International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 07 Issue: 02 | Feb 2020 www.irjet.net p-ISSN: 2395-0072

IDENTIFICATION OF SCENE IMAGES USING CONVOLUTIONAL NEURAL NETWORKS - A SURVEY

Rachita[1], Kavitha A S[2], Zehra Ali[3], Thejaswini Ram[4], Yuktha Lakshmi A B[5]

[1,3,4,5]Student, Dept. of Information Science and Engineering, Dayananda Sagar College of Engineering, Bangalore, Karnataka, India
[2]Assistant Professor, Dept. of Information Science and Engineering, Dayananda Sagar College of Engineering, Bangalore, Karnataka, India

---

Abstract - *A real-world image identification system that detects and describes features in image and video data is an emerging subject in computer vision. These systems develop computer vision approach for the classification of scene images. This paper mainly focuses on the application of deep learning in object identification. One important concept of deep learning is Convolutional Neural Networks (CNN). It is now possible to accurately identify and classify objects with the advancements in neural networks. Then we focus on various object detection algorithms that have been developed.*

***Key Words*: convolutional neural networks, deep learning, computer vision, object detection, YOLO, R-CNN, Fast R-CNN, Faster R-CNN**

1. INTRODUCTION

In recent years, a number of research areas have produced good results with the rapid development of deep learning, and followed by the continuous improvement of convolutional neural networks, computer vision has reached a new level. Object Identification is an important field in computer vision, and convolutional neural networks has made great progress in object identification and detection. AlexNet [1] is a convolutional neural network, designed by Alex Krizhevsky. Advancement and development started when Alex won the ImageNet Large Scale Visual Recognition Challenge in 2012. Since AlexNet, the architecture of convolutional neural networks is improving by every passing year [2].

Identifying scene images is complex because of the different illumination conditions, viewpoints and scales. For this purpose, its better to use classifier systems based on features extracted from convolutional neural networks.

In this paper, we will see the existing systems and algorithms and their drawbacks for object identification and the proposed system(convolutional neural networks). Further, we will summarize some of the object detection algorithms related to deep learning.

Conventional Object detection models have three stages [3] starting from the first stage in which instructive regions are selected by scanning the image using a multi-scale sliding window. This technique has various disadvantages. Firstly this technique is high-priced and it produces unessential windows.

Next stage is Feature Extraction which is used for extraction of derived values or features. SIFT(Scale-invariant Feature Transform) [4] is a feature detection algorithm used in computer vision. HOG(Histogram of Oriented Gradients) [5] is feature descriptor for object detection. Haar-like features, proposed by Alfred Haar in 1909, are popularly used in face detection algorithms. Its mainly used in various organisations, educational institutions, surveillance etc, to verify a person's identity. Nonetheless, due to the complexity of identifying objects because of different illumination conditions, viewpoints and scales, it is difficult to rely on the traditional methods as described above and to build feature descriptors to classify objects efficiently.

Third method is classification. This includes the need of creating a classifier to differentiate a desired object from rest of the objects. Support vector machines(SVM) [6] are supervised machine learning algorithm mainly used in classification and regression problems. It uses the labelled training data set and creates a hyperplane that separates the plane into two parts that categorizes new data. AdaBoost, is also known as Adaptive Boosting is a machine learning meta-algorithm. It creates a strong classifier by modifying numerous weak classifiers. Another popular method used is Deformable Part-based Model(DPM) which is the most efficient model amongst all models mentioned above. All the above mentioned methods and feature descriptors are not efficient. So the proposed model used is convolutional neural networks(CNN).

2. CONVOLUTIONAL NEURAL NETWORKS

The massive complexity of object recognition means that our problem can not defined by ImageNet [1] which is a very large dataset. So our model should have lots of previous knowledge to recompense for all the missing data. Convolutional neural network is an example of such a model.

Convolutional neural network is also known as CNN or ConvNet and it is a class of deep neural networks. CNNs are made with learnable weights and biases. Each neuron receives multiple inputs, takes over them a weighted sum, passes it through an activation function and responds with an output. CNNs comparatively use little pre-processing as compared to other image classification algorithms. CNNs are used to extract features, learn and classify them.

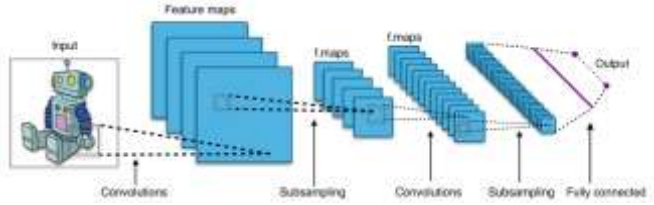

Fig. 1 CNN architecture [7]

CNN is composed of convolutional layer, pooling layer and fully connected layer. Convolutional layer is the first layer where features are extracted from the given input image. Convolution preserves the relation between pixels by using small squares of input data to learn image features. It requires two inputs such as matrix of images and a kernel or filter. Let us consider a 5x5 image matrix whose image pixel values are 0,1 and a 3x3 filter matrix. Convolution of 5x5 image matrix when multiplied with 3x3 filter matrix results in feature map as output. Strides decide the number of pixels that are shifted over the input matrix. If stride is 1, we move 1 pixel at a time. If stride is 2, we move 2 pixels at a time

and so on. If the filter does not fit perfectly into the input image, we can use two methods, the picture is padded with zeros so that it fits and part of the image where the filter did not fit is dropped. This is also known as valid padding as it keeps only the valid part of the image. ReLu(Rectified Linear Unit for a non-linear operation) is used to introduce non-linearity in our convolutional neural networks. (1) is the output of ReLu [8].

$$f(x) = max(0,x) \quad (1)$$

For negative values, the function is zero and for positive values, the function grows linearly. Section of pooling layers would lower the number of parameters when images are too large. Spatial pooling also known as subsampling or down sampling reduces each map's dimensionality but retains important information. Three types of spatial pooling are max pooling, average pooling and sum pooling. Fully connected layer is also known as FC layer. The last few layers in the network form fully connected layers. The output from the convolutional and pooling layers is flattened and then fed into this layer.

3. OBJECT DETECTION ALGORITHMS

3.1 R-CNN:

This method was proposed by Ross Girshick in 2014. He proposed a method which extracted 2000 regions from the image and named them as region proposals. There is no need to classify a huge number of regions as selective search algorithm is used to extract just 2000 regions. A 4096-dimensional feature vector is produced as an output using 2000 candidate region proposals that are distorted into a square and fed into a convolutional neural network. The algorithm predicts presence of an object within the region proposals. The algorithm also predicts offset values which are four in number that increases the precision of the bounding box. R-CNN has some problems. As we need to classify 2000 region proposals per image, it takes a lot of time to train the network. Its real time implementation is not possible as it takes 47 seconds per image approximately. Lastly selective search algorithm is a fixed algorithm. Due to this generation of bad candidate region proposals can take place.

3.2 Fast R-CNN:

Keeping in mind the disadvantages of R-CNN, a faster object detection algorithm was built which is known as Fast R-CNN. This approach differs from R-CNN by the fact that rather than feeding the region proposals, input image is fed to the CNN to create a convolutional feature map. We need not feed 2000 region proposals to the convolutional neural network every

time. Therefore, fast R-CNN is faster than R-CNN. Region proposals becomes an obstruction in Fast R-CNN, therefore affecting its performance.

3.3Faster R-CNN:

Fast R-CNN and R-CNN use selective search algorithm to create region proposals. The selective search algorithm is very slow and time consuming. So, to get rid of it another algorithm was proposed named as Faster R-CNN which lets the network to search for the region proposals. An RoI pooling layer is used to reshape the region proposals that were predicted. Then it helps in classifying the image within the proposed region and it is also used to find the offset values for bounding boxes. Faster R-CNN can be used for real-time object detection.

3.3 YOLO (You Only Look Once):

Redmon et al. [9] proposed an object detection algorithm. This algorithm is different from the above mentioned algorithms. Yolo is having three models till now Yolo v1, Yolo v2(YOLO90000 and Yolo v3. YOLO is different from region based algorithm. Rather than considering the complete image, we consider the parts that have high probabilities of finding the object. How YOLO works is that the input image is split into an SXS grid. We consider m bounding boxes within each grid. The network issues a class probability and offset values for each of the bounding box, we consider the hie bounding boxes that have a class probability above a threshold value that are used for object location within the image. YOLO is quite faster as compared to other object detection algorithms.

4. CONCLUSION

This paper elucidates the importance of Deep Learning methods like Convolutional Neural Networks over the traditional algorithms and methods. Detailed explanation of CNN along with various other object detection algorithms are expressed in this paper.

REFERENCES

[1] A. Krizhevsky, I. Sutskever, and G. Hinton. ImageNet classification with deep convolutional neural networks. In NIPS,2012.

[2] Xinyi Zhou, Wei Gong, WenLong Fu, Fengtong Du: Application of deep learning in object detection, 2017 IEEE/ACIS 16th International Conference on Computer and Information Science (ICIS)

[3] Zhong-Qiu Zhao, Member, IEEE, Peng Zheng, Shou-tao

Xu, and Xindong Wu, Fellow, IEEE: Object Detection with Deep Learning: A Review, April 2019

[4] D. G. Lowe, "Distinctive image features from scale-invariant key-points." Int. J. of Comput. Vision, vol. 60, no. 2, pp. 91–110, 2004. [5] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in CVPR, 2005.

[6] C. Cortes and V. Vapnik, "Support vector machine," Machine Learning, vol. 20, no. 3, pp. 273–297, 1995.

[7] A. Gulli and S. Pal, Deep Learning with Keras, Birmingham: Packt, 2017.

[8] Reagan L.Galvez, Elmer P. Dadios, Argel A. Bandala, Ryan Rhay P. Vicerra, Jose Martin Z.Maningo: Object Detection Using Convolutional Neural Networks , Proceedings of TENCON 2018 - 2018 IEEE Region 10 Conference (Jeju, Korea, 28-31 October 2018)

[9] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in CVPR, 2016.