Object Detection with Deep Learning

Yash Bhadiyadra

Student, Department of Information Technology, School of Engineering, P. P. Savani University, India

Abstract: ***Object detection has been a significant research hotspot and an extensively utilized problem in computer vision during the past 20 years. In a given image, it seeks to rapidly and precisely detect and locate a large number of items according to predetermined categories. The algorithms may be split into two categories based on the model training method: single - stage detection algorithms and two - stage detection algorithms. The typical algorithms for each level are thoroughly introduced in this work. Then, numerous typical techniques are examined and contrasted in this area while public and special datasets that are frequently utilized in target detection are introduced. The probable difficulties in target detection are therefore anticipated.***

Keywords: Deep Learning, Object Detection, Neural Network

1. Introduction

In the disciplines of computer vision, deep learning, artificial intelligence, etc., object detection is a fundamental study area. For more difficult computer vision tasks like target tracking, event detection, behaviour analysis, and scene semantic comprehension, it serves as a crucial precursor. It seeks to properly identify the category, locate the target of interest inside the image, and provide the bounding box of each target. Vehicle automated driving, video and image retrieval, intelligent video surveillance, medical image analysis, industrial inspection, and other sectors have all made extensive use of it.

Based on several methods and methodologies, this study proposes a unique framework for item recognition in crowded settings. It is a fundamental area of research in artificial intelligence, deep learning, computer vision, etc. For more challenging tasks like behaviour analysis, scene semantic interpretation, and object and event identification, this is essential. The identification of real - world objects, such as faces and people, presents severe challenges since modelling them is problematic due to the variation of colour and texture, as well as the unrestricted nature of the background against which the items lie. The goal of this is to provide computational methods and approaches that offer the basic data that computer vision requires. New techniques for object detection have been introduced in recent years as a result of the rapid advancement in deep learning, which has produced amazing advancements.

The goal of object detection is to identify the item of interest inside the image, as well as its category and bounding box. It has been employed in a variety of industries, including industrial inspection, autonomous vehicles, and medical image analysis, and video and image retrieval applications.

Region selection with information: It is a logical decision to scan the entire image with a multi - scale sliding window since distinct items may appear in any places of the image and have varying aspect ratios or sizes. Although this thorough approach can determine every conceivable position for the items, it also has clear flaws. It is computationally intensive and creates an excessive number of redundant windows dueto the enormous number of candidate windows. However, undesirable areas could be created if a set number of sliding window templates are used.

Extracting features: We need to extract visual characteristics that can offer a stable and meaningful representation in order to distinguish between distinct things. The representative ones include SIFT, HOG, and Haar - like characteristics. This is because these characteristics are capable of generating representations linked to complex brain cells. However, it's challenging to manually build a comprehensive feature descriptor to accurately characterize all types of objects because to the range of looks, lighting circumstances, and backdrops.

Classification: A classifier is also required to separate a target item from all other categories and to improve the hierarchical, semantic, and informative qualities of the representations for visual recognition. Typically, the Deformable Part - based Model (DPM), AdaBoost, and Supported Vector Machine (SVM) are suitable options. The DPM is one of these classifiers that can manage extreme deformations by combining object components with deformation cost. In DPM, component decompositions and carefully crafted low - level characteristics are merged with the help of a graphical model. Additionally, generating high - precision part - based models for a range of object classes is possible using discriminative learning of graphical models.

2. Two - stage Object Detection

2.1 R - CNN

Girshick introduced the R - CNN model, the first actual object detection model based on convolutional neural networks, to avoid the issue of choosing a large number of regions. Here, 2000 areas are extracted from the picture using selective search; Girshick dubbed these regions suggestions. Consequently, we can deal with only 2000 areas rather than trying to categories all of them.

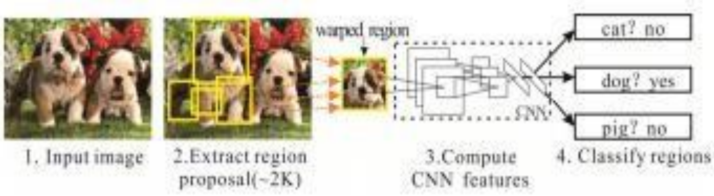


Figure 1: R - CNN Architecture

All of these 2000 area suggestions are then twisted into a square and fed into a convolutional neural network, which generates an output that is a 4096dimensional feature vector. In this case, the CNN performs the function of a feature extractor, and the output layer is packed with features that have been taken from the picture and then fed into the Support Vector Machine (SVM). The SVM assigns the object's presence in the area proposals a classification. The bounding box precision is increased by this algorithm's prediction of four offset values.

The issue with R - CNN is that training the network requires a significant amount of time (about 47 seconds for each test picture), since the categorization requires 2000 region recommendations per image. Because the R - CNN method, or the selective search algorithm, is a fixed algorithm, no learning is taking place at this point, which might result in the development of poor candidate region recommendations.

2.2 SPP - NET

He proposed the Spatial Pyramid Pooling (SPP) paradigm. This approach enhances the bounding box prediction performance when compared to R - CNN and fixes the issue with fixed input size in R - CNN. This rendered the convolutional neural network model independent of the size of the input picture.

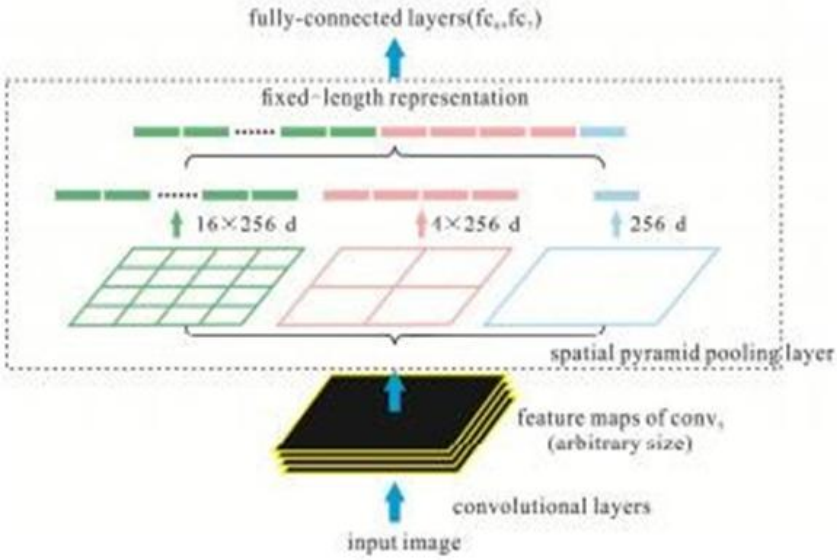


Figure 2: SPP - Net Architecture

This approach extracts the features of the region suggested on the feature map after the original picture is fed through the convolutional neural network. The final convolutional layer is simultaneously given the SPP layer, and the region proposal's feature extract is processed via the SPP layer to extract the fixed - size feature vector.

2.3 FAST R - CNN

Girshick also suggested this idea. The R - CNN model and this one is comparable. We send the picture to CNN directly rather than sending it region ideas and CNN creates a convolutional feature map as a result. The ROI pooling layer is used to restructure the squares that were warped into the convolutional feature map's identified regions into fixed - size regions (it is a max pooling, where the pool size depends on input size). We employ a SoftMax layer from this feature vector to forecast the class of the suggested region as well as the offset values for the bounding box.

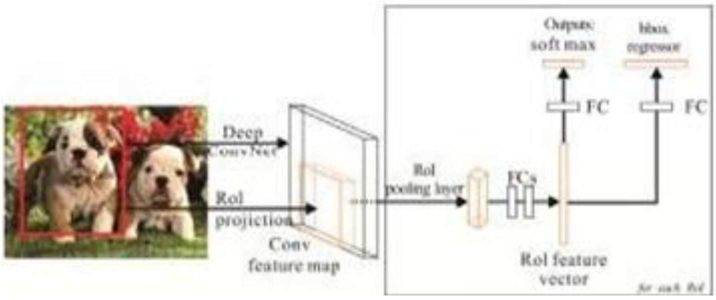


Figure 3: Fast R - CNN Architecture

After doing all these things, it still cannot meet the needs of real - time detection. This is because Fast RCNN also uses selective search to find RoI (Region of Interest) which takes 2 seconds to detect objects in every image (it's still better than RCNN but considering large datasets, then Fast R - CNN cannot be considered that fast).

2.4 FASTER R - CNN

Selective search strategies are used by R - CNN and Fast R - CNN to locate the region suggestions. because a lengthy process, targeted search Ren thus developed the Faster R - CNN model, which does away with the selective search technique and detects objects in images in 0.2 seconds.

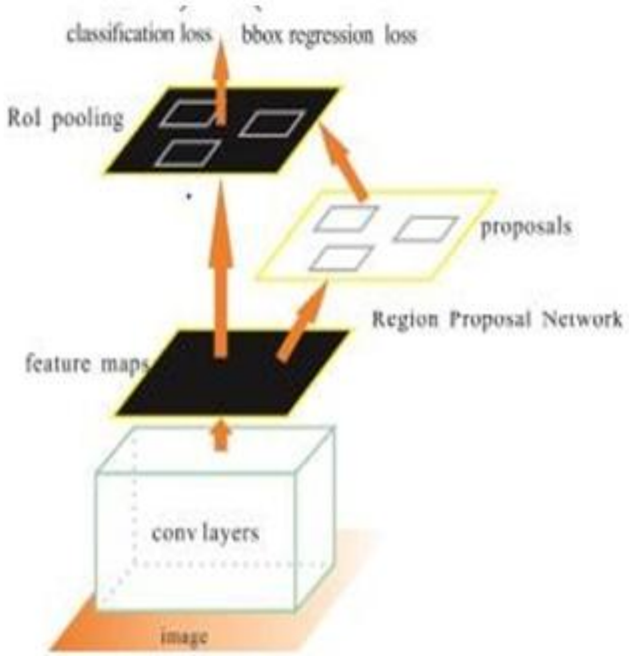


Figure 4: Faster R - CNN Architecture

Utilizing Region Proposal Networks, the Faster R - CNN (RPN). This method gives the feature map for the picture after receiving the image as input. Following that, RPN applies a sliding window to these feature maps and creates k Anchor boxes - fixed - sized boundary boxes scattered throughout the image - in a variety of shapes and sizes at each window. and this function delivers the score for objectness together with the object recommendations. These suggestions are given a RoI pooling layer to make them all the same size. The FC (completely connected) layer, which has a SoftMax and linear regression layer and identifies bounding boxes for objects, receives these ideas last.

3. One- Stage Object Detection Algorithm

3.1 YOLOv1

Joseph Redmon proposed the YOLOv1 object detection model in 2016. The extraction procedure of region proposals is not necessary for the YOLOv1 detection model. Simply said, the whole detection model is a CNN network structure. The fundamental concept is to immediately return the position and category of the bounding box at the output layer by feeding the network the whole graph as input. An S*S grid is first used to break up the picture, and each grid cell predicts a B bounding box and the confidence scores for these boxes. In other words, each cell forecasts a total of B* (4+1) values. Its real - time detection rate on a single TitanX may be as fast as 45 frames per second.

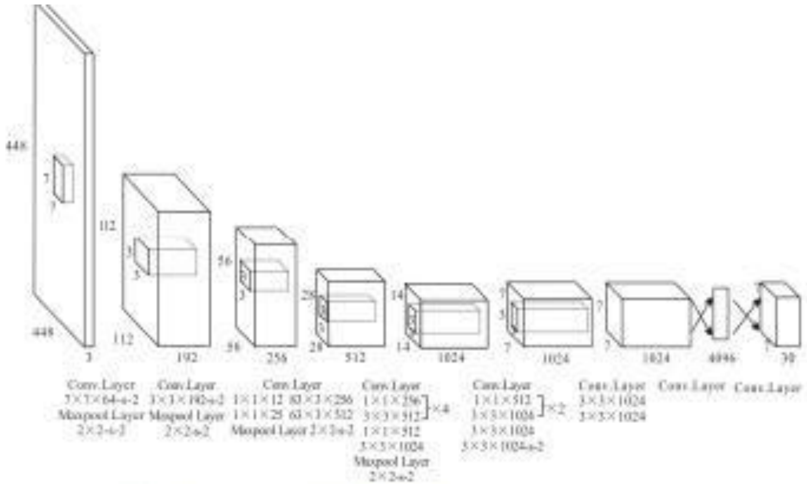


Figure 5: YOLOv1 Architecture

3.2 YOLOv2

Redmon put out the YOLOv2 concept in 2016. Enhancing recollection and localisation while keeping classification accuracy is the major objective. Darknet19, a novel fully convolutional feature extraction network with a total of 19 convolutional layers and 5 maximum pooling layers, is the network used by YOLOv2. The recall and accuracy are considerably increased by adding a batch normalisation layer to the convolutional layer, eliminating dropout, implementing an anchor box method, applying k - means clustering on the training set bounding box, and multi - scale training. However, there is still room for improvement in the identification of targets with significant overlap and tiny targets.

3.3 YOLOv3

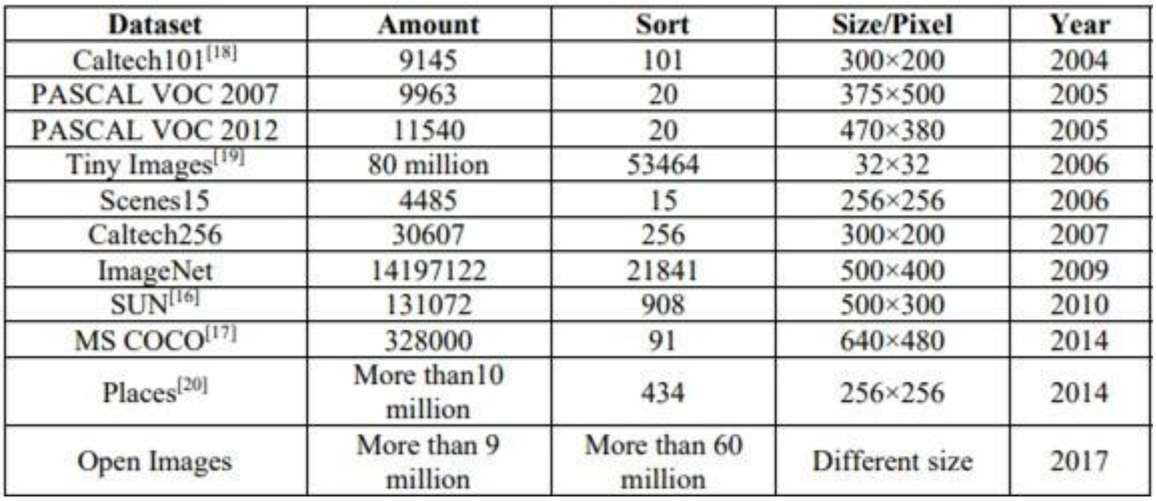
By far, Redmon's YOLOv3 model for object detection is the most balanced in terms of both detection speed and accuracy. In terms of categorization prediction, YOLOv3 primarily consists of converting the original single - label classification into a multi - label classification and swapping the original SoftMax layer for single - label multi - classification with a logistic regression layer. The model simultaneously combines a number of scales to make predictions. It uses an up - sampling fusion technique similar to FPN and blends three scales at the end, greatly enhancing the identification of tiny objects. This model's network structure uses the Darknet - 53 feature extraction network for deeper feature extraction. The detection impact of tiny objects has also been greatly enhanced, even though the YOLOv3 model further enhances detection speed.

4. Datasets and Performance Comparison of Various Algorithms

4.1 Dataset

The idea of "artificial intelligence" was put up as early as 1956. However, it wasn't until 2012 that the rise of artificial intelligence really took off. The development of machine learning techniques, increased processing power, and growing data volumes are the key causes of this. The growth of data volume and the development of detecting technologies go hand in hand. This is due to the fact that datasets are required for performance testing and algorithm assessment, and datasets are also a strong motivator for the advancement of the detection techniques study area. Table 1 displays the characteristics of popular public data sets.

Table 1: Public Dataset and Its Parameters

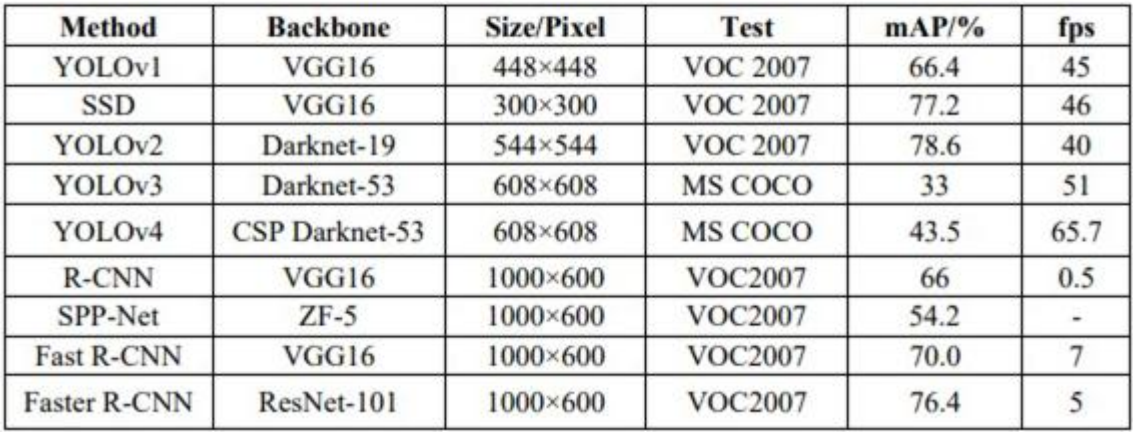


Dataset	Amount	Sort	Size/Pixel	Year
Caltech101[18]	9145	101	300×200	2004
PASCAL VOC 2007	9963	20	375×500	2005
PASCAL VOC 2012	11540	20	470×380	2005
Tiny Images[19]	80 million	53464	32×32	2006
Scenes15	4485	15	256×256	2006
Caltech256	30607	256	300×200	2007
ImageNet	14197122	21841	500×400	2009
SUN[16]	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

Statistics and comparisons of single - stage and two - stage detection techniques are provided in Table 2.

Table 2: 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

Object identification is one of the most fundamental and difficult issues in computer vision, and it has attracted a lot of attention lately. Although deep learning - based detection techniques have been widely used in various domains, there are several issues that need to be investigated.

1) Lessen reliance on data.
2) To obtain effective tiny object detection.
3) Multi - category object detection is implemented.

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: 211252.

[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] P. F. Felzenszwalb, R. B. Girshick, D. Mcallester, and D. Ramanan, “Object detection with discriminatively trained part - based models, ” IEEE Trans. Pattern Anal. Mach. Intell., vol.32, no.9, p.1627, 2010.

[12] K. K. Sung and T. Poggio, “Example - based learning for view - based human face detection, ” IEEE Trans. Pattern Anal. Mach. Intell., vol.20, no.1, pp.39–51, 2002.

[13] C. Wojek, P. Dollar, B. Schiele, and P. Perona, “Pedestrian detection: An evaluation of the state of the art, ” IEEE Trans. Pattern Anal. Mach. Intell., vol.34, no.4, p.743, 2012.

[14] H. Kobatake and Y. Yoshinaga, “Detection of spicules on mammogram based on skeleton analysis. ” IEEE Trans. Med. Imag., vol.15, no.3, pp.235–245, 1996.

[15] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell, “Caffe: Convolutional architecture for fast feature embedding, ” in ACM MM, 2014.

[16] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “Imagenet classification with deep convolutional neural networks, ” in NIPS, 2012.

[17] Z. Cao, T. Simon, S. - E. Wei, and Y. Sheikh, “Realtime multi - person 2d pose estimation using part affinity fields, ” in CVPR, 2017.