International Journal of Innovative Research in Computer Science & Technology (IJIRCST)
ISSN: 2347-5552, Volume-10, Issue-1, January 2022
https://doi.org/10.55524/ijircst.2021.9.6.64
Article ID IRPD12012, Pages 290-294
www.ijircst.org

Object Detection Using Convolutional Neural Networks: A Review

Sushil Bhardwaj

RIMT University, Mandi Gobindgarh, Punjab, India
Correspondence should be addressed to Sushil Bhardwaj; sushilbhardwaj@rimt.ac.in

ABSTRACT- The amount of data on the Internet has increased dramatically as a result of the advent of intelligent devices and social media. Object detection has become a popular international study topic as an important element of image processing. Convolutional Neural Network's (CNN) remarkable capacity with feature learning and transfer learning has piqued attention in the computer vision field in recent years, resulting in a series of significant advancements in object identification. As a result, it's an important study on how to use CNN to improve object detection performance. The article began by explaining the core concept and architecture of CNN. Second, techniques for resolving current difficulties with traditional object detection are examined, with a focus on assessing detection algorithms based on region proposal and regression. Finally, it provided various methods for improving object detecting speed. The study then went on to discuss various publicly available object identification datasets as well as the notion of an assessment criterion. Finally, it went over existing object detection research results and ideas, highlighting significant advancements and outlining future prospects.

KEYWORDS- Convolutional Neural Network, Datasets, Object detection, Region proposal, Regression.

I. INTRODUCTION

The amount of picture data on the Internet has risen fast as a result of the growth of mobile internet and the popularity of various social media, yet humans are unable to handle such large amounts of image data efficiently [1][2,3]. As a result, it is envisaged that this data processing will be carried out automatically with the help of a computer to address large-scale visual issues. With a better knowledge of image processing technologies, complete image interpretation and precise identification of the picture's target item become increasingly crucial. People are concerned not only with simple picture categorization, but also with properly obtaining the semantic category of an item and its placement in the image, therefore object detection technology has gotten a lot of attention. With the theories and methods of image processing and pattern recognition, object detection technology seeks to identify the target objects, determine the semantic categories of these items, and mark the exact position of the target object in the picture [4][5][6].

Using computer technology to automatically recognize items in a real-world application is a difficult challenge. Background complexity, noise disruption, occlusion, low-resolution, size and attitude changes, and other variables will all have a significant impact on object recognition ability [7][8][9]. The traditional object detection technique relied on a hand-crafted feature that was not resistant to changes in light and lacked generalization capabilities. Object detection progress has been sluggish in the PASCAL VOC challenge between 2010 and 2012, with tiny advances gained by creating ensemble systems and using slight variations of conventional approaches. As a result, a number of approaches for improving object detection performance have been presented. As a successful model of deep learning, the convolutional neural network (CNN) has the capacity to learn hierarchical features, and research demonstrates that the feature extracted by CNN has a higher ability of discriminating and generalization than hand-crafted features [10][11][12][13].

In several areas of computer vision, the CNN has had significant success. Hinton and his student Krizhevsky utilized CNN to picture classification in the 2012 ImageNet Large Scale Visual Recognition Challenge (ILSVRC) and obtained top5 error of 15.3 percent vs. 26.2 percent for the conventional technique.

Because of its evident success, CNN became dominant in later computer vision tasks after this turning point. Cheng et al. introduced the R-CNN (Regions with CNN features) technique, which he successfully used in object detection [14][15][16]. According to recent academic and technological advancements, the deep learning technique may achieve greater precision and reduce test time compared to the prior method [17][18][19].

II. DISCUSSION

A. *Convolutional Neural Network*

Feature extraction and categorization has long been a major research focus in the field of computer vision [20]. The retrieved features in traditional image processing jobs are frequently pre-designed features based on statistical regularities or previous information. As a result, it is

unable to correctly and fully reflect the original image data. The gradient descent approach may be used to train parameters in CNN's end-to-end learning model. A well-trained CNN can learn the image's characteristics more thoroughly, and we may consider it a superior black Box for extracting features.

CNN is a type of neural network that expands the notion of receptive field and shared weights, which not only decreases the number of training parameters but also the complexity of the network model [2]. By sharing the weight of the convolutional kernel, the features of each layer are produced from the local region of the preceding layer. These qualities make CNNs better than other neural networks at learning and representing visual features, and they can also preserve translation and scale invariance to a degree.

The initial few layers of a typical CNN are generally alternating convolution and pooling layers, while the latter levels towards the output layer are usually full-connected networks. The forward propagation and BP (Back Propagation) algorithms are primarily used in CNN training to learn layer-connection weights, bias, and other parameters. The training is a supervised learning procedure that uses picture data as input and labels as output to optimize the network parameters and produce an optimized-weight model.

CNN is made up of many functional layer structures [21]. The convolutional layer, pooling layer, and fully-connected layer are all present in a typical CNN. In the course of evolution and enhancement, CNN adds several additional layers, such as the SPP-layer from the SPP-net, the ROI (Region of Interest)-pooling layer from Fast R-CNN, and the Region Proposal Network (RPN) layer from Faster R-CNN. Improved performance can be achieved by modifying the traditional CNN structure based on the unique issues. We'll go through the fundamental network topology of a standard CNN and the BP algorithm in this part.

1) The fundamental structure of CNN

The input layer, convolutional layer, pooling layer, full connect layer, and output layer make up a typical CNN structure, as depicted in Fig. 1.

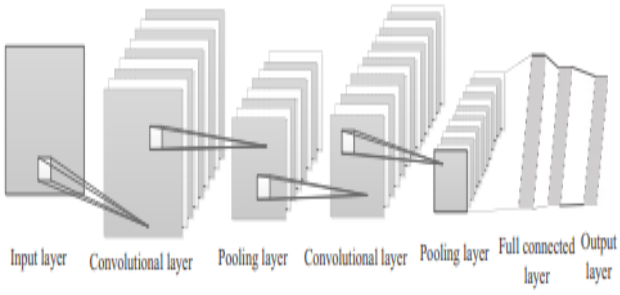

Figure 1: Typical CNN structure [22]

The original picture XIn is generally used as the input to the convolutional neural network. The j-th feature map of the l-th layer is denoted by Xj. We'll pretend that Xj is a feature map in the convolutional layer, and that Xj is created using the following formula:

$$X_j^l = f\left(\sum_{i \in M_j} X_i^{l-1} * W_{ij}^l + b_j^l\right)$$

Here,

W1 is the appropriate weight matrix for the l layer feature map,

* denotes a convolution with particular filters for the l-1 layer's feature maps,

After adding the computed results to the bias bj, nonlinear activation function f(x) procedures such as Rectified Linear Units (ReLU) are used to get Xj.

Following the convolutional layer, the pooling layer, also known as the down-sampling layer, down samples the preceding feature map according to a predetermined rule. Max-pooling, average-pooling, stochastic pooling overlapping pooling, and so on are the particular rules. The function of the pooling layer is divided into two parts:

- Feature map dimensionality reduction.
- Maintain scale invariance. Assume that Xj is a feature map in the pooling layer, and that the pooling process is defined using the following formula:

$$X_j^l = f(\beta_j^i pooling(X_j^{l-1}) + b_j^l)$$

The rule of down-sampling is denoted by pooling (x), and the weight of pooling is denoted by Bj. In general, bias b and the activation function f(x) are not utilized since B is a fixed value. As a result, the formula for pooling operations is:

$$X_j^l = \beta_j^i pooling(X_j^{l-1})$$

The feature maps of pictures are concatenated into a one-dimensional feature vector as input to a fully-connected network in a fully-connected network. By applying a weighted summation to the input and responding with the activation function, the output of the fully-connected layer may be produced, as indicated in the formula:

$$X^l = f(w^l X^{l-1} + b^l)$$

2) Back Propagation

To change the weight parameters of a neural network, the BP (back propagation) method is employed [23]. Convolution kernel parameters, pooling layer weights, full-connected layer weights, and bias parameters are the most important optimization parameters for CNN. The basic idea behind BP is to calculate the partial derivative of the residuals for each layer parameter, learn an association rule between the residuals and the network weights, and then modify the network weight to bring the network output closer to the anticipated value.

The CNN's objective during training is to reduce the network's loss function E(w,b).

MSE (Mean Squared Error), NLL (Negative Log Likelihood), and other loss functions are examples of common loss functions. The formula for calculating MSE and NLL is:

$$MSE(W,b)=\frac{1}{|Y|}\sum_{i=1}^{|Y|}(Y(i)-\bar{Y}(i))^2$$

$$NLL(W,b)=-\sum_{i=1}^{|Y|}\log Y(i)$$

The residuals are transmitted backwards by gradient descent during the training phase, and the trainable parameters of each layer are updated layer by layer in CNN. The strength of BP is controlled by the learning rate n, the weight Wi is updated by a formula, and the bias bi is updated by a formula:

$$W_i = W_i - \eta\frac{\partial E(W,b)}{\partial W_i}$$

$$b_i = b_i - \eta\frac{\partial E(W,b)}{\partial b_i}$$

B. Object Detection Based On CNN

The object detection challenge was added to the ImageNet competition in 2013, with 200 items in 40,000 pictures. The competition's champions, on the other hand, employ hand-crafted features with a mean Average Precision of just 22.581 percent (mAP). R-CNN achieves a substantial improvement in ILSVRC 2014, up to 43.933 percent mAP, thanks to deep learning and a region proposal method. R-CNN came up with the widely utilized CNN-based object identification approach initially.

This section examines and summarizes CNN-based object detection algorithms, and it is separated into four sections: The typical object detection process is introduced in Part 1. The CNN object detection framework is introduced in Part 2 along with a region suggestion. Part 3 explains how to convert a detection problem into a regression problem. Part 4 discusses various approaches for improving object detection performance.

1) Traditional Object Detection Method

The typical object detection framework is separated into four phases, as shown in Fig. 2.

- Creating candidate areas on a given picture using a sliding window,
- Extracting important features from these regions,
- Classifying and identifying the regions using the trained classifier, and
- Reviewing and optimizing the NMS detection findings.

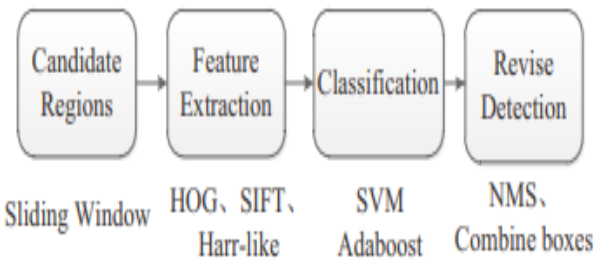

Figure 2: Traditional object detection pipeline

- *Generating Candidate Regions:*

The object's position is determined at this stage. The item, on the other hand, might appear anywhere in the photograph. Because the size and aspect ratio of the item are both unknown, a sliding window technique was used to traverse the whole image with a succession of scale and aspect ratio sliding windows. This comprehensive method covers the majority of the object's probable locations, but there are several apparent drawbacks, such as the enormous complexity of time and the excessive number of redundancy windows. This would have a significant influence on the speed and efficiency of feature extraction.

- *Feature Extraction:*

The design and performance of the classifier are directly influenced by feature extraction (24). However, a robust feature is difficult to develop due to a number of external elements such as item movement, lighting changes, and a complex and changing environment. Scale Invariant Feature Transform (SIFT), Histograms of Oriented Gradients (HOG), and Local Binary Patterns are some of the hand-crafted features popular at this time (LBP).

- *Classification:*

In this stage, SVM or AdaBoost classifiers are commonly employed to classify the extracted features.

- *Revise Detection Results:*

After classification, there are still many redundant windows, therefore it's important to eliminate them and enhance the detection results using Non-Maximum Suppression (NMS) and merging overlapping Bounding boxes.

Traditional object identification has two major flaws: first, the sliding window technique is not purposeful enough, has a high temporal complexity, and has too many redundant windows; and second, the hand-crafted features are not robust enough for a wide range of modifications.

2) Object Detection Based On Regions with CNN features(R-CNN)

To tackle the problem of redundant windows, region suggestion offers a suitable option, which is to determine the probable placement of items in the image ahead of time. The region concept makes effective use of picture information such as colour, edge, and texture to guarantee that the system has a high recall even when there are fewer windows (thousands or even hundreds). It significantly reduces the time-complexity of subsequent activities, and the quality of region proposals is far greater than with sliding windows. The most frequent region suggestions algorithm uses selective Search, edge Boxes, and other techniques.

The rest of the work with the candidate areas is essentially the picture classification task for the candidate regions (feature extraction plus classification). AlexNet's performance in picture categorization demonstrates the CNN's capacity to extract features. So, in 2014, Ross B. Girshick advocated using CNN instead of traditional sliding windows with hand-crafted features, and created the R-CNN detection framework. It was a huge

achievement in object detection, and it opened the way to CNN-based object detection.

The initial module in R-CNN is to create category-independent region suggestions. A big CNN retrieves a fixed-length feature vector from each area in the second module. Specific linear SVMs would divide the region into the object and background in the last module. To enhance localization accuracy, the author used a linear regression model to alter the coordinates of detection boxes, which was inspired by DPM's bounding box regression. These enhancements are clearly successful, as R-CNN achieves 53.7 percent mAP in VOC2012 datasets against 35.1 percent mAP with DPM HSC.

During testing, however, the R-CNN generated two thousand candidate windows each image. Each region would use CNN to extract features, making feature calculation time-consuming (50s per image). Another issue is that most CNNs require a set input size, therefore R-CNN crops or warps the input image to suit the predetermined size. However, picture information such as aspect ratio and size will be lost. To address the shortcomings of R-CNN, He Kaiming et al. introduced the SPP-Net. SPP-net executes the convolutional layer only once on the whole picture to generate feature maps. Because of the shared calculation, the test time is reduced by 24-64 times compared to R-CNN. Although the proposed areas vary in size, the full-connected layer requires a constant input size. As a result, He Kaiming suggested the SPP-layer (Spatial Pyramid Pooling). The position of the candidate windows generated by the selective search relative to the original picture is mapped to the final convolutional layer features maps, and the SPP-layer is placed below the last convolutional layer. After that, a multi-level spatial pyramid pools the patch features and creates a fixed-length feature vector representation for each window. The fully linked network receives these representations. Because SPP-Net can maintain both global and local image information, it has a greater mAP than R-CNN.

C. Datasets and Evaluation

For training a Deep Neural Network, a significant quantity of labelled data is required; currently, the most often used object identification datasets are ImageNet, PASCAL VOC, and MS COCO (25). PASCAL VOC is an image labelling and assessment system that is industry standard. The PASCAL VOC image dataset has 20 categories and includes a high-quality, fully labelled image that is ideal for evaluating algorithm performance. Because of a specialized boundary labelling training set, ImageNet provides an essential source of data for object detection. The annotation contains categories, geographical information, and a semantic text description of the image. COCO is funded by Microsoft. The COCO dataset, which is open-source, also contributes to the advancement of object detection.

The object detection results can be represented as follows: The detector examines the input picture I to determine the bounding box B of each item, as well as the matching category label c and confidence level f. The ground truth boxes Bg are used to evaluate multi-Objects detection in the same picture as independent detection results. The projected bounding box is considered accurate if it meets the following formula.

$$a = \frac{area(B \bigcap B_g)}{area(B \bigcup B_g)} \geq a_0$$

The overlap rate of the ground truth box and the object window predicted by the detector is represented by the evaluation parameter Intersection Over Union (IOU). The value of a0 is a previously determined threshold value of 0.5.

To compute average precision (AP), we'll need to understand a few concepts, which are listed in Table 1.

Table 1: Concepts about binary classification

Concepts	Explanation
True Positive (TP)	The number of samples which the true is predict as positive
True Negative (TN)	The number of samples which the true is predict as negative
False Positive (FP)	The number of samples which the false is predict as positive
False Negative (FN)	The number of samples which the false is predict as negative

III. CONCLUSION

The study focuses on object detection using CNN, and it introduces the structure of CNN, the framework for object detection using CNN, and strategies for increasing detection performance. Because CNN is so good at extracting features, it can compensate for the flaws in hand-crafted features. CNN also outperforms traditional approaches in terms of real-time, accuracy, and flexibility, but it still has a lot of potential for development. Improving the structure of CNN can decrease feature information loss, while properly leveraging object and context relationships and creating fuzzy inference can help the computer deal with issues like occlusion and poor resolution. The main aspect in future study will be to improve intelligence and the practicability of object detection using CNN.

REFERENCES

[1]. Gupta H, Varshney H, Sharma TK, Pachauri N, Verma OP. Comparative performance analysis of quantum machine learning with deep learning for diabetes prediction. Complex Intell Syst. 2021;

[2]. Ouyang W, Zeng X, Wang X, Qiu S, Luo P, Tian Y, et al. DeepID-Net: Object Detection with Deformable Part Based Convolutional Neural Networks. IEEE Trans Pattern Anal Mach Intell. 2017;

[3]. Jung D, Son JW, Kim SJ. Shot category detection based on object detection using convolutional neural networks. In: International Conference on Advanced Communication Technology, ICACT. 2018.

[4]. Kishore N, Singh S. Torque ripples control and speed regulation of Permanent magnet Brushless dc Motor Drive using Artificial Neural Network. In: 2014 Recent Advances in Engineering and Computational Sciences, RAECS 2014. 2014.

[5]. Bakker EM. Image and video retrieval : Second International Conference, CIVR 2003, Urbana-Champaign,

IL, USA, July 24-25 2003 : proceedings. Lecture notes in computer science. 2003.
[6]. Goel AR, Ranjan A, Wajid M. VLSI architecture and implementation of statistical multiplexer. In: Proceedings of the International Conference on Innovative Applications of Computational Intelligence on Power, Energy and Controls with Their Impact on Humanity, CIPECH 2014. 2014.
[7]. Guo W, Yang W, Zhang H, Hua G. Geospatial object detection in high resolution satellite images based on multi-scale convolutional neural network. Remote Sens. 2018;
[8]. Khanna R, Verma S, Biswas R, Singh JB. Implementation of branch delay in Superscalar processors by reducing branch penalties. In: 2010 IEEE 2nd International Advance Computing Conference, IACC 2010. 2010.
[9]. Gupta H, Kumar S, Yadav D, Verma OP, Sharma TK, Ahn CW, et al. Data analytics and mathematical modeling for simulating the dynamics of COVID-19 epidemic—a case study of India. Electron. 2021;
[10].Jain N, Awasthi Y, Jain RK. Ubiquitous sensor based intelligent system for net houses. In: Proceedings - IEEE 2021 International Conference on Computing, Communication, and Intelligent Systems, ICCCIS 2021. 2021.
[11].Sood R, Kalia M. Cloudbank: A secure anonymous banking cloud. In: Communications in Computer and Information Science. 2010.
[12].Bala L, K. Vatsa A. Quality based Bottom-up-Detection and Prevention Techniques for DDOS in MANET. Int J Comput Appl. 2012;
[13].Gupta P, Tyagi N. An approach towards big data - A review. In: International Conference on Computing, Communication and Automation, ICCCA 2015. 2015.
[14].Cheng B, Wei Y, Shi H, Feris R, Xiong J, Huang T. Revisiting RCNN: On awakening the classification power of faster RCNN. In: Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). 2018.
[15].Gulista K, Kumar GK, Rahul R. RE2R—Reliable energy efficient routing for UWSNs. In: Advances in Intelligent Systems and Computing. 2016.
[16].Kushawaha JS, Misra BK. Improved imposition of displacement boundary conditions in element free Galerkin method using penalty method. Int J Comput Aided Eng Technol. 2016;
[17].Khan G, Gola KK, Ali W. Energy Efficient Routing Algorithm for UWSN - A Clustering Approach. In: Proceedings - 2015 2nd IEEE International Conference on Advances in Computing and Communication Engineering, ICACCE 2015. 2015.
[18].Sharma R, Goyal AK, Dwivedi RK. A review of soft classification approaches on satellite image and accuracy assessment. In: Advances in Intelligent Systems and Computing. 2016.
[19].Saleem A, Agarwal AK. Analysis and design of secure web services. In: Advances in Intelligent Systems and Computing. 2016.
[20].Zhang Q, Wan C, Han W, Bian S. Towards a fast and accurate road object detection algorithm based on convolutional neural networks. J Electron Imaging. 2018;
[21].Gu J, Wang Z, Kuen J, Ma L, Shahroudy A, Shuai B, et al. Recent advances in convolutional neural networks. Pattern Recognit. 2018;
[22].Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification with deep convolutional neural networks. Commun ACM. 2017;
[23].Nagarajan S, Perumal K. A deep neural network for information extraction from web pages. In: IEEE International Conference on Power, Control, Signals and Instrumentation Engineering, ICPCSI 2017. 2018.
[24].Lindeberg T. Scale Invariant Feature Transform. Scholarpedia. 2012;
[25].Ren S, He K, Girshick R, Sun J. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. IEEE Trans Pattern Anal Mach Intell. 2017;