An Overview of Deep Learning Based Object Detection Techniques

Bhagya C
Department of Computer Science and Engineering
TKM College of Engineering
Kollam, India
bhagyacbose@gmail.com

Prof. Shyna A
Department of Computer Science and Engineering
TKM College of Engineering
Kollam, India
shyna@tkmce.ac.in

***Abstract*—Recent years have witnessed a boundless growth in the field of deep learning. With the preferment in the field of deep learning, the task of object detection has become more exciting and challenging. Object detection focuses on detecting the presence of entire objects within a given image. Deep learning based object detection techniques have shown an efficacy to learn the object features directly from the data. The paper mainly focuses on providing a survey on various state-of-the-art deep learning based object detection techniques. The work also concentrates on providing an extensive comparison regarding the opportunities and obstacles faced by different object detection techniques. The paper concludes by identifying the future golden scopes for research in these fields.**

Keywords—Deep learning, Object detection, Computer vision.

I. INTRODUCTION

Computer vision acts as a tool to perceive and grasp knowledge. The three basic works undertaken by a computer vision algorithm include classification, detection and localization. These three fundamental tasks provide both opportunities and challenges to the field of computer vision. They have started seeking the attention of researches and there came a tremendous increase in the work over these fields. One reason for this tremendous growth is the integration of deep convolutional neural network to the field of computer vision. Image classification intends to convert an image to a label. The process of object detection, another important task in computer vision, points to detect the presence of entire objects within an image. For example, driverless cars otherwise called as self-driving cars are a coupe that use a combination of software and sensors, not only to detect the presence of other cars but also the presence of trees, human beings, animals, other vehicles etc. on the road.

The third main task in computer vision is called object localization. The localization task is similar to the task of detection. Object detection deals with detection of a variable number of object classes whereas localization deals with the identification of a fixed number of object classes. The localization task is considered to be a superset of the computer vision task of image classification. With this, we are not only required to know whether there is an object in the image of the given class but also to know where exactly the object is in the image. Object detection is regarded as the first process in most computer vision systems. It has several applications and some of them include security systems, human-computer interactions, robotics, product detection in manufacturing industries, people counting and so on.

From the last decade, the field of machine learning has started to move forward in a purposeful way, as a result, the researches have started to concentrate more over these techniques especially for the matter of object detection. Machine learning-based object detection consists of two phases, a training phase as well as a testing phase. An extensively large number of images are used for the purpose of training. Increasing the number of images enables the computer to learn the class of objects correctly and helps to properly identify the objects belonging to different classes. Testing phase involves the task of testing whether the machine responds correctly by giving different inputs or test cases. By giving an image to a computer it tries to learn every bit of object features within the image by a process called feature extraction.

People learn and respond in different ways. They can learn through experience, but in the case of machines, we need to train them properly. An efficient training on machines enables them to respond quickly and correctly over different environments. Thus, these tasks are done by training a classifier that is capable of extracting even the minute differences in how the object looks. A set of regions called candidates or proposals are provided as the input to the classifier. These region proposals are found to be an important stage and inaccurate proposals can harm the performance of the system on work. Thus came a subset of machine learning called deep learning, which concentrates on these issues in a more efficient manner.

Deep learning is simply a kind of algorithm that founds to work well for the task of prediction. With the immense performance of deep learning based object detection techniques in the past few years, we have decided to provide a survey on various state-of-the-art deep learning based object detection techniques. The paper focuses on providing an overview of some real-world high valued object detection applications. It also gives a tabular comparison of various deep learning based object detection techniques and the scope of object detection in remote sensing images.

II. TECHNIQUES EMPLOYED

One of the most popular and simple object detection techniques is called a sliding window [1-4]. As the name suggests sliding window is a small window that moves across the full image starting from the top left corner of the image. This is a simple technique that focuses on making a window of size usually much smaller than the actual image size. After examining the full image content, the size of the window is incremented by some steady value. This process is repeatedly done with the new windows until the stopping condition is reached. In sliding window technique the amount of regions that are taken into consideration is extremely high. Sliding window technique is computationally expensive and also give inaccurate bounding boxes.

In order to beat the constraints of sliding window technique, R Girshick [5] proposed a method for object detection using R-CNN (Regional Convolutional Neural Networks). It focuses on choosing those regions that make some sense to run on CNN. The approach used here for region proposal is selective search algorithm [6] for the purpose of partitioning the whole image into groups based on similarity. Selective search algorithm works better as compared with traditional methods such as the sliding window. Even though RCNN gives less number of candidate region proposals, the drawback is that it requires an expensive multistage training and the technique mostly gives slow detection. To solve these issues P Sermanet et al. [7] introduced an interesting point of view for object recognition, localization and detection, and this strategy is called as OverFeat.

The ILSVRC2013 localization competition announced this method as one of the thriving object detection strategy. The main idea of overfeat is to train a convolutional neural network to simultaneously perform the three basic tasks of computer vision. It is treated as a unique method for localization and detection tasks by gathering the predicted bounding boxes. Bounding boxes are simple rectangular boxes used to mark the identified objects in the image. Most of these approaches uses the concept of sliding window techniques and hence need to compute for each window of the input image. This inefficiency can be solved by using ConvNets because they share the computations which are common to overlap regions. This leads to the arrival of a new technique for object detection called Spatial Pyramid Pooling or SPP.

Spatial Pyramid Pooling suggested by Kaiming He et al. [8] is a different pooling strategy and the entire network structure is called SPP-net. Feature extraction process mainly aims to extract a large collection of features from each of the region proposals of the image. This came to be another major reason for introducing spatial pyramid pooling into the convolutional neural network architecture. The ability to manage images of different scales, sizes and aspect ratios made SPP an adaptive technique. The initial step is to generate feature maps of the input images using a number of convolutional layers and the feature maps are generated only once in the SPP-net. These feature maps are allowed to pass through the SPP layer which outputs n number of M-dimensional vectors. Here n is the number of filters available in the final convolutional layer. Fig. 1 describes the spatial pyramid pooling methodology. SPP is found to be faster than RCNN but it gives reduced accuracy for very deep neural networks. There came the need to have a generalizable object detection technique.

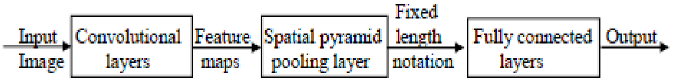

Fig. 1. SPP-net Architecture.

Gidaris S and Komodakis N [9] thus proposed a generalizable method for object detection called MultiRegion CNN or simply MRCNN. The central idea is to extract as many features as possible from all around and within the region proposals. All of these features are then simply merged together. The various regions under consideration for an object proposal include half regions, central regions, border regions, contextual regions and so on. It also integrates the semantic segmentation aware features for the work of object detection and is not found suitable for all kind of real-time applications. Hence Donggeun Yoo et al. [10] has introduced an iterative procedure for object detection called AttentionNet. The policy is to narrow down the bounding box starting from the image boundary to the exact object location. But experiments revealed the fact that AttentionNet is suited to scale for multiple classes and also results in a reduced recall.

Girshick [11] has proposed a new method, called fast R-CNN, to overcome the limitations of R-CNN and SPP-Net. This turned out as one of the main aim of fast R-CNN. Feature vectors of fixed-length is learned from the feature map for the entire candidate region proposals. This is done with the help of a pooling layer called the region of interest (RoI) pooling layer, which came to be an attractive advantage for fast R-CNN. Initially, it allows the entire image to pass through the ConvNet to create the regions of interest. It adopts a single stage strategy which extracts features from the proposed regions, classifies them and finally returns the bounding boxes together with the output class label. It gives high-quality object detection than SPP-Net and R-CNN.

One of the shortcomings of fast R-CNN is its high computation time due to slow region proposals by the selective search algorithm. Inorder to overthrow this situation, Ren S et al. [12] has suggested a new methodology called faster R-CNN. The framework offers a more accurate and efficient way of generating region proposals via region proposal networks. The input image is made to pass through the ConvNet to obtain feature maps and is then applied to the region proposal network to generate the object proposals. For classification, these candidate regions are moved to the fully connected layer. The highlight of faster R-CNN is its speed but faces many challenges such as inaccurate bounding boxes.

In order to overcome the computational expense and to provide more accurate bounding boxes, Redmon et al. [13] has introduced the YOLO (You Only Look Once) method. The main idea is to divide the image into multiple number of grid cells. The classification and localization algorithm is then applied over each of these grid cells. The class label or tag corresponding to each object present in a grid is resolved by the centre of the object. For each of the class instance, YOLO predicts a fixed number of bounding boxes and scores or class probabilities for multiple objects. It is extremely fast and provides less amount of background errors. But it can't detect multiple objects within a single grid, also can detect one object multiple times. There came a loss in accuracy rate with YOLO.

Thus there arise a need to detect objects of different scales, and hence Anguelov D et al. [14] has proposed the idea of single shot multibox detector, in short, called SSD. It passes boxes of different scales to the different layers of CNN and allows each layer to predict objects based on the scale value. Scale value is nothing but a furnishing parameter of images. SSD works well for larger objects but does not generate enough amount of higher-level features for smaller objects. A good new generalizable strategy which can provide rich semantics in all levels is needed to improve the accuracy. Thus Wanli Ouyang et al. [15] introduced DeepIDNet to properly learn deformation features of objects with varying size and shape. Verification issue is a challenge to this method. To overcome the problems of SSD and DeepIDNet and to improve the accuracy, Dai et al. [16] have proposed Region based Fully Convolutional Network or in short RFCN. One significant advantage of RFCN over methods such as RCNN, fast RCNN etc. is that it share the computation results thereby reducing the complexity of the entire system. It deals with an easier training strategy with reduced complexity and furnishes with acceptable accuracy.

Lin T et al. [17] puts forward a pyramid concept to generate better quality features for object detection known as feature pyramid network or FPN. FPN is actually a feature extractor which can be used to alter the extractors of other detectors. It is a layered architecture where the semantic values increase as each layers proceeds. Semantically rich layers can be used to generate high resolution layers. The performance of FPN is highly influenced by the semantic layers and one challenge of FPN is its connections. For example removing the top-down connection can reduce the accuracy of the entire network. Several works are going on to solve these challenges of FPN to use for many socially relevant applications.

Later, a scene adaptive object detection model was designed by TychsenSmith L and Petersson L [18] called DeNet. The basic idea of this approach is to predict the object candidate proposals by providing a provision for the bounding box corner estimation. The main advantage is that it doesn't require any predefined anchors and is much faster than many other object detection techniques. One problem associated with this approach is that it requires more time for generating corners and for evaluating the base network. By solving this problem, the DeNet can achieve a steep improvement in performance.

III. APPLICATIONS OF OBJECT DETECTION

Object detection is a current trending area having a wide range of applications over different fields. One real-time application is the case of self-driving cars which points to detect the presence of trees, other vehicles, humans, animals, etc. on the road. A popular trend is the use of object detection techniques over remote sensing images to detect the presence of desert oasis, forest fires, crashed flights etc. and researchers have started working on it. It also has significance in the field of medical imaging for tumor detection, computer tomography (CT) and so on. Some common applications include the following:

A. Face detection.

One of the most popular applications of object detection is face detection which includes detecting the presence of faces from the given images. Convolutional neural networks play a crucial role in extracting the facial features, spotting them and reporting the final output. Social media makes use of this application to detect the face image when a photo is uploaded and is of prime importance in the current world.

B. Vehicle detection.

Another real-time application is to detect the presence of vehicles such are cars, scooters, bus etc. which act as the objects in this scenario. In order to track the vehicles moving in a road, the speed act as a crucial factor to keep track of the objects and is proved to be successful.

C. People counting.

This application aims to count the number of people in a given scene. Counting the number of people in a video is difficult but of high importance. It is mainly used to analyze the crowd during an event. These are currently in active use in many countries.

D. Security and surveillance.

This application is of high value in current world due to the rise in anti-social activities. Remote sensing image based detection of intruders, explosives etc. fall under this category. Anomaly detection is another application in which firms spend lots of money, also comes under this category of object detection application. A lot of research works are carried out to automate these systems for improved performance.

IV. COMPARATIVE STUDY

Table 1 provides a brief idea about the various techniques used for object detection, its pros and cons.

TABLE I
BRIEF OVERVIEW OF VARIOUS OBJECT DETECTION TECHNIQUES

No.	Technique	Authors	Year	Advantages	Limitations
1.	Sliding Window[1]	Viola P, Jones M	2001	Simple Easy to implement	Time consuming
2.	R-CNN[5]	Girshick R, Donahue J, Darrell T, Malik J	2014	Number of regions proposed is less as compared with sliding window technique	Multi-stage training Expensive training in terms of space and time Slow object detection
3.	OverFeat[7]	Sermanet P, Eigen D, Zhang X, Mathieu M, Fergus R, LeCun Y	2014	High speed than RCNN	Less accurate
4.	SPP-net[8]	He K, Ren S, Zhang X, Sun J	2014	Faster than R-CNN Avoid repeated computation of features.	Reduced accuracy for very deep neural network
5.	MRCNN[9]	Gidaris S, Komodakis N	2015	Easy to train Generalize well Small overhead	Not suited for all kind of real time applications
6.	AttentionNet[10]	Donggeun Yoo, Sunggyun Park, Joon-Young Lee, Anthony S. Paek, In So Kweon	2015	More accurate detection	Unable to scale to multiple classes Low recall
7.	Fast R-CNN[11]	Girshick R	2015	High quality detection than SPP-net and R-CNN Single stage training	Slow clustering Selective search is slow so still high computation time
8.	Faster RCNN[12]	Ren S, He K, Sun J, Girshick R	2015	Faster than fast R-CNN	Slow object proposal Slow implementation than YOLO
9.	DeepIDNet[15]	Ouyang W, Wang X, Zeng X, Qiu S, Luo P, Tian Y, Li H, Loy C, Yang S, Wang Z	2015	Learn deformation of objects with varying size and meaning	Verification issues occur
10.	YOLO[13]	Redmon J, Divvala S, Girshick R, Farhadi A	2016	Efficient unified object detector Extremely fast Less amount of background errors than fast R-CNN	Can't detect multiple objects within the same grid Loss in accuracy rate Possibility to detect one object multiple times. Unable to localize small size objects
11.	SSD[14]	Liu W, Anguelov D, Erhan D, Szegedy C, Reed S, Fu C, Berg A	2016	Faster than faster RCNN Works well for bigger objects	Doesn't generate much enough amount of higher level features for small objects
12.	RFCN[16]	Dai J, Li Y, He K, Sun J	2016	Faster than RCNN with acceptable accuracy Easier training Reduced complexity	Need more computation resources
13.	FPN[17]	Lin T, Dollar P, He K, Girshick R., Hariharan B, Belongie S	2017	Rich semantics in all levels	Removing top-down connection reduce accuracy
14.	DeNet[18]	TychsenSmith L, Petersson L	2017	Much faster than RCNN Predefined anchors not needed	More time spend for generating corners and for evaluating base network

V. SCOPE OF OBJECT DETECTION

Recent years have observed a tremendous refinement in the domain of object detection especially over remote sensing images. As the quantity and the quality of the remote sensing images have increased, it leads to both opportunities and challenges in this field. Remote sensing images are high quality images taken from the top of the earth at varying heights, at varying light conditions for variety of applications. The accuracy of the object detection results can be improved by increasing the count of the images. Even though the quantity of images has been increased, it is not up to the level of having an efficient system trained using a large number of training images. And also the increased number of objects over these images has increased the complexity of the image background which results in difficulty to analyze these images properly. Thus object detection is however a vital task for remote sensing images. Object detection is the task of detecting the presence of entire objects in the image such as roads, water sources, vehicles, buildings, forest fire etc. Detecting all the objects in the image came out to be an efficient and effective way to analyze these images.

Remote sensing images are of high importance and useful as it is used for environment monitoring and tracking, change detection and so on. Tracking the growth of a city, reduction in agricultural lands, decreasing the area of forest lands, disappearing water resources all these comes under the category of change detection and environment monitoring. Here we are actually analyzing images over a period of time say for over 20 years. From these images we can analyze and make some predictions over it such as the condition of any river. These are some exclusive applications of remote sensing images and is of research oriented. Monitoring the oil reserves, volcano eruption, military surveillance, monitoring and controlling the forest fire, plotting areas of oasis in desert, weather forecasting all those demand remote sensing images of high resolution in nature.

Object detection over remote sensing images module can be obtained by embedding or simply concatenating a three step process of candidate region proposals, feature extraction and final classification of converting these image proposals to labels. Candidate region proposal is the art of drawing regions of interest. These techniques must be more concentrating on regions of interest, in sense that there is some chance of having an object in the region of our interest, and not on regions that make no sense. Feature extraction deals with extracting the unique, high level feature of each proposal. The more proper and efficient region proposal leads to a better extraction of features in turn leads to an accurate object detection and classification.

VI. CONCLUSION

Deep learning based object detection has made a great advent over the past few years in the development as well as the use of advanced object detection techniques. This review provides a short and crisp idea about different object detection techniques and lists the highlights and challenges of each. The work also concentrates on the various applications of object detection. Apart from this, the paper also mention the promising future of object detection in the field of remote sensing images. This survey provides a valuable observation in object detection and promote new research.

REFERENCES

[1] Viola, P. and Jones, M., 2001. Rapid object detection using a boosted cascade of simple features. *CVPR (1)*, *1*, pp.511-518.

[2] Lampert, C.H., Blaschko, M.B. and Hofmann, T., 2008, June. Beyond sliding windows: Object localization by efficient subwindow search. In *2008 IEEE conference on computer vision and pattern recognition* (pp. 1-8). IEEE.

[3] Vedaldi, A., Gulshan, V., Varma, M. and Zisserman, A., 2009, September. Multiple kernels for object detection. In *2009 IEEE 12th international conference on computer vision* (pp. 606-613). IEEE.

[4] Felzenszwalb, P.F., Girshick, R.B., McAllester, D. and Ramanan, D., 2010. Object detection with discriminatively trained part-based models. *IEEE transactions on pattern analysis and machine intelligence*, *32*(9), pp.1627-1645.

[5] Girshick, R., Donahue, J., Darrell, T. and Malik, J., 2014. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 580-587).

[6] Uijlings, J.R., Van De Sande, K.E., Gevers, T. and Smeulders, A.W., 2013. Selective search for object recognition. *International journal of computer vision*, *104*(2), pp.154-171.

[7] Sermanet, P., Eigen, D., Zhang, X., Mathieu, M., Fergus, R. and LeCun, Y., 2013. Overfeat: Integrated recognition, localization and detection using convolutional networks. *arXiv preprint arXiv:1312.6229*.

[8] He, K., Zhang, X., Ren, S. and Sun, J., 2015. Spatial pyramid pooling in deep convolutional networks for visual recognition. *IEEE transactions on pattern analysis and machine intelligence*, *37*(9), pp.1904-1916.

[9] Gidaris, S. and Komodakis, N., 2015. Object detection via a multi-region and semantic segmentation-aware cnn model. In *Proceedings of the IEEE International Conference on Computer Vision* (pp. 1134-1142).

[10] Yoo, D., Park, S., Lee, J.Y., Paek, A.S. and So Kweon, I., 2015. Attentionnet: Aggregating weak directions for accurate object detection. In *Proceedings of the IEEE International Conference on Computer Vision* (pp. 2659-2667).

[11] Girshick, R., 2015. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision* (pp. 1440-1448).

[12] Ren, S., He, K., Girshick, R. and Sun, J., 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems* (pp. 91-99).

[13] Redmon, J., Divvala, S., Girshick, R. and Farhadi, A., 2016. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 779-788).

[14] Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.Y. and Berg, A.C., 2016, October. Ssd: Single shot multibox detector. In *European conference on computer vision* (pp. 21-37). Springer, Cham.

[15] Ouyang, W., Wang, X., Zeng, X., Qiu, S., Luo, P., Tian, Y., Li, H., Yang, S., Wang, Z., Loy, C.C. and Tang, X., 2015. Deepid-net: Deformable deep convolutional neural networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 2403-2412).

[16] Dai, J., Li, Y., He, K. and Sun, J., 2016. R-fcn: Object detection via region-based fully convolutional networks. In *Advances in neural information processing systems* (pp. 379-387).

[17] Lin, T.Y., Dollár, P., Girshick, R., He, K., Hariharan, B. and Belongie, S., 2017. Feature pyramid networks for object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 2117-2125).

[18] Tychsen-Smith, L. and Petersson, L., 2017. Denet: Scalable real-time object detection with directed sparse sampling. In *Proceedings of the IEEE International Conference on Computer Vision* (pp. 428-436).