Medicon Engineering Themes
Volume 2 Issue 4 April 2022
Review Article

Deep Learning based Object Detection Methods: A Review

Divya Mishra*

Department of Electronics and Communication Engineering, Jamia Millia Islamia, New Delhi, India

***Corresponding Author:** Divya Mishra, Department of Electronics and Communication Engineering, Jamia Millia Islamia, New Delhi, India.

Received: March 23, 2022; **Published:** March 29, 2022

DOI: 10.55162/MCET.02.027

Abstract

Recently, object detection has become one of the effective and popular trends in computer vision to deal with numerous applications such as in medical image processing of breast cancer, skin cancer, brain injuries, blood cells, and more. Also, it is used in video surveillance stations for real-time monitoring of crowd and anomaly detection. The application is widely used in satellite images and astronomy, fraud detection, and in the field of remote sensing to detect disaster-prone areas from satellite images so that important measures can be taken at the correct time to overcome or reduce the loss of life and property. For example, in medical applications earlier diagnosis procedures usually tend to figure out early diabetes, cancer and a few more diseases. Despite many existing object detection methods in state-of-the-art literature, it is getting harder to identify the best fit model for the specific application or dataset. Therefore, it is highly important and much needed to bring all the existing techniques to a single platform and mention their advantages and limitations. In this paper, a thorough literature review and comparison of various existing deep learning-based object detection methods are presented using three different parameters named mean average precision, frames per second, and data set used. Such information is useful for researchers and practitioners to identify the better approaches among the others easily according to the dataset in hand.

Keywords: Object detection; Deep learning; Convolution neural networks

Abbreviations

- ML – Machine Learning
- LR- Logistic Regression
- SVM- Support Vector Machine
- RF- Random Forest
- KNN- K- Nearest Neighbor
- MSA- Mean Shift Algorithm
- DT- Decision Tree
- RNN- Recurrent Neural Networks
- CNN- Convolutional Neural Networks
- GANS- Generative Adversarial Networks
- LSTM- Long Short-Term Memory
- FNN- Feed Forward Neural Network
- R-CNN- Region Based CNN
- SSD- Single Shot Detector
- YOLO- You only look once

- FPS- Frame Per Second
- SIFT- Shift Invariant Feature Transform
- HOG- Histogram of Oriented Gradients
- RoI- Region of Interest
- RPN- Region Proposal Network
- RFPN- ResNet Feature Pyramid Network
- CLU-CNNs- Clustering CNNs
- ANCF- Agglomerative Nesting Clustering Framework

Introduction

In the last few years, advancement in image processing and computer vision drastically helps modern applications and devices, specifically for medical imaging applications. Medical imaging consists of a group of techniques to create a visual depiction of the interior parts of the human body such as tissues for clinical purposes to diagnose and treat diseases and injuries [5, 16]. Traditional medical imaging incorporates radiology which includes the technologies such as radiography, magnetic resonance imaging, ultrasound, endoscopy, thermography, nuclear medicine imaging, and tomography, and more [5, 14, 15]. It is evident that the healthcare industry is a high-priority and sensitive sector where the majority of the analysis of medical data is done by medical experts. Normally, such analyses are quite limited to particular experts due to their complexity and sensitivity towards patient's safety.

The applications are also widely used in automatic detection and extraction systems for smart vehicles and devices; in live object tracking for surveillance and security monitoring systems, in underwater imaging systems for marine biological research and tracking underwater cables, in biometric recognition for security purpose, in industrial inspection systems helps to reduce the manpower and increase the accuracy of auditing, in satellite imagery systems and many more endless potential applications. Additionally, traditional Machine Learning (ML) algorithms cannot grasp the complexity of object detection problem statements due to the subject matters and complexity. A similar trend implies for the other fields such as surveillance stations for real-time monitoring of crowd and anomaly detection, satellite images and astronomy, fraud detection, and in the field of remote sensing to detect disaster-prone areas from satellite images.

The best existing state-of-the-art solutions to date are the traditional ML applications in computer vision [23]. Usually, ML algorithms are based on the hand-crafted features by experts or practitioners since they have hands-on experience in relevant subject matters. This can be done easily on a small-scale basis. However, this becomes complex and difficult as the data size grows, varies from subject to subject, and the quality of data analysis also varies with regard to the experience of the expert. For that reason, traditional learning methods were not trustworthy. Moreover, these different machine learning algorithms of Logistic Regression (LR), Naive Bayes, Support Vector Machine (SVM), Random Forest (RF), K-Nearest Neighbour (KNN), Mean Shift Algorithm (MSA), Decision Tree (DT) and more are taking raw image data into account without any learning of hidden delegations [5]. Besides, the data preprocessing and reshaping is also based on the knowledge of experts which ultimately consumes a lot of time and can be treated as labor-intensive work.

To address the limitations of the aforementioned traditional methods, deep learning has shown promising potential. In recent years, some of the greatest successes of deep learning have been in the field of computer vision [6, 18, 23]. Deep learning in computer vision has shown breakthroughs in capturing hidden patterns and extract features from them in the most accurate way. It has the advantage of automatically learning the most meaningful features directly from the images or data, rather than features extracted from them like in ML [24]. Furthermore, deep learning methods consists of algorithms, namely, Recurrent Neural Networks(RNNs), Convolutional Neural Networks(CNNs), Generative Adversarial Networks (GANs), Long Short-Term Memory (LSTM) Networks, Fully-connected Feed forward Deep Neural Networks(FNNs), which do not require manual preprocessing or handcrafted feature extraction on raw data.

Furthermore, deep learning-based object detection methods play a vital role in many domains for various applications [13]. Object

detection is one of the most substantial and challenging divisions of computer vision, which has been extensively enforced in numerous applications for the betterment. Alongside the speedy development of deep learning algorithms for detection tasks, the work of object detection has been considerably enhanced. As a result, deep learning models have been widely chosen in the entire computer vision field. There exist few surveys or reviews about object detection methods in the state-of-the-art literature [13, 11, 26, 25, 1, 16].

The existing review studies mainly focusing on the review of each methodology. To the best of the authors' knowledge, there is no proper study that performs a comparison among all existing deep learning-based object detection methods according to the dataset and application available and identify the better model in particular scenario. This is mainly to grasp the fundamental progress status of such models. Moreover, these object detection methods are categorized into two types, the first one is a two-stage detector, the conventional one, Faster Region-based Convolutional Neural Networks (R-CNN) [22]. The second one is a single-stage detector, such as Single-Shot Detector (SSD) [17] and You Only Look Once (YOLO) [19]. A detailed review of those techniques is presented in Section 3.

Hence, the contributions of this paper are as follows:

- We gathered all the existing deep learning-based object detection methods from state-of-the-art literature and presented them in this paper.
- We performed a review and then detailed comparison among different object detection methods so that the better method can be easily chosen for the relevant application or dataset.
- Our study is different from other existing studies, as it presents the comparison among different deep learning-based model accuracy through mean Average Precision (mAP) and speed computed in terms of Frames Per Second(FPS) along with the test dataset used.
- Conclusions drawn from this review are helpful for computer vision researchers, engineers, automation industry, live tracking security systems and medical practitioners.

The remainder of this paper is organized as follows. In Section 2, we briefly describe the traditional approaches for object detection. Section 3, describes the deep learning-based approaches for object detection. Section 4, performs the comparison between different deep learning-based object detection mechanisms from state-of-the-art research. Section 5 details our future work plans. Section 6, draws the conclusions of this paper.

Traditional Methods

Traditional approaches for object detection are not real-time due to large processing time. Also, the accuracy is not up to the mark as required for the implementation of practical applications. Non-neural networks for object detection require first feature extraction through any one of the methods like Viola-Jones object detection framework based on Haar-features, Scale-Invariant Feature Transform(SIFT), and Histogram of Oriented Gradients(HOG) features followed by an SVM classifier for classification. Deep learning-based approaches resolved these unnecessary steps of feature engineering by using Convolutional Neural Networks (CNNs) for image processing. CNN does feature extraction by automatically training and updating the network parameters without human interference with improved speed, accuracy, and performance.

Deep Learning Based Methods

Deep learning-based methods often suffer from problems with multiple parameters and hence are computationally complex and expensive with a slow rate of convergence. The classic traditional methods are not only computationally expensive but also have poor detection performance. To overcome this, sparse representation-based methods have been introduced [4].

R-CNN

The first breakthrough model using Convolution Neural Networks(CNN) for object detection was R-CNN [8]. The model generates 2000 region proposals per image for classification and resizes them in 227 x 227 dimensions. Further, CNN is used for feature

extraction and model training followed by an SVM classifier for object classification. The model is slow, taking 45 to 50 seconds to process a single image. Later, Fast R-CNN is proposed in 2015 to overcome the accuracy by 8% and speed by 9 times [7]. Followed in the same series, region proposal networks are proposed in Faster R-CNN for feature extraction and to get rid of storage cost [22]. The model is good in terms of accuracy and speed than previous models but suffers from misalignment of bounding boxes of ground truth and predicted one. To address this, Mask R-CNN is introduced by authors to avoid error due to the quantization process in the Region of Interest (RoI) pooling layer [9]. The overall detection process is shown in Fig.1.

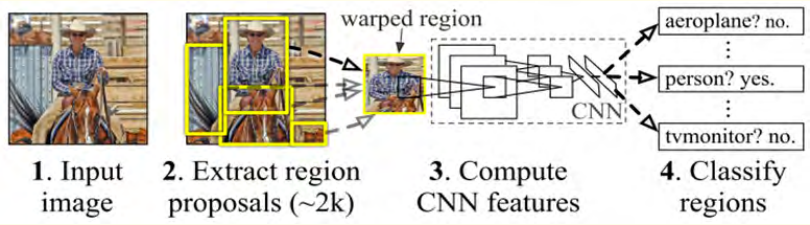

Figure 1: Object detection overview in R-CNN [8].

FAST R-CNN

In 2015, the model Fast R-CNN with improved speed and accuracy was presented [7]. R-CNN takes a lot of time to classify each region proposals separately by multiple linear SVM classifiers for each object. This causes large computation costs and resources and more time to process. All these limitations were improved by Fast R-CNN in 2015. The network is single-stage streamlined rather than multi-stage as in R-CNN. Here, a finite number of object proposals are directly used as an input, and convolution operations are used only once per image as shown in Fig. 2. This significantly decreases the computations and enhances the time of processing and hence speed of detection from 2s to 47s in R-CNN.

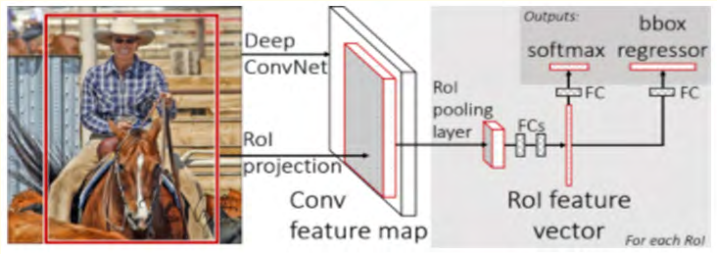

Figure 2: Object detection overview in Fast R-CNN [7].

FASTER R-CNN

Region-based CNN's require some storage space and hence increases the cost of the network. Here, in Faster R-CNN cost-free solution is used for detection tasks using Region Proposal Networks (RPN) followed by classification network same as used in Fast R-CNN [22]. The model has 3% more accuracy and 10 times faster in detection performance than Fast R-CNN. The drawback is it still suffers problems in the detection of small-sized objects. The complete architecture has been shown in Fig. 3.

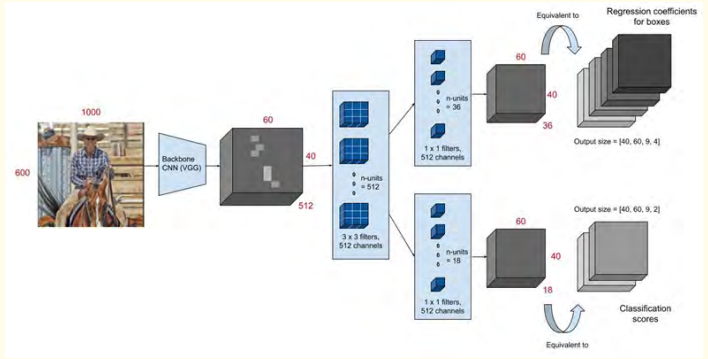

Figure 3: Architecture Overview of Faster R-CNN.

MASK R-CNN

The model is more particularly for instance segmentation tasks that help to detect small-sized objects. The network introduced a segmentation mask to classify and predict on a pixel-to-pixel basis. The ResNet-Feature Pyramid Network (R-FPN) is used as a feature extraction network for both improved accuracy and speed as shown in Fig. 4.

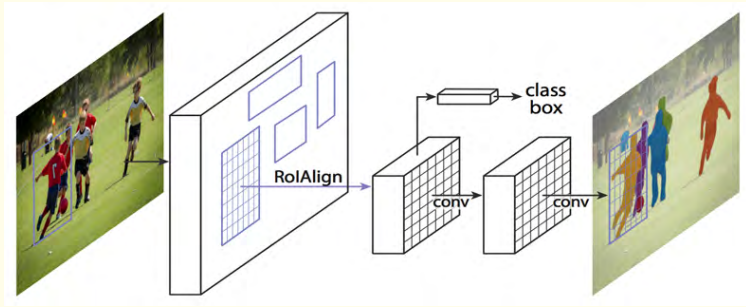

Figure 4: Architecture Overview of Mask R-CNN.

YOLO

Previous deep learning-based object detectors are two-stage detectors followed by multiple sequential processes that hurdle in real-time applications. To mitigate this, a single-stage network with fast frame processing speed is required. YOLO was such a model that encapsulated both, the regression task for object localization and the classification task to detect an object class in a pipeline as shown in Fig. 5. Later, successive versions of YOLO came YOLO-v2 [20], YOLO-v3 [21], YOLO-v4 [2] for progressive improvement in real-time application for object detection. The YOLO versions are good in accuracy, fast, easy, and better generalization ability on other unseen datasets.

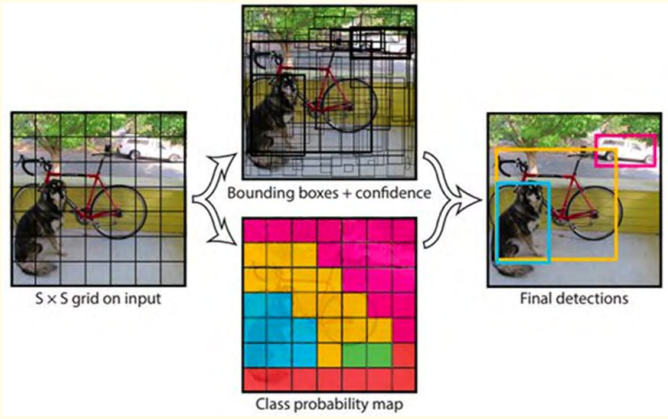

Figure 5: Architecture Overview of YOLO Model.

SSD

SSD is a single-stage network that has two modules [17]. First, the pre-trained image classification network like ResNet trained on ImageNet followed by the second module, the SSD head. The model is faster than Faster R-CNN due to the removal of bounding boxes proposals and image sub-samples. Also, it is faster and more accurate than YOLO due to the introduction of multi-scale feature maps rather than fully connected layers in YOLO as shown in Fig. 6.

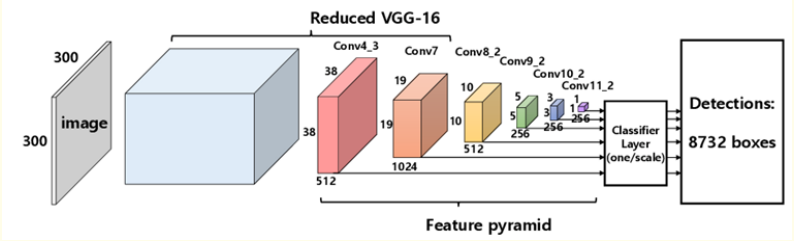

Figure 6: Architecture Overview of SSD Model.

CLU-CNN

Features of medical images are different from natural images, so normal image object detection methods would perform poorly on medical images. Apart from that, the data availability of medical images is tougher and costly than other natural images. Clustering-Convolutional Neural Networks (CLU-CNNs), particularly for medical images based on adaptation framework is proposed without specific domain adaptation training [15]. The model is fast, easy to modify, accurate, and performs well on a small training dataset. The model is fast than Faster R-CNN [22] and SSD [17] and produces more accurate results when there is a significant difference between source and target domain due to the presence of BN-IN Net wrapping for the model's stability. The model is fast than Faster R-CNN and SSD and produces more accurate results when there is a significant difference between source and target domain due to the presence of BN-IN Net wrapping as shown in Fig. 7. BN-IN Net is used for model's stability.

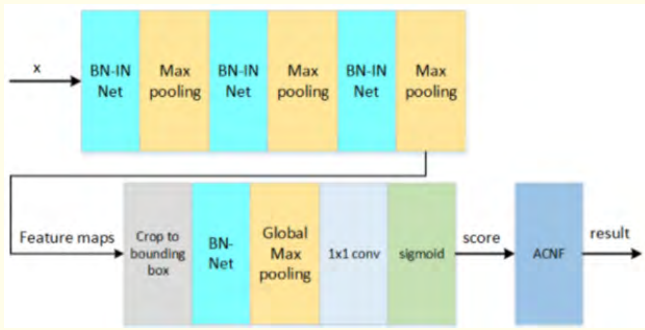

Figure 7: The Framework of CLU-CNN.

CLU-CNN consists of two parts, the first one is the fully convolutional networks to improve its performance and generating feature maps from input images. The other one is Agglomerative Nesting Clustering Framework (ANCF), an error correction mechanism without special training in object detection. It deals with a dataset having different distributions and improving performance for those networks whose size is limited and are not potential for positional accuracy. It is a hierarchical clustering algorithm. Firstly, it assigns each element as a cluster, then measures the distance between two clusters. The two clusters closed are merged based on the minimal distance between them in the first iteration. Similarly, multiple iterations are performed till the expected count of clusters reduced to the value we are interested in. The process of ANCF is shown in Fig. 8.

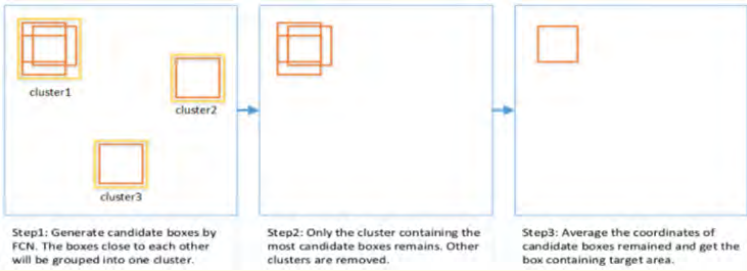

Figure 8: The process of ANCF.

CLU-CNN is smaller, easier, and faster than Faster R-CNN. It can be modified to cope with different tasks. It is fast and avoids unnecessary complex computational costs which is a hurdle in medical image diagnosis.

Comparison

The section is focused on the comparison of speed and accuracy (in mAP%) on different existing deep learning object detection methods. We did not include CLU-CNN here for comparison purposes since it is particularly used in medical image datasets while the other datasets are natural image datasets. The performance is shown here on standard image datasets used by previous authors in YOLOv2 [20]. Table 1. presents a comparison of various deep learning-based object detection methods.

Detection Framework	mAP(%)	FPS	Test Dataset
Fast R-CNN [7]	70.0	0.5	PASCAL VOC-2007
Faster R-CNN VGG-16 [22]	73.2	7	PASCAL VOC-2007
Faster R-CNN ResNet [22]	76.4	5	PASCAL VOC-2007
YOLO [19]	63.4	45	PASCAL VOC-2007
SSD (300 x 300) [17]	74.3	46	PASCAL VOC-2007
SSD (500 x 500) [17]	76.8	19	PASCAL VOC-2007
YOLOv2 (288 x 288) [20]	69.0	91	PASCAL VOC-2007
YOLOv2 (352 x 352) [20]	73.7	81	PASCAL VOC-2007
YOLOv2 (416 x 416) [20]	76.8	67	PASCAL VOC-2007
YOLOv2 (480 x 480) [20]	77.8	59	PASCAL VOC-2007
YOLOv2 (544 x 544) [20]	78.6	40	PASCAL VOC-2007
YOLOv3 (320 x 320) [21]	28.2	22	MS-COCO
YOLOv3 (416 x 416) [21]	31.0	29	MS-COCO
YOLOv3 (608 x 608) [21]	33.0	51	MS-COCO
YOLOv4 (512 x 512) [2]	43	83	MS-COCO
YOLOv4 (608 x 608) [2]	43.5	65	MS-COCO

Table 1: Comparison of various deep learning-based object detection frame-works.

For seeking the best object detection model, firstly it is important to consider the application. There are use-cases where accuracy is a prime concern rather than fast detection speed like surveillance systems, medical image analysis, navigation systems, and many more similar critical applications. For all tested cases where accuracy is a prime concern, Faster R-CNN with 300 proposals using Inception ResNet [10] architecture gives the best accuracy. Similarly, for real-time processing applications SSD on MobileNet [12] outcomes with the highest accuracy. Again, object size matters a lot in detection, for large-sized objects, SSD proves to be a good detector with a simple extractor network while it performs worse on small-sized object detection. For small object detection, YOLO-v3 has significant growth in accuracy.

One other way is to use image super-resolution as a pre-processing step to enhance the image quality and hence detection. It has been observed that decreasing the image resolution by a factor of 2 results in a reduction of both accuracies by 15.88% and inference time by 27.4% on an average. If speed is required then R-FCN [3], SSD and YOLO models are best to use. Again, for real-time scenarios, YOLO models are best in terms of speed and generalization of images to other unseen datasets. Again, for medical images, CLU-CNN gave a very good response with decent accuracy and even performs well on the small training dataset. The trade-off between accuracy and speed is always there, but we have to choose wisely according to the condition and datasets under the experiment.

Future Work

In this review paper, we presented the findings and limitations of deep learning-based object detection from the existing state-of-the-art literature. One of the further research plans is to propose an image super-resolution model as a pre-processing task for image enhancement to improve the visual quality and hence object detection. Another research direction can be more accurate image segmentation at the pixel level, learning and training of neural network from a small dataset and generalize on the large unseen dataset. The area of object detection for images can be extended for video in action recognition and live monitoring of small objects and the public.

Conclusion

Although, there are few review papers presenting similar approach but our idea is unique since it presented all detailed and thorough literature review on various existing deep learning-based object detection methods with their strengths and limitations comparision according to the potential application and dataset availability. The presented object detection approaches are very useful for medical imaging processing and its applications for medical researchers and practitioners, in navigation and marine systems, automatic self-driving vehicles like driver-less metro trains and cars are live examples, in surveillance and security systems and people detection,

so that they can integrate the best-fit approach into an application.

References

1. Bochkovskiy Alexey., et al. "YOLOv4: Optimal Speed and Accuracy of Object Detection". ArXiv (2020).
2. C Bhagya and A Shyna, "An Overview of Deep Learning Based Object Detection Techniques". 2019 1st International Conference on Innovations in Information and Communication Technology (ICIICT) (2019): 1-6.
3. Dai Jifeng., et al. "R-fcn: Object detection via region-based fully convolutional networks". Advances in neural information processing systems 29 (2016).
4. Dong Minghui., et al. "Sparse fully convolutional network for face labeling". Neurocomputing 331 (2019): 465-472.
5. Dutta S. "A 2020 guide to deep learning for medical imaging and the healthcare industry". Nanonets.com (2020).
6. Esteva Andre., et al. "A guide to deep learning in healthcare". Nature medicine 25.1 (2019): 24-29.
7. Girshick Ross. "Fast r-cnn". Proceedings of the IEEE international conference on computer vision (2015).
8. Girshick Ross., et al. "Rich feature hierarchies for accurate object detection and semantic segmentation". Proceedings of the IEEE conference on computer vision and pattern recognition (2014).
9. He Kaiming., et al. "Mask r-cnn." Proceedings of the IEEE international conference on computer vision (2017).
10. He Kaiming., et al. "Deep residual learning for image recognition". Proceedings of the IEEE conference on computer vision and pattern recognition (2016).
11. Hechun Wang and Zheng Xiaohong. "Survey of deep learning based object detection". Proceedings of the 2nd International Conference on Big Data Technologies (2019).
12. Howard AG. "Mobilenets: E±cient convolutional neural networks for mobile vision applications". Computer Science (2017).
13. Jiao Licheng., et al. "A survey of deep learning-based object detection". IEEE access 7 (2019): 128837-128868.
14. Kieu Phat Nguyen., et al. "Applying Multi-CNNs model for detecting abnormal problem on chest x-ray images". 2018 10th International Conference on Knowledge and Systems Engineering (KSE). IEEE (2018).
15. Li Zhuoling., et al. "CLU-CNNs: Object detection for medical images". Neurocomputing 350 (2019): 53-59.
16. Litjens Geert., et al. "A survey on deep learning in medical image analysis". Medical image analysis 42 (2017): 60-88.
17. Liu Wei., et al. "Ssd: Single shot multibox detector". European conference on computer vision. Springer, Cham (2016).
18. Maier Andreas., et al. "A gentle introduction to deep learning in medical image processing". Zeitschrift für Medizinische Physik 29.2 (2019): 86-101.
19. Van Etten Adam. "You only look twice: Rapid multi-scale object detection in satellite imagery". arXiv (2018).
20. Redmon Joseph and Ali Farhadi. "YOLO9000: better, faster, stronger". Proceedings of the IEEE conference on computer vision and pattern recognition (2017).
21. Redmon Joseph and Ali Farhadi. "Yolov3: An incremental improvement". arXiv (2018).
22. Ren Shaoqing., et al. "Faster r-cnn: Towards real-time object detection with region proposal networks". Advances in neural information processing systems 28 (2015).
23. Shanahan James G and Liang Dai. "Introduction to computer vision and real time deep learning-based object detection". Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (2020).
24. Sharma Alok., et al. "DeepInsight: A methodology to transform a non-image data to an image for convolution neural network architecture". Scientific reports 9.1 (2019): 1-7.
25. Xiao Youzi., et al. "A review of object detection based on deep learning." Multimedia Tools and Applications 79.33 (2020): 23729-23791.
26. Zhao Zhong-Qiu., et al. "Object detection with deep learning: A review". IEEE transactions on neural networks and learning systems 30.11 (2019): 3212-3232.

Volume 2 Issue 4 April 2022