JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

A PERSPECTIVE REVIEW ON OBJECT DETECTION USING COMPUTER VISION

[1]P.Nagaraju, [2]Dr.Manchala Sadanandam

[1]Researach Scholar, [2]Professor
[1,2]Department of Computer Science & Engineering,
[1,2]Kakatiya University, Warangal, Telangana, India

Abstract: Recent years have seen deep neural networks emerge as the most significant advancement in the field of computer vision (CV), producing outstanding results for picture categorization. Convolution neural networks (CNNs) are thought to be an intriguing tool for researching biological vision since this class of artificial vision systems has visual identification abilities that are comparable to those of human observers. In the world of Internet applications nowadays, CV technology has taken on significant importance. Object detection (OD) has evolved into one of the fundamental issues in CV and is now the foundation of many vision tasks. It offers reliable data, whether we want to understand how pictures and text interact or identify precise classifications. The advancement of OD networks is examined in this article. We provide candidate region-based OD, including R-CNN, Retina Net, YOLO, and SSD, etc. Currently, detectors are the best approach. It offers recommendations for the future development trend and OD research by examining the existing research state of OD networks. We are concentrating on the fundamental components of the architecture of CNN in this article and CNN-based OD techniques are presented.

Index Terms **- Deep Learning (DL), Convolutional Neural Network (CNN), Object Detection (OD), Computer Vision (CV), object recognition (OR), Machine learning (ML), Neural Network (NN).**

I. INTRODUCTION

CNN is a sparkling treasure in the growing deep neural network treasure house and has made significant strides in recent years. Additionally, AI can now see and comprehend visual information thanks to CV technologies [1]. DL-based CV algorithms have been very successful in recent years at performing well in traditional CV tasks like image classification, OD, and image segmentation. This is due to improvements in computer hardware performance and the development of large-scale image annotation data sets. CV is a crucial area of computer science that enables a machine to comprehend and anticipate visual data in order to produce the intended results, much like the human brain does with retinal input [2]. By identifying natural patterns in data, ML algorithms produce insight and assist us in making better judgments and predictions. OD in images, computational biology, energy production, natural language processing, automotive, aerospace, and manufacturing are just a few of the areas where ML has found considerable use recently [3].

Currently, OD is heavily employed in both academia and industry, including video fire detection [4], autonomous vehicle operations [5], security monitoring [6], and UAV scene analysis [7-9]. Currently, there are two primary categories of OD algorithms: CNN-based algorithms and more conventional methods based on image processing. Girshick et al. presented R-CNN [10] in 2014. CNNs were employed for OD for the first time, and the results were very well received. The detection accuracy was increased by around 30% in comparison to conventional detection techniques. According to recent academic studies and real-world applications, CNN based OD algorithm is more accurate and takes less time to test than the conventional methodology, which has almost entirely supplanted [11].

ML is becoming nearer to AI because of the enormous results in a variety of DL applications. The term "AI" refers to more than simply computers and robots. As technology advances, brilliant thinkers from all around the world are embracing its revolutionary potential to aid people in ways that, not even very long ago, may have looked like moonshots. AI and ML are used to help the blind and visually challenged better evaluate data from cameras and other sensors. DL algorithms are used in many technologies [12]. These tools enable vision impaired persons to recognize objects and navigate independently by providing in-depth descriptions of images. One of the technologies with the quickest growth in this expanding market is DL. It processes data and generates patterns for use in decision-making using neural network (NN) functions that mimic the actions of the human brain. In the real-time applications, we are detecting and following objects using DL techniques. OD and Tracking System makes an effort to find, follow, and identify objects of interest throughout several movies as well as, more broadly, to decipher object behaviors and activities. Different tracking approaches are used to compare object recognition across numerous consecutive frames of a movie and compare the movement of an item. This methodical technique involves watching how items behave and utilizing a DL algorithm to process such behaviors in order to identify objects [13].

II. CONVOLUTIONAL NEURAL NETWORK

CNNs, commonly known as ConvNets, are a subclass of feedforward NNs that are particularly effective in object recognition and other CV related tasks. The main advantage of a CNN over a NN is its unique structure, which is depicted in Fig. 1. This structure uses shared weights to capture the local properties of the signal and sparse local connectivity between layers to reduce the number of parameters used in calculations [14]. Similar to regular NNs, CNNs include many consecutive layers where the inputs from one layer become the outputs of the following layer. On CNNs, the majority of NN ideas are used, including back

propagation and stochastic gradient descent for weight estimation. CNN employs the three key concepts of local receptive fields, pooling, and shared weights and biases to make training quicker, deeper, and with more layers [15-16].

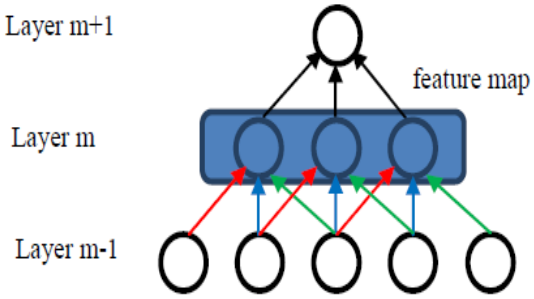

Fig. 1. Basic structure of CNN

Fig. 2 shows an example of CNN architecture. Input layer, convolutional layer (CL), ReLU layer, pooling layer (PL), and fully connected layer (FCL) make up the majority of the layers of CNNs. These layers can be combined to create a full CNN. The CL also undergoes the activation function following the convolution operation because in practical applications, both the CL and the ReLU layer are frequently referred to as the CL. More specifically, when the CL and the FCL convert the input, various parameters, including the neuron's weight and deviation, will also be employed. The ReLU layer and the PL then execute a fixed function operation. In order to the classification score given by the CNN to match the label of each picture in the training set, the parameters in the CL and the FCL will be learned as the gradient decreases. The ideas of local receptive fields [17], sparse weights, and parameter sharing are included in CNNs. CNNs are more suited for learning from picture input because of these three ideas, which provide them a certain degree of translation and scale invariance compared to other NNs [18].

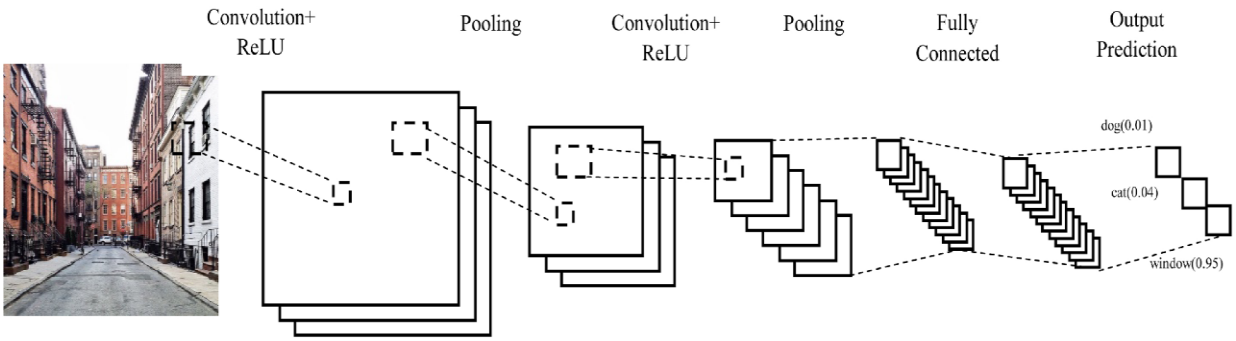

Fig. 2. CNN framework- Architecture [11]

2.1 Convolutional Layer

There are several filters in the CL. The learnable parameters of the layer are these filters' values. When referring to CNNs, the objective of a convolution is to extract the features from an image while maintaining the spatial relationship between the pixels and the learnt features inside the picture by using small, evenly-sized tiles. The main layer of a CNN, which produces the majority of the network's computations, is the CL. Keep in mind that the number of parameters does not equal the amount of calculation. Convolutional networks are simpler to train than fully connected networks of the same size because convolutional networks may effectively minimize the training difficulty of the network model and lower the network connection and parameter weights.

In a procedure known as "ordinary convolution," the convolution kernel is applied to the picture, and after a series of matrix operations, the task of determining the grey value of every pixel in the image is eventually accomplished. Transposed Convolution (dilated convolution), which is the opposite of ordinary convolution, is the convolution approach from low-dimensional feature mapping to high-dimensional feature mapping. Semantic analysis, image recognition, and other domains make extensive use of it.

The lightweight network model MobileNets [19] uses depth wise separable convolution, which applies a single filter to each input channel. Point wise convolution is then used to merge the outputs of several depth convolutions. In typical convolution procedures, depth separable convolution makes it possible to separate channels and regions. The amount of computation and the size of the model may both be significantly reduced by this decomposition procedure.

2.2 Activation Layer

Artificial neural networks (ANNs) may learn complicated patterns in data with the use of the activation function. The activation function eventually selects the material to be broadcast to the next neuron, much like the neuron-based model in the human brain. As seen in Fig. 3, the rectification function in linear form One of the most important unsaturated activation mechanisms is ReLU. Its mathematical formulation is as follows, $f(x) = \max(0, x)$.

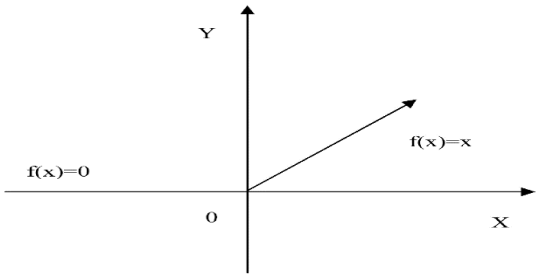

Fig. 3. ReLU function image.

2.3 Pooling Layer

Pooling is one of the ConvNets' distinguishing principles, as previously indicated. The goal of the pooling stage is to minimize the dimensionality of each feature map while maintaining the crucial data by removing redundant and noisy convolutions from the computing network. There are several forms, including Max, Sum, and Average, but max-pooling is the most popular and widely used. In max-pooling, a spatial neighborhood is constructed, and the maximum unit is obtained from the feature map depending on the filter dimension, which may be, for instance, a 2x2 window.

The PL was introduced for the first time in the LeNet [20] article, which was given the term Subsample following the release of the AlexNet [21] paper. It is one of the elements of modern CNNs that is often employed. The PL is layered between subsequent CLs to minimize over fitting by compressing the quantity of input and parameters. The PL's primary job when the input is an image is to compress the picture. The matrix's size can be efficiently decreased using the PL. On the special diagnosis at several locations within the local region of the picture, it may carry out collective statistical operations.

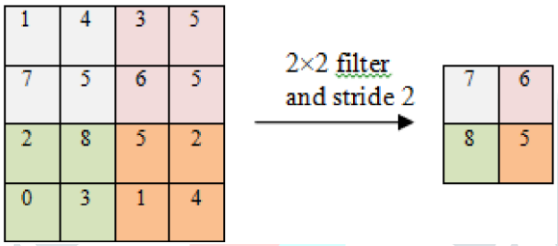

Fig. 4. Example of max-pooling

2.4 Fully connected layer

The FCL, one of the last layers of a CNN that comes before the output layer, functions like a CNN after the CL and PLs. There is a connection between every neuron in the layer before the FCL and every neuron in the FCL. The FCL's objectives are to categorize the picture using the training dataset and the output features from the preceding layer. In essence, a CNN's FCLs act as a classifier, using the outputs of the CLs as the classifier's input.

III. Object Detection Algorithms and Network Model

DL has made significant advancements in OD systems in recent years. There are many algorithms categories that may be applied to the most used OD techniques nowadays; R-CNN, Fast R-CNN, Faster R-CNN, Yolo and SSD [22-26]. The one-stage object identification technique executes feature extraction, target classification, and position regression in the full NN before obtaining the target location and category. It does not extract candidate areas through the intermediary layer. The two-stage OD algorithm's OD accuracy is somewhat better than the recognition accuracy. According to the concept, the pace has significantly increased.

One of the most obvious applications of CNNs is object recognition in pictures. Due to differences in each object or the particular image, such as the lighting or angle, it might be difficult to identify an object in an image. Identifying and locating items in a picture is what OD means [27-28]. After Krishevsky et al. (2012) used a CNN to exhibit considerable performance increases on the ImageNet classification test, several successes for OD using CNN were possible. The following subsections explore and contrast several CNN-based approaches to OD [29].

3.1 Region-based CNN

In 2014, Ross Girshick introduced R-CNN [10], which eliminates window duplication and lowers the technique's temporal complexity by using a selective search algorithm in place of the sliding window. CNNs, which can more accurately extract an image's features and enhance a network's capacity to resist interference, are used in place of the conventional hand-made feature extraction component. With the help of the Selective Search algorithm, RCNN first chooses potential object frames from a group of object candidate frames. It then resizes the pictures in these chosen object frames to a fixed size image and sends them to CNN. The classifier analyses the extracted features to determine if the picture in the object frame contains the target to be detected. The regression is then used to determine which category the detection target belongs to. Using a convolutional network, the characteristics from each area proposal are extracted. Each proposal's sub-image is twisted to achieve the necessary CNN input size before being sent to the CNN. The final classification is then produced when the extraction features are fed into Support Vector Machines (SVM). The R-CNN steps are shown in Fig. 5.

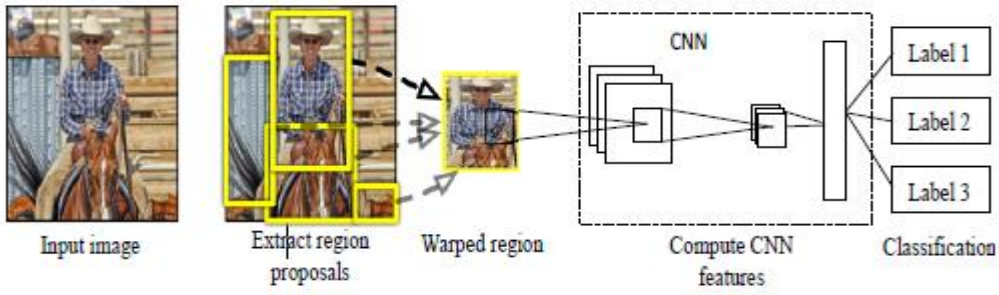

Fig. 5. R-CNN Stages

3.2 Fast R-CNN

In 2015, Fast R-CNN was introduced. When Fast R-CNN and R-CNN frameworks are compared, there are two key differences: first, a ROI PL is added after the final CL, and second, the loss function uses a multi-task loss function, with the Bounding Box Regression being added directly to the CNN network for training. Instead of extracting features for each image block numerous times, Fast R-CNN initially extracts the features of the entire picture using the CNN network. The approach of making candidate areas may then be directly applied to the extracted feature maps. To create a convolutional feature map, the entire picture is first processed using a number of CL and PLs. Instead of conducting a CNN for each region of interest, Fast R-CNN runs a single CNN for the whole picture, creating the feature map at the conclusion. This initial operation is one of the speedups that Fast R-CNN accomplishes when compared to R-CNN. After the first step is finished, the second stage of the framework starts, where a RoI PL uses maximum pooling to obtain a tiny fixed size vector from the feature map for each item proposition.

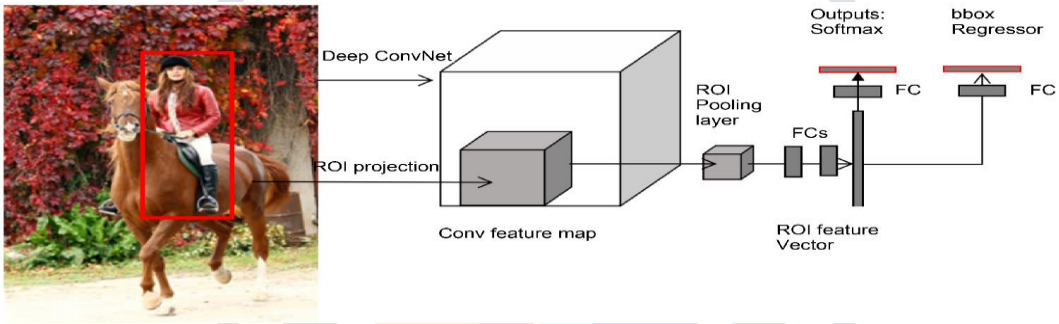

Fig. 6. Fast R-CNN

3.3 Faster R-CNN

Faster R-CNN developed by Microsoft Research in 2016[30], has demonstrated excellent performance in OD tasks. The Region Proposal Networks (RPN) are presented in this framework. For the purpose of identifying areas from an input picture that most likely contain objects, a separate RPN is employed. In order to get a final determination on an object's presence and bounding boxes, these area proposals are given to a detection network. The crucial factor is that because RPN and the OD network share CLs, calculations will be more effective. By considering both objectness ratings and map coordinates, it generates several region recommendations with various sizes and aspect ratios. Selective search approach coupled with R-CNN and Fast R-CNN is significantly slower than RPN combined with Fast R-CNN in the Faster R-CNN model. Using a notion called RPN, which is a slight expansion of the original recognition network, faster R-CNN combines calculations for object identification and recognition back into one.

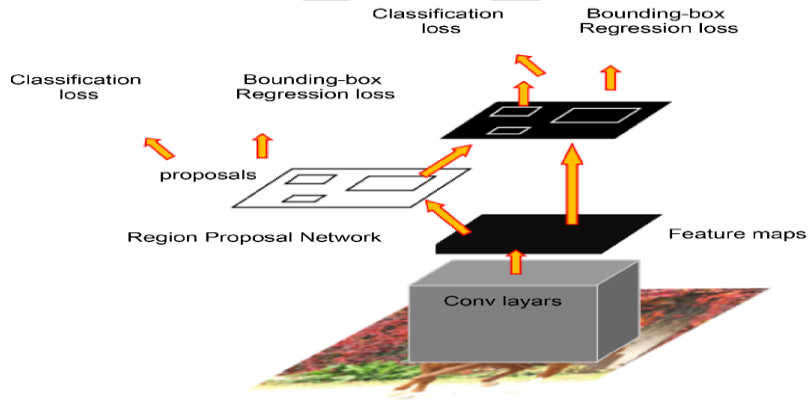

Fig. 7. Faster R-CNN

3.4 Mask R-CNN

HeKaiming introduced Mask R-CNN in 2017[31]. Faster-original RCNN's modification, Mask R-CNN, adds a branch to utilize current detection to forecast the target concurrently. In addition, this network structure can be readily extended to other sectors, such as OD, segmentation, and key point detection of humans, and it is reasonably simple to design, train, and operate. ResNet-FPN is utilized to extract features from the image initially, followed by the RPN network to forecast proposals, RoI Align to extract features from the image, classification, and detection heads, and lastly the mask detection head, meaning that each category predicts a mask.

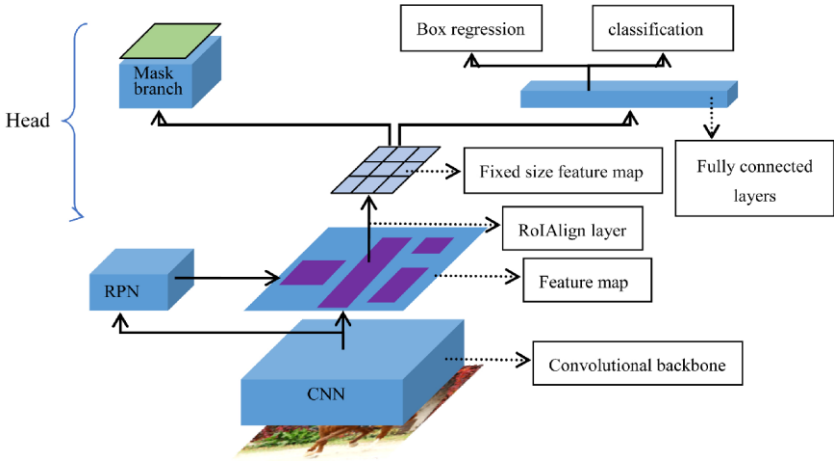

Fig. 8. Mask R-CNN framework

3.5 Trident Net

TridentNet developed in 2019 and conducted pertinent experimental verifications [32]. TridentNet was the first to suggest the effect of receptive fields on objects of various sizes in OD tasks. In order to prevent the insertion of extra parameters and an increase in computation during inference, the parameter sharing approach is used to suggest three branches during training and only one of them is used during testing.

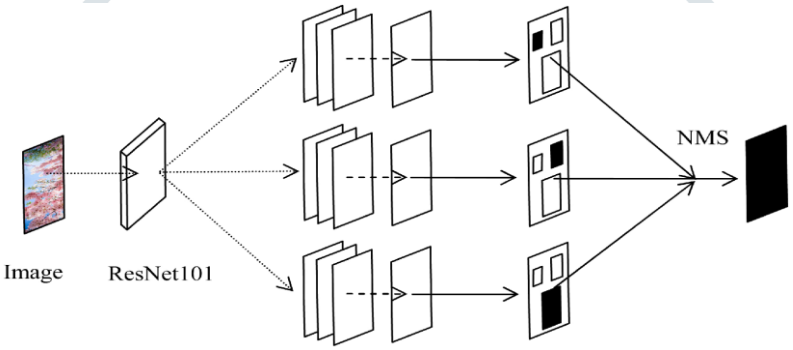

Fig. 9. Trident Net framework

3.6 D2Det

Based on the two-stage technique, the classification and regression branches were upgraded in 2020 to increase the object identification and instance segmentation accuracy even further. They suggested D2Det, a technique that is accurate in both location and classification [33]. This study presents a dense local regression approach for accurate placement, which forecasts several dense box offsets for each target candidate box. This work presents a discriminative RoI pooling strategy for precise classification. It may take samples from several sub-regions within a candidate area and use adaptive weighting throughout the computation to provide discriminative features.

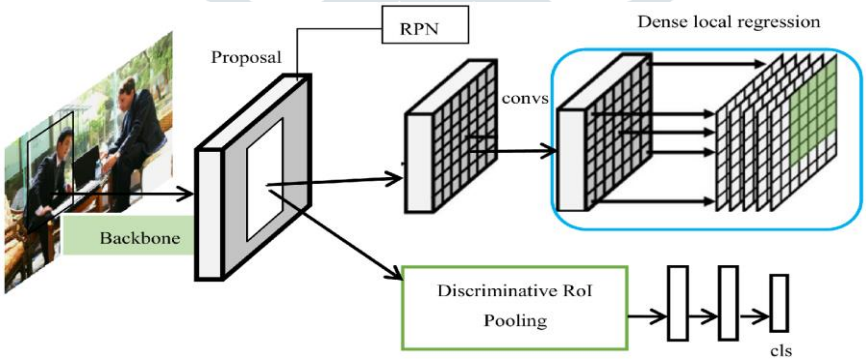

Fig. 10. D2Det framework

3.7 Sparse R-CNN

The majority of the earlier target detectors are dense detectors that are built on dense suggestions that are already present in the feature map network or picture grid [34]. These ideas' grid scores and offsets are anticipated, evaluated by the intersection ratio (IOU), and then filtered using non-maximum suppression (NMS). The dense-sparse detector, which first extracts a small number of foreground boxes from the dense suggested regions, or regional candidate boxes, before classifying and regressing the position of each regional candidate box, narrows the field of candidates down from thousands to a select few prospects, makes up the small portion. Sparse R-CNN avoids many-to-one positive and negative sample allocation as well as the manual selection of several hyper-parameters for candidate boxes. Furthermore, the ultimate prediction outcome may be generated directly without NMS [35].

3.8 You Only Look Once (YOLO)

In 2016, YOLO was presented [25]. YOLO's prediction is based on the complete image, and it will output all detected target information at once, including category and position, in contrast to the R-CNN series, which requires to determine the candidate area first before identifying the objects in the candidate area. YOLO's initial step is to partition the image. It creates grids that are of the same size and splits the image into grids. Using the complete image as the network's input and only passing through a NN to obtain the position of the bounding box and its category, YOLO's central notion is to transform object recognition into a regression issue. It has a very quick detection rate, good generalization potential, quick delivery, and decreased accuracy. The drawback is that overlapping items cannot be recognized for tiny objects.

In 2017, Joseph Redmon and Ali Farhadi significantly improved YOLOv1 and put forward YOLOv2, which concentrated on addressing YOLOv1's recall rate and placement accuracy issues [36]. In contrast to YOLOv1, which incorporates the concept of Faster R-CNN and adds the Anchor mechanism, YOLOv2 employs the FCL to indirectly forecast the coordinates of the Bounding Box. The training data is clustered and a better Anchor template is calculated using the K-means clustering approach, which significantly raises the algorithm's recall rate. The shallow and deep features of the picture are connected simultaneously when the fine-grained characteristics are combined; this aids in the identification of small-scale objects.

Based on YOLOv2, Redmon made various upgrades in 2018. The feature extraction portion employs the feature pyramid network structure to achieve multi-scale detection in place of the darknet-19's original network structure. Instead of using softmax, the classification approach substitutes logistic regression, which ensures OD accuracy while accounting for real-time performance. The classifier or location is used again by YOLOv3's earlier detection system to carry out detection operations. They apply the model to various picture sizes and locales. The test results might be those sections with higher scores. Additionally, they employ a totally distinct methodology in comparison to previous OD techniques. To the entire image, they apply a single NN. The network segments the picture into many areas and forecasts the probability and bounding box of each region. The expected probability is weighted while determining these boundary boxes. Comparing the model to classifier-based systems, there are several benefits. During the test, it examines the entire image, so its prediction makes use of the overall data in the picture.

Bochkovskiy and others introduced YOLOv4 [27] in 2020. As a result of extensive testing, YOLOv4 chose the following practical Tricks: WRC, CSP, CmBN, SAT, Mish activation, Mosaic data augmentation, CmBN, DropBlock regularization, and CIoU loss. To achieve the optimum balance between detection speed and accuracy, YOLOv4 expands on the conventional YOLO by incorporating these useful talents.

3.9 RetinaNet

RetinaNet was proposed by Lin et al. in 2017 [38]. Because the positive and negative samples are not balanced, they think the one-stage technique is quicker but less accurate than the two-stage method. The cross-entropy inaccuracy of the regression task was modified to focal loss by the one-stage detector to address the obstacle difficulty of category unbalance during the training stage. A cross entropy loss that may be dynamically zoomed is called focal loss. The zoom factor diminishes to zero as the confidence in the proper category rises. When used during training, the zoom factor can automatically **lessen** the weight of the loss caused by simple instances.

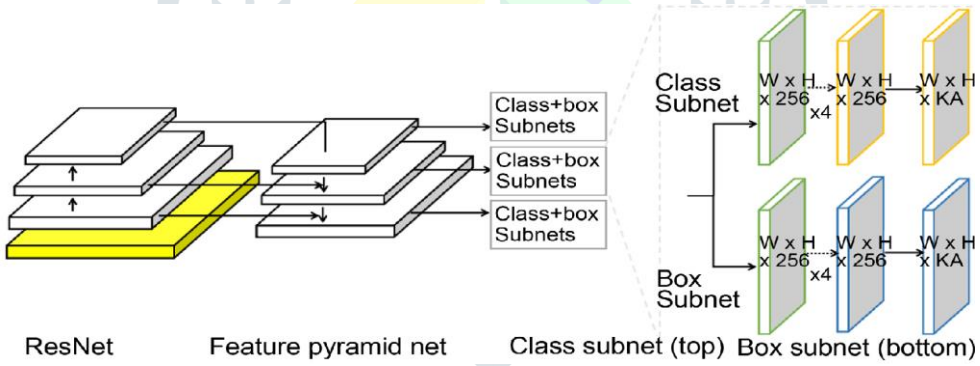

Fig. 11. RetinaNet framework.

3.10 CornerNet

CornerNet [39] was published on ECCV2018 by Hei et al. They suggested solving the object identification problem as a key point detection problem, which would include finding the two key points in the top left corner and bottom right corner of the target frame in order to produce the prediction frame. Using this method, we can detect the objects. The industry is rather inventive and capable of producing quality outcomes. The complete detection network is being trained from scratch without using a previously taught classification model.

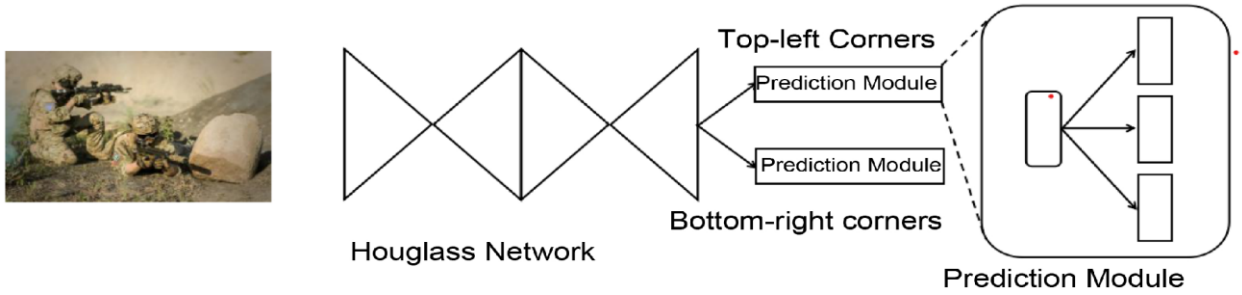

Fig. 12. CornerNet framework

IV. Conclusion

This article presents the best deep-learning models for real-time object detection and recognition. Universal object identification technology has advanced significantly in recent years due to the quick development of DL technology. The efficiency and speed of the detection model and the humanized performance, however, remain far apart. By adding convolution layers, CNNs considerably improve the capabilities of feed-forward networks like MLP. CNNs have two key advantages: (1) sharing parameters, which lowers network complexity and computing costs; (2) sparsity of connection, this makes CNN translation invariant and facilitates training with fewer training sets because each output value only depends on a limited number of inputs. Every current CV model is built on CNNs. CNNs have been used to perform a variety of perceptual tasks, including handwriting identification, traffic sign recognition, pedestrian detection, human action recognition, object recognition, scene parsing, and the detection of breast cancer cell mitosis. Several object detection methods were covered in this study along with their benefits and drawbacks. In the future, the system can be enhanced by swapping cutting-edge procedures with a more dependable and straightforward system. This has a number of real-time uses, including manufacturing, space exploration, and criminal investigations.

References

[1] Krizhevsky, A., Sutskever, I. and Hinton, G.E. (2017) ImageNet classification with Deep Convolutional Neural Networks. Communications of the ACM, 60, 84-90. https://doi.org/10.1145/3065386

[2] Akhtar N. & Mian A., 2018. Threat of Adversarial Attacks on Deep Learning in Computer Vision: A Survey. IEEE Access. :14410-14430.

[3] Heba Hakim and Ali Fadhil 2021 J. Phys.: Conf. Ser. 1804 012095

[4] Kim, B. and Lee, J. (2019) A Video-Based Fire Detection Using Deep Learning Models. Applied Sciences, 9, Article No. 2862. https://doi.org/10.3390/app9142862

[5] Li, P., Chen, X. and Shen, S. (2019) Stereo R-CNN Based 3D Object Detection for Autonomous Driving. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, 15-20 June 2019, 7636-7644. https://doi.org/10.1109/CVPR.2019.00783

[6] Zhang, X., Yi, W.-J. and Saniie, J. (2019) Home Surveillance System Using Computer Vision and Convolutional Neural Network. 2019 IEEE International Conference on Electro Information Technology (EIT), Brookings, 20-22 May 2019, 266-270. https://doi.org/10.1109/EIT.2019.8833773

[7] Zhang, R., Shao, Z., Huang, X., Wang, J. and Li, D. (2020) Object Detection in UAV Images via Global Density Fused Convolutional Network. Remote Sensing, 12, Article, No. 3140. https://doi.org/10.3390/rs12193140

[8] Vaigandla, K. K., Thatipamula, S. & Karne, R. K. (2022). Investigation on Unmanned Aerial Vehicle (UAV): An Overview. IRO Journal on Sustainable Wireless Systems, 4(3), 130-148. doi:10.36548/jsws.2022.3.001

[9] Karthik Kumar Vaigandla, SandyaRani Bolla, Radha Krishna Karne, "A Survey on Future Generation Wireless Communications-6G: Requirements, Technologies, Challenges and Applications", International Journal of Advanced Trends in Computer Science and Engineering, Volume 10, No.5, September - October 2021, pp.3067-3076, https://doi.org/10.30534/ijatcse/2021/211052021.

[10] Girshick, R., Donahue, J., Darrell, T. and Malik, J. (2013) Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation. 2014 IEEE Conference on Computer Vision and Pattern Recognition, Columbus, 23-28 June 2014, 580-587. https://doi.org/10.1109/CVPR.2014.81

[11] Ren, J.S. and Wang, Y. (2022) Overview of Object Detection Algorithms Using Convolutional Neural Networks. Journal of Computer and Communications, 10, 115-132, https://doi.org/10.4236/jcc.2022.101006

[12] Hanbin Luo, Chaohua Xiong, Weili Fang, Peter E.D. Love, Bowen Zhang, Xi Ouyang, "Convolutional neural networks: Computer vision-based workforce activity assessment in construction," Automation in Construction 94 (2018) 282–289, https://doi.org/10.1016/j.autcon.2018.06.007

[13] Deepika B, Dorthy Shaji, Sivapriya K, Vendamani N,Mohanambal K, "Object Detection And Tracking Using Deep Learning Algorithms," IJARIIE, Vol-6 Issue-2, 2020

[14] Pratt H., Coenen F., Broadbent D., Harding S. & Zheng Y., 2016. Convolutional Neural Networks for Diabetic Retinopathy. ELSEVIER, Procedia Computer Science. :200 – 205.

[15] Krebs S., Duraisamy B. & Flohr F., 2017. A survey on Leveraging Deep Neural Networks for Object Tracking. Proceedings of the 20th IEEE International Conference on Intelligent Transportation Systems (ITSC). :411-418.

[16] Zhi-Peng F. & Yan-Ning Z., 2014. Survey of Deep Learning in Face Recognition. Proceedings of the IEEE International Conference on Orange Technologies. :5-8.

[17] Fukushima, K and Miyake, S (1982) Neocognitron: A New Algorithm for Pattern Recognition Tolerant of Deformations and Shifts in Position Pattern recognition, 455-469. http://doi.org/10.1016/0031-3203(82)900243

[18] Singh, B. and Davis, L.S. (2018) An Analysis of Scale Invariance in Object Detection Snip. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, 18-23 June 2018, 3578-3587. https://doi.org/10.1109/CVPR.2018.00377

[19] Howard, A.G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., et al . (2017) Mobilenets: Efficient Convolutional Neural Networks for Mobile Vision Applications. https://arxiv.org/abs/1704.04861

[20] 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, 1904-1916. https://doi.org/10.1109/TPAMI.2015.2389824

[21] Krizhevsky, A., Sutskever, I. and Hinton, G.E. (2012) ImageNet Classification with Deep Convolutional Neural Networks. Advances in Neural Information Processing Systems, 25, 1097-1105.

[22] Girshick, R. (2015) Fast R-CNN. Proceedings of the IEEE International Conference on Computer Vision, Santiago, 7-13 December 2015, 1440-1448. https://doi.org/10.1109/ICCV.2015.169

[23] Ren, S., He, K., Girshick, R. and Sun, J. (2017) Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. IEEE Transactions on Pattern Analysis & Machine Intelligence, 39, 1137-1149. https://doi.org/10.1109/TPAMI.2016.2577031

[24] Uijlings, J.R., van de Sande, K.E.A., Gevers, T. and Smeulders, A.W.M. (2013) Selective Search for Object Recognition. International Journal of Computer Vision, 154-171. http://doi.org/10.1007/s11263-013-0620-5
[25] Redmon, J., Divvala, S., Girshick, R. and Farhadi, A. (2015) You Only Look Once: Unified, Real-Time Object Detection. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, 27-30 June 2016, 779-788. https://doi.org/10.1109/CVPR.2016.91
[26] Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.-Y. and Berg, A.C. (2016) SSD: Single Shot Multibox Detector. European Conference on Computer Vision, Amsterdam, 8-16 October 2016, 21-37. https://doi.org/10.1007/978-3-319-46448-0_2
[27] Pei T. & Xiaoyu W., 2016. Object Proposals Detection. Proceedings of the IEEE Conference on Computer and Communication. :445-448.
[28] Zhou X., Gong W., Fu W. & Du F., 2017. Application of Deep Learning in Object Detection. Proceedings of the 16th International Conference on Computer and Information Science (ICIS). :631-634.
[29] Krizhevsky A., Sutskever I. & Hinton G.E., 2012. Imagenet Classification with Deep Convolutional Neural Networks. Advances in Neural Information Processing Systems (NIPS).:1097–1105.
[30] He K., Zhang X., Ren S. & Sun J., 2016. Identity Mappings in Deep Residual Networks. arXiv:1603.05027.
[31] He, K., Gkioxari, G., Dollár, P. and Girshick, R. (2017) Mask R-CNN. 2017 IEEE International Conference on Computer Vision (ICCV), Venice, 22-29 October 2017, 2980-298. https://doi.org/10.1109/ICCV.2017.322
[32] Li, Y., Chen, Y., Wang, N. and Zhang, Z.-X. (2019) Scale-Aware Trident Networks for Object Detection. Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, 27 October-2 November 2019, 6053-6062. https://doi.org/10.1109/ICCV.2019.00615
[33] Cao, J., Cholakkal, H., Anwer, R.M., Khan, F.S., Pang, Y. and Shao, L. D2det: Towards High Quality Object Detection and Instance Segmentation. Proceedings of the IEEE / CVF Conference on Computer Vision and Pattern Recognition, Seattle, 13-19 June 2020, 11482-11491. https://doi.org/10.1109/CVPR42600.2020.01150
[34] Neubeck, A. and Van Gool, L. (2006) Efficient Non-Maximum Suppression. 18th International Conference on Pattern Recognition, Hong Kong, 20-24 August 2006, 850-855. https://doi.org/10.1109/ICPR.2006.479
[35] Sun, P., Zhang, R., Jiang, Y., Kong, T., Xu, C., Zhan, W., et al. (2021) Sparse R-CNN: End-to-End Object Detection with Learnable Proposals. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, 20-25 June 2021, 14449-14458. https://doi.org/10.1109/CVPR46437.2021.01422
[36] Redmon, J. and Farhadi, A. (2017) YOLO 9000: Better, Faster, Stronger. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, 21-26 July 2017, 6517-6525. https://doi.org/10.1109/CVPR.2017.690
[37] Bochkovskiy, A., Wang, C.-Y. and Liao, H.-Y.M. (2020) YOLOv4: Optimal Speed and Accuracy of Object Detection. https://arxiv.org/abs/2004.10934
[38] Lin, T.-Y., Goyal, P., Girshick, R., He, K. and Dollár, P. Focal Loss for Dense Object Detection. Proceedings of the IEEE International Conference on Computer Vision, Venice, 22-29 October 2017, 2999-3007. https://doi.org/10.1109/ICCV.2017.324
[39] Law, H. and Deng, J. (2018) Cornernet: Detecting Objects as Paired Key points. Proceedings of the European Conference on Computer Vision (ECCV), Munich, 8-14 September 2018, 765-781. https://doi.org/10.1007/978-3-030-01264-9_45.