International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 06 Issue: 05 | May - 2022 Impact Factor: 7.185 ISSN: 2582-3930

CONVOLUTIONAL NEURAL NETWORK STRATEGIES FOR REAL-TIME OBJECT DETECTION

Prof. Dr. Sonali Kadam[1], Aastha Sharma[2], Anjali Bari[3], Bhumika Trivedi[4], Amruta Shinde[5]

[1]*Professor, Department of Computer Engineering, Bharati Vidyapeeth's College of Engineering for Women, Pune*

[2,3,4,5]*Student, Department of Computer Engineering, Bharati Vidyapeeth's College of Engineering for Women, Pune*

Abstract - Segmentation, extraction of features, and object recognition from picture data are all areas where computer vision excels. Object detection is attracting a lot of attention from a variety of industries, including healthcare, traffic monitoring, surveillance, robotics, and so on. Because of its employment in sensitive fields, the capacity to identify the item more accurately is a key element. Recent improvements of \sdeep learning approaches provide appealing performance to fine-grained picture classification which tries to identify \ssubordinate-level categories. This job is exceedingly tough because of strong intra-class and minimal inter-class variation. Convolutional Neural Networks have recently endeavored to deliver a greater degree of efficiency and accuracy in all of the sectors where they have been used, the most prominent of which being Object Detection, Digit, and Image Recognition. It uses a well-defined set of procedures to follow, including Backpropagation, Convolutional Layers, Feature Formation, and Pooling. In addition, this post will look at how to leverage several frameworks and tools that employ the CNN model.

Key Words: Convolutional Neural Network, Fast R-CNN, Activation map, YOLO, Feature detector, Single Shot Detector.

1. INTRODUCTION

Although the human eye is capable of instantaneously and exactly distinguishing a given visual, including its content, position, and nearby visuals, human-made, systems using computer vision are very poor in accuracy and speed. Any developments in this subject that lead to increased efficiency and performance might open the way for the development of more intelligent systems, similar to humans. These developments, in turn, would make human existence easier by allowing humans to execute activities with little to no conscious thinking. Driving an automobile outfitted with computer vision-enabled assistive technology, for example, might forecast and report a driving collision before the incidence, even if the driver is unaware of their activities. Computer vision, as well as to object identification, are significant topics of machine learning that are likely to help unleash the potential of general-responsive robotic systems in the future. Object recognition is a method for detecting meaningful things in digital photos and movies. Self-driving automobiles are one of their real-time uses. Our objective, in this case, is to recognize many things in one picture. The automobile, motorbike, and pedestrian are the most prevalent objects to identify in this program. We utilize Object Localization to detect things in images and must locate many objects in real-time systems. There are several approaches for detecting objects, which may be divided into two groups. The first is algorithms relying on classifications. This category includes CNN and RNN. In this case, we must identify the regions of interest in the image and categorize them using a Convolutional Neural Network. This technique is exceedingly slow since we have to execute a forecast for each specified location. The following category is based on regressions. The task has been greatly simplified thanks to neural networks. All models, from Fast RCNN neural networks to Faster R-CNN, have played a vital role in the context of computer vision. This work focuses on the categorization and detection of single and multi-class items. YOLO plays a role here since there is little need to choose the regions in the image.

YOLO, on the other hand, anticipates the classes and grayscale of numerous objects in a single image using only a single neural network. When compared to other categorization methods, the YOLO algorithm is the fastest. Our algorithm processes 45 frames per second in real-time. In the background, the YOLO method computes localization mistakes but predicts fewer false positives.

2. Literature Review

[5]"Real-Time Object Detection and Tracking", this study discusses the growing influence of object detection in security systems. They gave a literature review of several methods for implementing object detection and tracking in this paper. Absolute Differences, Census Transform Method, Feature-Based Method, Kanade-Lucas Technique, and Mean Shift Method are some of the methodologies discussed in this paper. Considering numerous simulation results for several approaches, the authors of this research propose the Kanade-Lucas algorithm since it is the fastest and consumes the least amount of memory. It is also the most accurate, with the fewest implementation complications.

In [8] "Real-Time Object Detection with Yolo", by Geethapriya. S, N. Duraimurugan, S.P. Chokkalingam. The YOLO: You Only Look Once Algorithm is the main focus of

this work. They provided a quick overview of the YOLO algorithm's operation and discussed processes such as bounding box prediction and accuracy enhancement. YOLO performs object detection as a regression problem and returns the probability for each class of the detected photos. To detect objects in real-time, the YOLO method leverages convolutional neural networks (CNN). To identify objects, the approach requires just one forward iteration through a neural network.

[10] "YOLOv2 based Real-Time Object Detection", by Sakshi Gupta, Dr. T. Uma Devi. The YOLOv2 technique is proposed in this work for object detection in images with localization and video recordings. The primary goal of this study is to detect things in real-time, i.e. live detection, utilizing a camera and video recordings. COCO, a dataset with 80 classes, was employed in this work. Using the YOLOv2 model, it is simple to recognize items with grids and boundary prediction, and it also aids in predicting extremely small objects or objects that are far away in the image. Darknet makes it easier to detect moving objects in video recordings and generates.avi files containing detections.

3. Convolutional Neural Network-

Deep learning is a subset of machine learning in which the neural network layers are expanded. This method has produced astounding outcomes in every sector. A basic neural network consists of three layers: one input layer, a second concealed layer, and a third layer where The weights have been determined in the second layer, and the output is the final step. The inclusion of it is termed a deep convolutional neural network if it has more layers. Deep learning algorithms have several levels of representation. The non-linear yet simple representation is achieved through these modules. These modules raise the level of presentation to a higher scale. Deep learning models master more complicated functions as a result of this transition.

A convolutional network is perhaps the most widely known in the field of computer vision; it has been in need for a lot longer and has gained increasing attention in recent years as hardware advancements have equipped machinery with much more computing capabilities; convolutional networks are moving towards deep learning, which provides better results. Fukushima first introduced convolutional neural networks (CNN) in 1998, and they have wide applications in activity recognition, sentence classification, text recognition, face recognition, object detection and localization, image characterization, and so on. They are composed of neurons, each of which has a learnable weight and bias. It has an input and output layer, and many hidden layers, the latter of which includes a convolutional layer, a pooling layer, a fully connected layer (FC), and numerous normalizing layers. To integrate two sets of information, the convolutional layer uses a convolution procedure. It stimulates the response of a single neuron to visual input. The pooling layer is being used to decrease dimensionality by connecting the output of one layer's neuron cluster with a single neuron.

The recurrent convolutional network, as well as the fast recurrent convolutional network, are the most popular and widely used deep learning models for object identification. These neural networks each have their own set of benefits, which are detailed in the subsections below.

4. CNN ARCHITECTURE-

1. **Convolutional layer-** In the receptive field, this layer is the basic building element of a convolutional neural network, which defines the outcome of linked inputs. Kernels are convolved throughout the length and girth of the datasets, estimating the dot product in between intake and filter values, and so constructing a two-dimensional activation map of that filter. The fundamental benefit of this type of map would be that it preserves all of an image's distinctive qualities while also minimizing the quantity of data that must be processed. The feature detector, which is essentially a collection of parameters with which the system is compatible, is the matrix in which the information is convolved. Using varying parameters of the feature detector, multiple versions of the picture are created. In order to ensure low error in each layer, the network is additionally trained by backpropagation. The entries on identical points in the dataset and feature map, i.e. values with value 1 or even more than 1, are maintained in convolution, while the rest are deleted. The picture data matrix is examined 3*3 at a time. The size of the feature detector changes depending on the CNN type.

2. **Pooling Layer-** Pooling is a crucial step in further decreasing the dimensionality of the activation map, preserving just the most relevant properties while diminishing spatial invariance. As a result, the range of learnable characteristics for the model is reduced. Its major goal is to reduce the parameters and operations in the model by shrinking the size of the representation. It accelerates computations along with preventing overfitting.The max-pooling tier is the most frequent type of pooling layer.
   Max Pooling- Pooling that chooses the maximum elements from the area of the map spanned by the filter is known as max pooling. As a result, the output of the max-pooling layer will be a feature map with some of the most significant characteristics of the preceding feature map.

3. **Fully connected layer-** FC layers are conventional deep NN layers that aim to create predictions from activations for classification or regression. It works

on the same principle as a Multi-layer Perceptron neural system (MLP). This is the outermost layer that the neural network receives. Matrixes are usually flattened before being sent to the neurons. Because there are so many hidden layers with varied weights for each neuron's output, it's difficult to follow the data after this stage. This is where all of the data reasoning and computing takes place. This layer obtains the complete connections to each engagement in the preceding layer, and also the activations are computed using mathematical operations with a bias offset.

5. Faster R-CNN

It is a single-developer object detector created by Ross Girshick, a senior Microsoft and Facebook AI researcher. R-CNN has a number of difficulties that Fast R-CNN solves. One benefit of it over R-CNN, as its name implies, is its speed. Faster R-CNN is made up of two channels: a region proposal network that generates areas of interest, and a network that uses these suggestions to detect the object. As a result, in Faster R-CNN, the same network is used to generate region proposals and identify objects. The concept of anchor boxes was presented.

The R-CNN is made up of three primary parts.

1. By using the Selective Search method, the first module creates 2,000 region recommendations.
2. The next module retrieves a feature vector of height 4,096 out of each area proposal after it has been shrunk to a set pre-defined size.
3. The last module employs a pre-trained SVM algorithm to categorize the region suggestion as either background or object.

Drawbacks- The R-CNN model has a number of flaws:

- It's a multi-stage approach with each level functioning independently. As a result, it can't be taught from beginning to end.
- It saves the extracted features from the pre-trained CNN to disc so that the SVMs may be trained later. Hundreds of gigabytes of storage are required.
- For producing region recommendations, R-CNN uses the Selective Search method, which takes a long time. Furthermore, this approach is not adaptable to the detection problem.
- The CNN fed each area suggestion separately for feature extraction. R-CNN cannot be run in real-time as a result of this.

6. YOLO (You Only Look Once)

YOLO is a new method for detecting numerous items in a picture in real-time and creating bounding boxes around them. The picture is only sent through the CNN algorithm once to obtain the output, hence the name. Despite being identical to R-CNN, YOLO is significantly quicker than Faster R-CNN because of its simplified design. YOLO, unlike Faster R-CNN, can classify and perform bounding box regression simultaneously. The position of items may be predicted using YOLO, the class label including them. YOLO addresses object identification as a regression issue by spatially isolating bounding boxes and their corresponding class probabilities, which are forecasted using a single neural network. This is a complete departure from the traditional CNN pipeline.

[1] The design of YOLO is comparable to that of a standard convolutional neural network, which was inspired by the GoogLeNet image categorization model. The initial layer of the network extracts the characteristics of the picture, and the fully connected layers anticipate the output probabilities and coordinates. The whole YOLO network model was developed using 24 convolutional layers, two fully connected layers, 1x1 reduction layers, and 3x3 convolutional layers.

Working of YOLO-

[2]Step 1: To begin, split a picture into grid cells. The image has been divided into grids of 7x7 matrices in this example. Depending on the image's complexity, it will divide it into any number of grids.

Step 2- After the picture has been separated, each grid cell undergoes classification and localization. If an item is detected, the probability of each grid vector is shown.

The dimension of the enclosing box and class are the results of this.

Step 3: Thresholding is now done, and the value grid cells with the highest probability are chosen. This stage results in the elimination of bounding boxes that don't have an item or have a confidence score less than 0.35.

Step 4- The YOLOv2 algorithm employs Anchor Boxes, which identify and locate items inside a single grid cell. Finally, for ultimate detection, Non-max suppression employs Intersection over Union.

7. Single Shot Detector

R-CNN was created to identify, localize, and classify objects, with the result being a set of bounding boxes. The training, on the other hand, is sluggish and involves several periods. R-CNN is slower than a single-shot detector. SSD, as the name implies, uses only one shot to recognize multiple objects in an image, implying that the convolutional network will only run once and predict the feature map. It also makes use of anchor boxes, which come in a variety of aspect ratios and scales. The network integrates numerous feature map predictions with different resolutions to anticipate variable-size objects.

SSD Architecture-

The SSD model employs a feedforward convolutional network to generate a network of bounding boxes and the

existence of object class instances within those boxes, which is then suppressed via non-max suppression. To produce detection, an auxiliary structure is developed on top of basic layers.

8. CONCLUSIONS

We looked at deep learning-based object identification algorithms in this paper. Deep learning-based object detection systems have made incredible progress in recent years, thus it will continue to be a hot topic of research. It's important to remember that every object detector model has two parts: one for creating the filter and the other for determining how that filter will perform. In the future, we can construct a CNN model by taking into account one or more existing models or by modifying existing models to improve object detection difficulties. We can also test the system to see if it can reduce the number of false positives by preprocessing the photos and utilizing movies.

REFERENCES

1. J. Redmon, S. Divvala, R. Girshick, and A. Farhadi,"You only look once: Unified, real-time objectdetection," 2016, doi: 10.1109/CVPR.2016.91.
2. Geethapriya. S, et al. "Real-Time Object Detection with Yolo" proceedings of the International Journal of Engineering and Advanced Technology (IJEAT) inFeb 2019
3. Domink Scherer, Andreas Muller, Sven Behnke: Evaluation of Pooling Operations in Convolutional Architecture for Object Recognition(2010). Available at: http://www.ais.uni-bonn.de
4. Wei Liu1(B) , Dragomir Anguelov2, Dumitru Erhan3, Christian Szegedy3, Scott Reed4, Cheng-Yang Fu1, and Alexander C. Berg: SSD: Single Shot MultiBox Detector
5. H. Naeem, J. Ahmad and M. Tayyab, "Real-time object detection and tracking," INMIC, 2013, pp. 148-153, doi: 10.1109/INMIC.2013.6731341.
6. "A Neural Algorithm of Artistic Style" by Leon A. Gatys, Alexander S. Ecker, Matthias Bethge.
7. Zhiqiang, W., Jun, L.: A review of object detection based on convolutional neural network. In: 2017 36th Chinese Control Conference (CCC), pp. 11104–11109 (2017)
8. Real-Time Object Detection with Yolo Geethapriya. S, N. Duraimurugan, S.P. ChokkalingamInternational Journal of Engineering and Advanced Technology (IJEAT) ISSN: 2249 – 8958, Volume-8, Issue-3S, February 2019
9. Domink Scherer, Andreas Muller, Sven Behnke: Evaluation of Pooling Operations in Convolutional Architecture for Object Recognition(2010).
10. YOLOv2 based Real Time Object Detection Sakshi Gupta [1] , Dr. T. Uma Devi