i.e., region proposal networks and unified networks [1]. The models based on region proposal networks are called as multi-stage or two-stage models. The unified models are called as single-stage models. The multi-stage models perform the detection task in two stages: (i) The region of interest (ROI) is generated in first stage and (ii) classification is performed on these ROI. Two-stage object detectors are accurate but somewhat slower. Single-stage detectors are faster than two-stage detectors as less computational work is needed

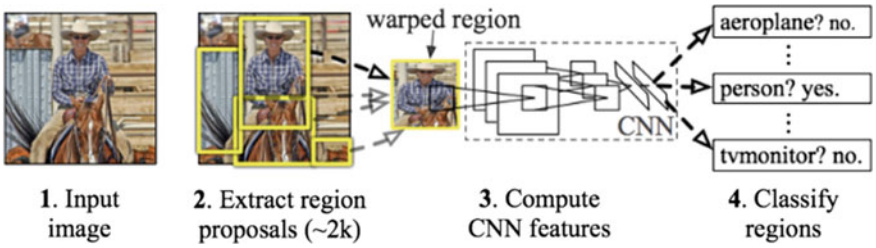

Fig. 1 R-CNN: a region-based CNN detector. *Image source* Girshick et al. [10]

to carry out. On the contrary, two-stage detectors are slow but more accurate. Mask R-CNN, R-CNN, Fast R-CNN, and Faster R-CNN are two-stage object detectors. YOLO and SSD are single-stage object detection models. Two-stage models detect the foreground objects first and then classify it to a specific class, whereas single-stage detector skips the detection of foreground objects and takes uniform samples of objects through a grid [9]. The following section describes the single-stage and two-stage detectors models briefly.

3 Two-Stage Detectors

3.1 Region-Based Convolutional Neural Network (R-CNN)

R-CNN, a short form of region-based convolutional neural network, is one of the most widely used object detection model that falls under two-stage object detectors. Girshick et al. [10] proposed R-CNN that is a first region-based CNN detector as shown in Fig. 1. R-CNN uses a selective search method that generates 2000 regions from the image, called region proposals. These regions are input to CNN that produces a 4096-dimensional feature vector as an output. SVM is applied on this generated vector to classify the objects. Moreover, the bounding box is also drawn surrounding to an object. The major problem with R-CNN is its speed as it is very slow. Also, selective search, the algorithm used to generate the proposals, is very fixed that discards the possibility of learning. The major drawbacks are selective search algorithm proposes 2000 regions per image; for each region of image, it generates CNN feature vector and there is no shared computation between these stages [11]. R-CNN obtained a value of 53.3% of mAP which is significant improvement over the previous work on PASCAL VOC 2012 [1].

3.2 Fast R-CNN

A faster version of R-CNN was released after a year by Girshick [12]. R-CNN uses convolution on each region proposal and it takes more time to complete the detection

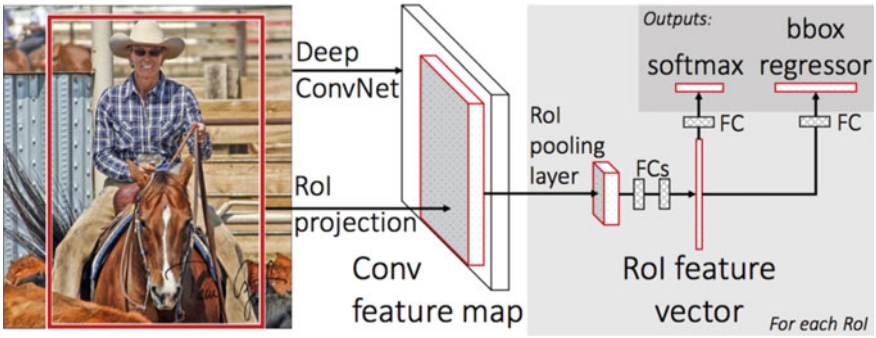

Fig. 2 Fast R-CNN: a faster version of R-CNN. *Image source* Girshick [12]

process. Fast R-CNN uses the concepts of region of interest (ROI) that reduced the time consumption. ROI pooling layer is used to extract a fixed-sized feature map with fixed height and width. For converting features from the regions, it performs the max pooling operation. Before applying a max pooling operation, the h x w RoI window is divided into set of small and fixed sub-windows, i.e., H X W. The size of each generated sub-window in the grid is h/H x w/W. Experimental results showed that Fast R-CNN obtained a mAP score of 66.9% whereas; R-CNN obtained 66.0%. The experiment was conducted on PASCAL VOC 2007 dataset [13]. It uses VGG16 as a pre-trained network model and follows the shared computation that makes the Fast R-CNN fast. It combines three independent processes that shared computational task. R-CNN extracts CNN feature vector from each region proposal, whereas Fast R-CNN group them and only one CNN forward pass is generated. Moreover, the same feature map is used by classifier and regressor, i.e., bounding box shown in the following Fig. 2.

3.3 Faster R-CNN

Faster R-CNN introduced a concept of a region proposal network (RPN) introduced by Ren et al. 2016 [14]. It replaces the slow selective search algorithm used in Fast R-CNN with RPN, a fully CNN that predicts the region proposals. First, a set of anchor boxes, which are in the form of rectangle, are generated around the object. In second step, loss functions are applied to calculate the likelihood of an error. At last, the feature map is generated by backbone network and RPN proposes set of region proposals. These set of proposals are sent as an input to the next layer, i.e., RoI pooling layer. ROI pooling layer converts the features obtained from the fine-tuned CNN layer to a fixed-sized of feature maps. At last, classification layer predicts the class, while bounding box regression creates the rectangular box surrounded to an object for localization. A mAP score of 69.9% is achieved on PASCAL VOC 2007

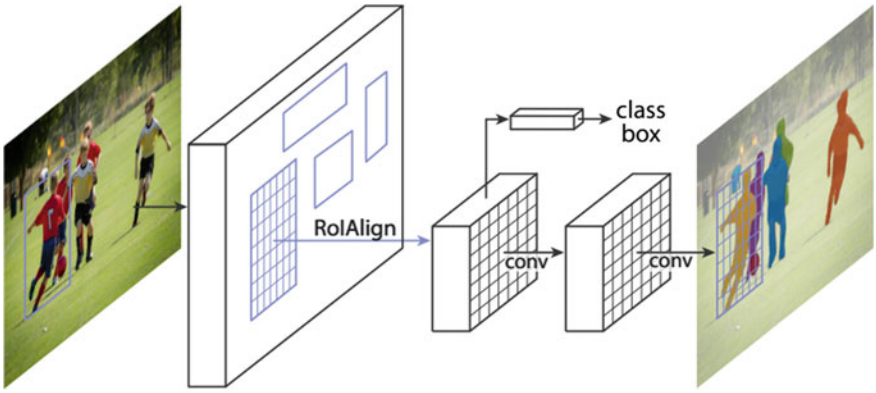

Fig. 3 Instance segmentation using mask R-CNN. *Image source* He et al. [15]

test set, which is a significant improvement over Fast R-CNN for both detection and prediction efficiency [14].

3.4 Mask R-CNN

Faster R-CNN is extended by Mask R-CNN that focuses on instance segmentation from an image and introduced by He et al. [15]. It is an extension of Faster R-CNN and in addition to the class label and bounding box, it also generates the object mask. The accurate detection is essentially required in an instance segmentation task. Therefore, it combines the two important aspects of computer vision task, object detection, which classifies and localizes the objects from an image and semantic segmentation, which classifies and assigns each pixel into a fixed set of category. Mask R-CNN introduced RoIAlign layer that maps the regions more precisely by fixing the location misalignment. The following Fig. 3 is a simple illustration of Mask R-CNN model.

4 Single-Stage Object Detectors

4.1 You Only Look Once (YOLO)

YOLO falls under single-stage object detection models and widely used for real-time object detection task and introduced by Redmon et al. [16]. It generates the bounding boxes and class predictions in single evaluation. It is widely known as unified network and very fast compared to Faster R-CNN and runs using single convolutional neural network. The CNN used in YOLO is based on GoogLeNet model originally and

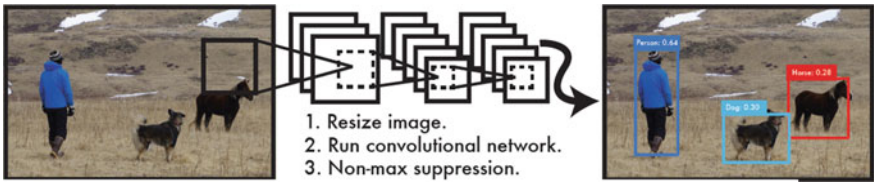

Fig. 4 YOLO: a single-stage object detector model. *Image source* Redmon et al. [16]

the updated version is called DarkNet based on VGG. As per shown in Fig. 4, it splits the input image into a grid of cells, where each cell directly classifies the object and predicts a bounding box. As a result, there are large numbers of bounding boxes generated that are integrated to a final prediction. The variations in YOLO are YOLOv1, YOLOv2, and YOLOv3, where YOLOv3 is the latest version. YOLO is a fast and good for real-time object detection tasks. It is possible to train it end-to-end for accuracy improvisation as it uses a single CNN for prediction. It is more generalized and performs well with generalization of natural and artwork images.

4.2 Single-Shot Detector (SSD)

Like YOLO, SSD also falls under single-stage object detection model and introduced by Liu et al. [17]. It takes only single shot to detect multiple objects within the image. Object localization and classification both performed in a single pass only. For extracting useful image features, SSD uses the VGG-16 model pre-trained on ImageNet dataset as its base model. At the end of the base model, it has additional convolutional layers for object detection. While predicting, score is generated for each object category presented in an image using default box. Also, to improve matching of object shape, it produces adjustments to the box. The network of SSD also pools the predations generated from multiple feature maps with different resolutions. This process helps to handle objects of different sizes.

The following Table 1 presents the features and limitations of the benchmark object detection models including R-CNN family, YOLO and SSD.

5 Benchmark Datasets for Object Detection

As object detection models required huge amount of data to be trained, dataset plays very crucial role in success of these models [2]. A generalized datasets available for object detection tasks are ImageNet [18], MS COCO [19], PASCAL VOC [20], and open images [21]. These dataset are in annotated form and used for benchmarking deep learning algorithms. The following Table 2 summarizes the available images and classes to each dataset and types of images pertain to that dataset.

Table 1 Benchmark single-stage and two-stage object detectors

Model	Features	Limitations
R-CNN	Performs classification and localization for object detection using selective search algorithm	Extracted 2000 region proposals per image, huge amount of time taken to train the network, Real-time implementation not possible as testing is very slow
Fast R-CNN	Region of interest (ROI) pooling layer that proposed fixed-sized regions, combines three models used in R-CNN, incorporated softmax in place of SVM for classification, more accurate and faster than R-CNN	Selective search algorithm for finding proposal makes it slow and time consuming
Faster R-CNN	Introduced region proposal network (RPN) in place of selective search, emerged as a very precise detection model for recognition, much faster and accurate than Fast R-CNN	Finding proposals takes time, many passes are required to complete the entire process
YOLO	Remarkably fast compare to R-CNN family of algorithms, single CNN for localization and classification, used for real-time object detection, better generalization for object representation	Struggles with detection of small objects comprises in a group, incorrect localizations are the main reason of error, problem with generalization of objects with uncommon aspect ratio
SSD	End-to-end training is possible, small CNN filters for prediction of category	Accurate than YOLO but somewhat slower, faster than Faster R-CNN but lesser in accuracy
Mask R-CNN	Used for instance segmentation, introduced RoIAlign layer for accurate segmentation	Classification depends on segmentation

6 Applications of Object Detection Models

Several researchers have applied object detection models so far for different application areas including agriculture, medical imaging, satellite imagery, transport system, etc. The following Table 3 summarizes the work done with different object models and pre-trained model, application area, dataset used, and accuracy achieved. To evaluate the effectiveness of any object detection model, various parameters are taken into consideration including average precision (AP), average recall (AR), mean average precision (mAP), and intersection over union (IoU).

Table 2 Benchmark datasets for object detection

Dataset	Number of images	Number of classes	Types of images
ImageNet—WordNet hierarchy	14,000,000	21,841	Animal, plant, and activity
Microsoft COCO (common objects in context)—Large scale and used for object detection and segmentation	330,000	80	250,000 people with key points, more than 200,000 are labeled, 5 captions per image
PASCAL VOC (Visual object classes)—Standardized image data sets obtained through VOC challenges	11,530	20	Common images of daily life, annotations are in XML file
Open Images—OICOD (open image challenge object detection)—training − 9 million + images, validation −41 k + images, test −125 k + images	9,000,000	6000	Diverse images of complex scenes, all images are annotated

7 Conclusion

With the increase number of usage of face detection systems, video surveillance, vehicle tracking and autonomous vehicle driving, fast and accurate object detection systems are heavily required. Object detection refers to locating and classifying object from digital image. With the progressive result from deep CNN architectures, CNN-based object detectors are used in variety of applications. Based on the methodology, it has been categorized as either single-stage or two-stage object detection model. This paper summarizes the different CNN-based models that include R-CNN, Fast R-CNN, Faster R-CNN, Mask R-CNN, SSD, and YOLO. Apart from this, it explains the different features of available datasets. It also covers the details of research work carried out so far that applied object detection models in various fields of applications.

Table 3 Applications of object detection models

Description of problem	Dataset	Object detection method	Object(s) identified	Accuracy obtained
The soldered dots identification of automobile door panels [22]	800 images collected using camera	Fast R-CNN and YOLOv3	Rectangle dot, semicircle dot, circle dot	Fast R-CNN: mAP-0.8270 Recall-0.8993 YOLOv3: mAP-0.7992 Recall-0.9900
Different scales face detection [23]	FDDB, AFW, PASCAL faces, Wider face	Improved Faster R-CNN	Human faces from different scales	Recall-96.69%
Using semantic segmentation pedestrian detection [24]	Caltech pedestrian dataset with 10 h of video consists 2300 pedestrian	Faster R-CNN + VGG16 model	Pedestrian segmentation and detection	IoU-0.75
Segmentation and detection of oral disease [25]	MSCOCO dataset for pre-training and 30 training and 10 validation	Modified Mask R-CNN	Cold sores and canker sores	AP- 0.744
Gastric cancer diagnosis [26]	1400 images with 1120 positive sample and 280 negative sample	Mask R-CNN	Prediction with masking and bounding box	AP-61.2
Ear detection [27]	AWE dataset	Mask R-CNN	Colored region on identified ear	Accuracy-99.7% IoU-79.24% Precision-92.04%
Abandoned baggage detection [28]	User created dataset	YOLO	Box around the person and abandoned bag	11 cases are correctly identified
Fast vehicle detection [29]	KITTI (15,000 images) + LSVH (16 videos)	Improved Fast R-CNN	Bounding box around the vehicles	AP Easy-89.20% Moderate-87.86% Hard-74.72%
Ship detection and segmentation [30]	42,500 images from the Airbus ship dataset	Mask R-CNN	Circle box and masking around the ship	mAP-76.1%
Bones detection in the pelvic area [31]	320 monochromatic images created out of two CT datasets	YOLO	Bones detection with labels	Precision-99% Recall-99%

References

1. Z, Zhao, P. Zheng, S. Xu, X. Wu, Object detection with deep learning: a review. IEEE Trans. Neural Netw. Learn. Syst. **30**(11), 3212–3232 (2019)
2. L. Liu, W. Ouyang, X. Wang et al., Deep learning for generic object detection: a survey. Int. J. Comput. Vis. **128**, 261–318 (2020)

3. A. Opelt, A. Pinz, M. Fussenegger, P. Auer, Generic object recognition with boosting. IEEE TPAMI **28**(3), 416–431 (2006)
4. A. Voulodimos, N. Doulamis, A. Doulamis, E. Protopapadakis, Deep learning for computer vision: a brief review. Comput. Intell. Neurosci. 1–13 (2018)
5. Y. LeCun, Y. Bengio, G. Hinton, Deep learning. Nature **521**, 436–444 (2015)
6. H.A. Rowley, S. Baluja, T. Kanade, Neural network-based face detection. PAMI (1998)
7. A. Krizhevsky, I. Sutskever, G.E. Hinton, ImageNet classification with deep convolutional neural networks. Commun. ACM **60**(6), 84–90 (2017)
8. A.R. Pathak, M. Pandey, S. Rautaray, Application of deep learning for object detection. Procedia Comput. Sci. **132**, 1706–1717 (2018)
9. C. Li, Transfer learning with Mask R-CNN, https://medium.com/@c_61011/transfer-learning-with-mask-r-cnn-f50cbbea3d29
10. R, Girshick, J, Donahue, T, Darrell, J, Malik, Rich feature hierarchies for accurate object detection and semantic segmentation, in *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 580–587 (2014)
11. L. Weng, Object detection for dummies part 3: R-CNN family. https://lilianweng.github.io/lil-log/2017/12/31/object-recognition-for-dummies-part-3.html
12. R. Girshick, Fast R-CNN, in *ICCV*, pp. 1440–1448 (2015)
13. M, Everingham, L. Van Gool, C.K.I. Williams, J. Winn, A. Zisserman, The pascal visual object classes (voc) challenge. Int. J. Comput. Vis. **88**, 303–338 (2010)
14. S. Ren, K. He, R. Girshick, J. Sun, Faster RCNN: towards real time object detection with region proposal networks. IEEE TPAMI **39**(6), 1137–1149 (2017)
15. K. He, G. Gkioxari, P. Dollár, R. Girshick, Mask RCNN, in *ICCV* (2017)
16. J. Redmon, S. Divvala, R. Girshick, A. Farhadi, You only look once: unified, real time object detection, in *CVPR*, pp. 779–788 (2016)
17. W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C. Fu, A. Berg, SSD: single shot multibox detector, in *ECCV*, pp. 21–37 (2016)
18. J. Deng, W. Dong, R. Socher, L. Li, K. Li, F. Li, ImageNet: a large scale hierarchical image database, in *CVPR*, pp. 248–255 (2009)
19. T. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, L. Zitnick, Microsoft COCO: common objects in context. in *ECCV*, pp. 740–755 (2014)
20. M. Everingham, S. Eslami, L.V. Gool, C. Williams, J. Winn, A. Zisserman, The pascal visual object classes challenge: a retrospective. IJCV **111**(1), 98–136 (2015)
21. A. Kuznetsova, H. Rom, N. Alldrin, J. Uijlings, I. Krasin, J. Pont-Tuset et al., The open images dataset v4: unified image classification, object detection, and visual relationship detection at scale. arXiv:1811.00982. (2018)
22. W. You, L. Chen, Z. Mo, Soldered dots detection of automobile door panels based on faster R-CNN model, in *Chinese Control And Decision Conference (CCDC)* (Nanchang, China, 2019), pp. 5314–5318
23. W. Wu, Y. Yin, X. Wang, D. Xu, Face detection with different scales based on faster R-CNN. IEEE Trans. Cybern. **49**(11), 4017–4028 (2019)
24. T. Liu, T. Stathaki, Faster R-CNN for robust pedestrian detection using semantic segmentation network. Front. Neurorobot. (2018)
25. R. Anantharaman, M. Velazquez, Y. Lee, Utilizing mask R-CNN for detection and segmentation of oral diseases, in *IEEE International Conference on Bioinformatics and Biomedicine (BIBM)* (Madrid, Spain, 2018), pp. 2197–2204
26. G. Cao, W. Song, Z. Zhao, Gastric cancer diagnosis with mask R-CNN, in *11th International Conference on Intelligent Human-Machine Systems and Cybernetics (IHMSC)* (Hangzhou, China, 2019), pp. 60–63
27. M. Bizjak, P. Peer, Ž. Emeršič, Mask R-CNN for ear detection, in *42nd International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)* (Opatija, Croatia, 2019), pp. 1624–1628
28. T. Santad, P. Silapasupphakornwong, W. Choensawat, K. Sookhanaphibarn, Application of YOLO deep learning model for real time abandoned baggage detection, in *IEEE 7th Global Conference on Consumer Electronics (GCCE)* (Nara, 2018), pp. 157–158

29. H. Nguyen, Improving faster R-CNN framework for fast vehicle detection. Math. Prob. Eng. 1–11 (2019)
30. N. Xuan, D. Mengyang, D. Haoxuan, H. Bingliang, W. Edward, Attention mask R-CNN for ship detection and segmentation from remote sensing images. IEEE Access 1–1 (2020)
31. Z. Krawczyk, J. Starzyński, Bones detection in the pelvic area on the basis of YOLO neural network, in *19th International Conference* (2020)