A Review on Object Detection Techniques using Deep Learning

Krutika Dandgaval[1], Swami Patil[2], Hemant Chavan[3], Nikhil Bankar[4], Megha Patil[5]

[1-4]Student, Dept. of Computer Engineering, K. K. W. I. E. E. R, Maharashtra, India
[5]Assistant Professor, Dept. of Computer Engineering, K. K. W. I. E. E. R, Maharashtra, India

---***---

Abstract - *Object Detection is one of the most challenging problems in computer vision. It attempts to find object instances in given images from a large number of categories that are already defined. Deep learning methods have developed to learn feature representations from data and have led to remarkable inventions in the field of object detection. As deep learning has observed rapid growth, more robust tools, which can learn semantic, high-level, deeper features, are proposed to address the issues existing in conventional architectures. Our review starts with an introduction of deep learning algorithms in the family Convolutional Neural Network (CNN). Then we focus on basic image segmentation techniques.*

Key Words: **Computer vision, Object detection, Deep learning, Convolutional Neural Network, Image segmentation.**

1. INTRODUCTION

Object detection has been an active area of research for several years. It aims to determine whether there are any instances of objects from predefined categories in an image. The applications of object detection include computer vision, surveillance, autonomous driving, human-computer interaction, and augmented reality. Image detection algorithms should focus on estimating the locations of objects in images and not only classify the images. There are two types of object detection: First is the detection of specific instances which aims to solve the matching problem by detecting instances of a particular object. And the other one is the detection of broad categories which aims to detect instances of predefined object categories. In recent years, there is a focus on developing a general-purpose detection system.

2. OBJECT DETECTION

The various techniques used for object detection are as follows:

2.1 Convolutional Neural Network (CNN)

Convolutional Networks are intended to process data that appear in the form of multiple arrays. CNN is a type of feed-forward neural network comprised of a set of convolutional and subsampling layers. The general architecture of the Convolutional Network is structured as a sequence of stages. The first few stages are composed of two types of layers: convolutional layers and pooling layers. Units in a convolutional layer are arranged in feature maps, inside which every unit is connected to local patches in the feature maps of the prior layer through a set of weights called a filter bank. Then the outcome of this local weighted sum is passed through a non-linearity such as a ReLU. Every unit in a feature map shares the same filter bank. Different feature maps in a layer utilize different filter banks[13].

An initial feature hierarchy is built with an interleave between convolution and pooling, which can be fine-tuned in a supervised manner by adding various fully connected (FC) layers to adapt to different visual tasks. The final layer with different activation functions [1] is added to get a specific conditional probability for each output neuron, according to the tasks involved. And the entire network can be optimized on an objective function via the stochastic gradient descent (SGD) method. The typical VGG16 has a total of 13 convolutional (Conv) layers, 3 fully connected layers, 3 max-pooling layers, and a softmax classification layer. The convolutional feature maps are generated by convoluting 3*3 filter windows, and feature map resolutions are reduced with 2 stride max-pooling layers. With the trained network, an arbitrary test image of the same size as training samples can be processed. If different sizes are provided, re-scaling or cropping operations may be needed[14].

The common perceptions of simple cells and complex cells in visual neuroscience inspired the convolutional and pooling layers in Convolutional Networks. Convolutional Networks have their origins in the neocognitron46, the architecture of which was somewhat similar but did not have an end-to-end supervised-learning algorithm such as [13] backpropagation.

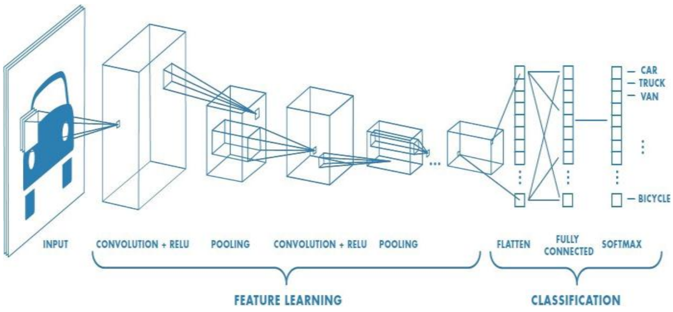

Fig-1: Convolutional Neural Network from https://towardsdatascience.com/

2.2 Region Proposal Network (RPN)

The purpose of an RPN is to propose object proposals[12]. The input to the RPN is images and it outputs a set of rectangular object proposals, including the probability of being an object in each proposal. RPN utilizes a CNN to extract a feature map and slide another Convolutional layer on the map. A rectified linear unit (ReLU) activation function follows the Convolutional layer [11], which provides nonlinearity and increases the rate of convergence [1]. Mapping the features of each sliding window to a vector is done by this Conv, which id followed by Relu. Then this mapping is fed into regression and softmax layers. The coordinates of the multiple bounding boxes and the probability of being an object in each box are predicted by regression and softmax layers respectively.

2.2.1 Region Based Convolutional Neural Network (R-CNN)

R- CNN was proposed by Ross Girshick in 2014. R-CNN can be divided into three stages as follows:

Region proposal generation:
To generate about 2k region proposals for each image, the R-CNN adopts a selective search [4]. To provide more accurate candidate boxes of arbitrary sizes quickly and to reduce the searching space in object detection, the selective search method depends on simple bottom-up grouping and prominent signals [5], [7].

CNN based deep feature extraction:
Every region proposal is warped or cropped into a fixed resolution. The CNN module [1] is utilized to extract a 4096-dimensional feature as the final representation in this stage. A high-level, semantic, and robust feature representation for every region proposal can be obtained due to extensive learning capacity, dominant expressive power, and hierarchical structure of CNNs.

Classification and localization:
Different region proposals are scored on a set of positive regions and background (negative) regions with pre-trained category-specific linear SVMs for multiple classes [14]. Then to produce final bounding boxes for preserved object locations, the scored regions are adjusted with bounding box regression and filtered with a greedy non-maximum suppression (NMS).

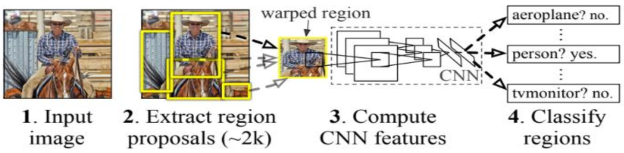

Fig-2: Region Based Convolutional Neural Network from Object Detection with Deep Learning: A Review[14]

2.2.2 Spatial Pyramid Pooling (SPP-Net)

R-CNN prefers to crop or warp each region proposal into the same size because Fully Connected layers must take a fixed-size input. The warping operation may produce unwanted geometric deformity cause the object may exist partly in the cropped region [14]. The recognition accuracy will be reduced due to these content losses or distortions, especially when the scales of objects vary.

To solve this problem, He et al. proposed a new CNN architecture named Spatial Pyramid Pooling (SPP-net) based on the theory of spatial pyramid matching (SPM) [3], [10]. To partition the image into several divisions and aggregating quantized local features into mid-level representations, SPM takes several finer to coarser scales.

Figure 3 represents the architecture of SPP-net for object detection. To project region proposals of arbitrary sizes to fixed-length feature vectors, SPP-net reuses feature maps of the 5th Conv layer (conv5). Feature maps can be reused because they involve the strength of local responses and they also have relationships with their spatial positions [6]. The layer after the final Conv layer is referred to as the spatial pyramid pooling layer (SPP layer).

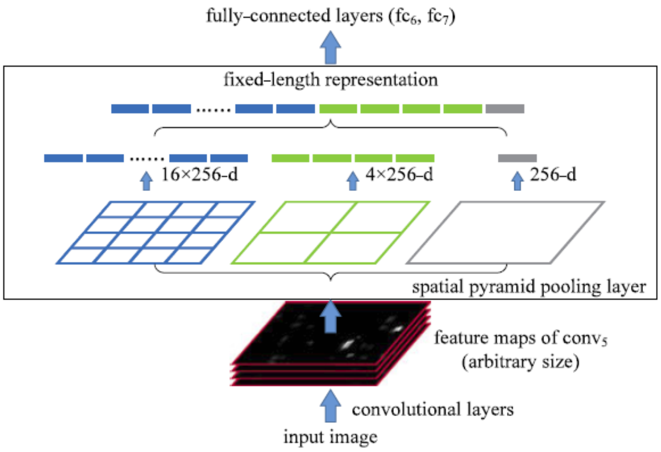

Fig-3: Spatial Pyramid Pooling from Object Detection with Deep Learning: A Review[14]

2.2.3 Fast Region based Convolutional Network (Fast R-CNN)

The schematic architecture of Fast R-CNN, is shown in Figure 4. To extract a features map from the input image, the Fast R-CNN takes precomputed object proposals from RPN and uses CNNs, like RPN. Features bound by object proposals are called a region of interest (RoI) [12]. Precomputed object proposals are overlaid on the feature map in the RoI pooling layer. The extraction of a fixed-size feature vector from each RoI is done by RoI pooling which takes RoIs and applies max-pooling operation. To calculate the location of bounding boxes and classify objects in the boxes, these vectors are fed into Fully Connected layers, followed by two regression and softmax layers.

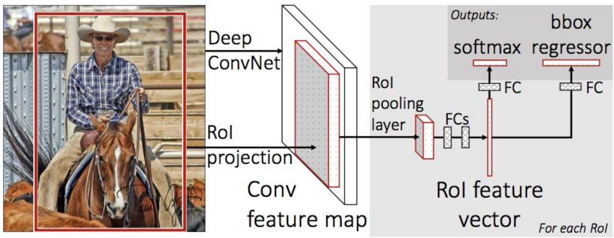

Fig-4: Fast Region based Convolutional Network from Object Detection with Deep Learning: A Review[14]

2.2.4 Faster Region based Convolutional Neural Network (Faster R-CNN)

In Faster R-CNN, computations of the CNN are shared between RPN and Fast R-CNN for feature extraction. The first nine layers of both RPN and Fast R-CNN have the same specifications, and their computations can be shared. Figure 5 shows the architecture of Faster R-CNN.

To fine-tune the parameters of the RPN and Fast R-CNN a four-step training process is used. In the first step, RPN is trained with initial weights, and object proposals are prepared for Fast RCNN. In step two Fast R-CNN is initialized with the trained weights from step one. In the third step, RPN is initialized with the final weights of the previous step and trained again. In the last step, Fast R-CNN takes the object proposals generated in step three and is trained with the initial parameters trained in step three. RPN may produce more than 2,000 object proposals for an image. Thus it leads to costly computations and may decrease the accuracy of object detection [2][8][9]. So, the outputs of RPN are sorted based on the score of its softmax layer. The first 2,000 objects proposals with the highest scores are fed into Fast R-CNN in the second stage of training. For the fourth stage of the training and testing, the methods of Ren et al. (2016) are used. Using which the first 300 object proposals with the highest scores are used to increase the speed of detection.

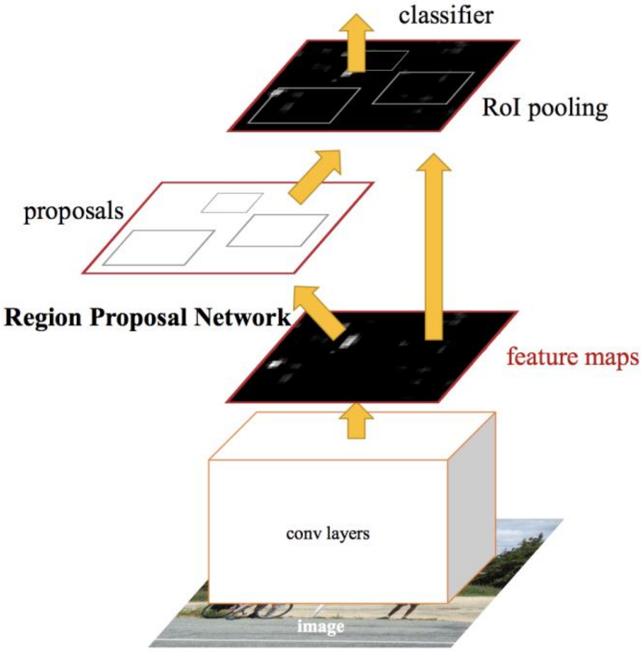

Fig-5: Faster Region based Convolutional Neural Network from https://www.analyticsvidhya.com/

Table -1: Summary

Summary			
Sr. No	**Methods**	**Merits**	**Demerits**
1.	Convolutional Neural Network (CNN)	1.Automatic calculation of damage sensitive features during training process.	1.Difficult to find the optimal size of the sliding Window because the testing images may have various sizes and scales.
2.	Region based Convolutional Neural Network (R-CNN)	1.To provide multiple objects detection and localization. 2.Increased the accuracy of object detection, in comparison with CNN-based methods.	1.Slow process. 2.Costly process.
3.	Spatial Pyramid Pooling (SPP-Net)	1.Increase in the speed of object detection in comparison to the	1. A difficult training process. 2.Limited

		R-CNN.	accuracy.
4.	Fast R-CNN	1.Higher speed than both the R-CNN and SPP-net. 2.Higher accuracy than both the R-CNN and SPP-net.	1.Limited accuracy.
5.	Faster R-CNN	1.Reduces computational costs. 2.Improves the accuracy 3. Provides real-time object detection	1.Very time-consuming. 2. Not skilled in dealing with objects with extreme scales or shapes.

All the above object detection techniques are summarized in Table 1.

3. CONCLUSION

In this paper, we offer a brief review of the Object Detection techniques using Deep Learning like CNN, RPN, SPP-Net, Fast RCNN, Faster RCNN. For every technique in object detection, we describe the existing methodologies in the literature and pinpoint both its advantages and disadvantages.

REFERENCES

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

[2] Fan, Q., Brown, L. & Smith, J. (2016), "A closer look at Faster R-CNN for vehicle detection", in Proceedings of 2016.

[3] F. Perronnin, J. S´anchez, and T. Mensink, "Improving the fisher kernel for large-scale image classification," in ECCV, 2010.

[4] J. R. Uijlings, K. E. Van De Sande, T. Gevers, and A. W. Smeulders, "Selective search for object recognition," Int. J. of Comput. Vision, vol. 104, no. 2, pp. 154–171, 2013.

[5] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in CVPR, 2009.

[6] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," IEEE Trans. Pattern Anal. Mach. Intell., vol. 37, no. 9, pp. 1904–1916, 2015.

[7] 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, pp. 1627–1645, 2010.

[8] Ren, S., He, K., Girshick, R. & Sun, J. (2016), Faster R-CNN: towards real-time object detection with region proposal networks, IEEE Transactions on Pattern Analysis and Machine Intelligence, 39(6), 1137–49. IEEE Intelligent Vehicles Symposium (IV), Gothenburg, Sweden, 19–22 June 2016, 124–29.

[9] R. Girshick, "Fast r-cnn," in ICCV, 2015.

[10] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," in CVPR, 2006.

[11] V. Nair and G. E. Hinton, "Rectified linear units improve restricted boltzmann machines," in ICML, 2010.

[12] Young-Jin Cha*, Wooram Choi, Gahyun Suh & Sadegh Mahmoudkhani , Oral Buy¨ uk¨ ozt¨ urk , 2017. "Autonomous Structural Visual Inspection Using Region-Based Deep Learning for Detecting Multiple Damage Types".

[13] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," Nature, vol. 521, no. 7553, pp. 436–444, 2015.

[14] Zhong-Qiu Zhao, Member, IEEE, Peng Zheng, Shou-tao Xu, and Xindong Wu, Fellow, IEEE ,2017. " Object Detection with Deep Learning: A Review".