IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 4, May 2022
Impact Factor: 6.252

Deep Learning for Computer Vision: A Brief Overview of YOLO

Kunal Patel[1], Akash Patil[2], Abhiraj Shourya[3], Rajesh Kumar Malviya[4], Prof. Maghana Solanki[5]
Students, Department of Computer Engineering[1,2,3,4]
Assistant Professor, Department of Computer Engineering[5]
D. Y. Patil College of Engineering, Pune, Maharashtra, India

Abstract: *Inspired by the brain, deep neural networks (DNN) are thought to learn abstract representations through their hierarchical architecture. However, at present, how this happens is not well understood. Deep learning has been overwhelmingly successful in computer vision (CV), natural language processing, and video/speech recognition. In this paper, our focus is on the CV. We provide a critical review of recent achievements and methods of in terms of techniques and applications. We identify eight emerging techniques, investigate their origins and updates, and finally emphasize their applications in four key scenarios, including recognition, visual tracking, semantic segmentation, and image restoration. A brief account of their history, structure, advantages, and limitations is given, followed by a description of their applications in various computer vision tasks, such as object detection, face recognition, action and activity recognition, and human pose estimation. Finally, a brief overview is given of future directions in designing deep learning schemes for computer vision problems and the challenges involved therein.*

Keywords: Deep Learning, Computer Vision, Convolutional Neural Network, Yolo

I. INTRODUCTION

Computer vision and machine intelligence (known as machine vision) paradigms are always promoted in the domain of medical image applications, including computer-assisted diagnosis, image-guided radiation therapy, landmark detection, imaging genomics, and brain connectomics. The complex real-life problems prevalent in medical image analysis and its understanding are daunting tasks owing to the massive influx of multimodal medical image data during routine clinical practice. In the field of medical science and technology, the objective of such advanced computational paradigms is to provide robust and cost-effective solutions for the emerging problems faced by humanity. Medical image analysis includes the fields of medical image enhancement, segmentation, classification, and object detection, to name a few. Advanced computer vision and machine intelligence approaches have been employed in the field of image processing and computer vision. A given computer vision system may require image processing to be applied to raw input, e.g. pre-processing images. Examples of image processing include

- Normalizing photometric properties of the image, such as brightness or colour.
- Cropping the bounds of the image, such as centring an object in a photograph.
- Removing digital noise from an image, such as digital artefacts from low light levels.

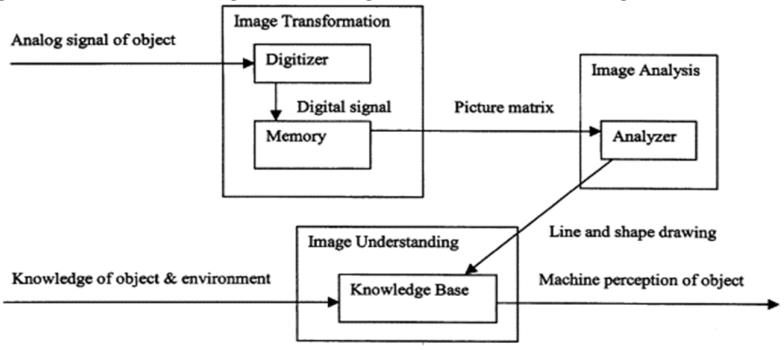

Figure 1: Components of a computer vision system [1]

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 4, May 2022
Impact Factor: 6.252

II. TECHNIQUES

2.1 Convolutional Neural Network

Convolutional neural networks are a specialized type of artificial neural network that uses a mathematical operation called convolution in place of general matrix multiplication in at least one of their layers They are specifically designed to process pixel data and are used in image recognition and processing.

When it comes to substantial training of multiple layers, the Convolutional Neural Network(CNN) is considered as the most momentous approach for a variety of applications [4]. The CNN architecture shown in fig. 2, consists of mainly three types of layers, namely [1]-convolutional layers, pooling layers and fully connected layers. Training the network can be divided into forwarding and backward stages. In the forward stage first, we classify the input image depending upon its weights and bias for each layer. The loss cost is calculated from the input data by using the predicted output. In the backward stage, depending on the loss cost measured, the gradients are calculated for each parameter. Using the gradients, it then updates the parameters for the next iteration. The training procedure can be halted after a satisfactory number of repetitions. The functionalities of the said network are described as follows:

2.2 Convolutional Layer

In this layer, the CNN uses numerous filters to convolve the entire image including intermediate feature maps & generating different feature maps. The major privileges [5] of the convolution operation are-

- Reduce number of parameters using weight sharing mechanisms.
- Correlation between neighboring pixels are easy due to local connectivity.
- Location of object is fixed.
- This advantages leads researchers to replace fully connected layers to put forward the learning process [6,7].

2.3 Pooling Layer

This layer is similar to convolution layer but minimizes the measurements of feature maps and also the parameters of the network. Generally average and max pooling are used. For average and max pooling, Boureu et al. [8] have demonstrated the theoretical details about their performances. Among all the three layers, the pooling layer happens to be the most profusely investigated. There exist mainly three approaches, with varied usage, that are related to the pooling layers. All of these different pooling approaches are discussed as follows

2.4 Fully-Connected Layer

The final layer of CNN consists of 90% of the parameters. The feed forward network forms a vector of a particular length to follow up processing. Since these layers contain most of the parameters, there is a high computational burden while training the data

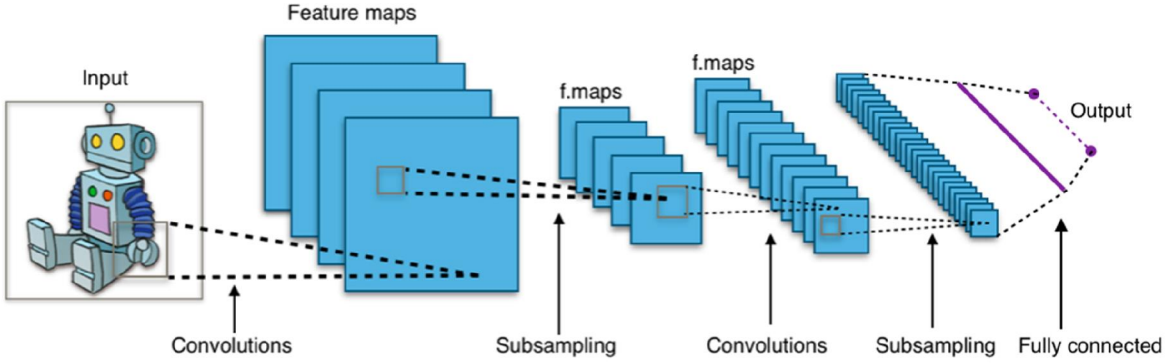

Figure 2: CNN architecture

2.5 Yolo

YOLO an acronym for 'You only look once', is an object detection algorithm that divides images into a grid system. Each cell in the grid is responsible for detecting objects within itself.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 4, May 2022
Impact Factor: 6.252

YOLO algorithm is important because of the following reasons:

- **Speed:** This algorithm improves the speed of detection because it can predict objects in real-time.
- **High accuracy:** YOLO is a predictive technique that provides accurate results with minimal background errors.
- **Learning capabilities:** The algorithm has excellent learning capabilities that enable it to learn the representations of objects and apply them in object detection.

YOLO algorithm works using the following three techniques:

- Residual blocks
- Bounding box regression
- Intersection Over Union (IOU)

A. Residual Blocks

First, the image is divided into various grids. Each grid has a dimension of S x S. The following image shows how an input image is divided into grids. In the image, there are many grid cells of equal dimension. Every grid cell will detect objects that appear within them. For example, if an object center appears within a certain grid cell, then this cell will be responsible for detecting it.

Figure 3: Residual block

B. Bounding Box Regression

A bounding box is an outline that highlights an object in an image. Every bounding box in the image consists of the following attributes:

- Width (bw)
- Height (bh)
- Class (for example, person, car, traffic light, etc.)- This is represented by the letter c.
- Bounding box center (bx,by)

The following image shows an example of a bounding box. The bounding box has been represented by a yellow outline.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 4, May 2022
Impact Factor: 6.252

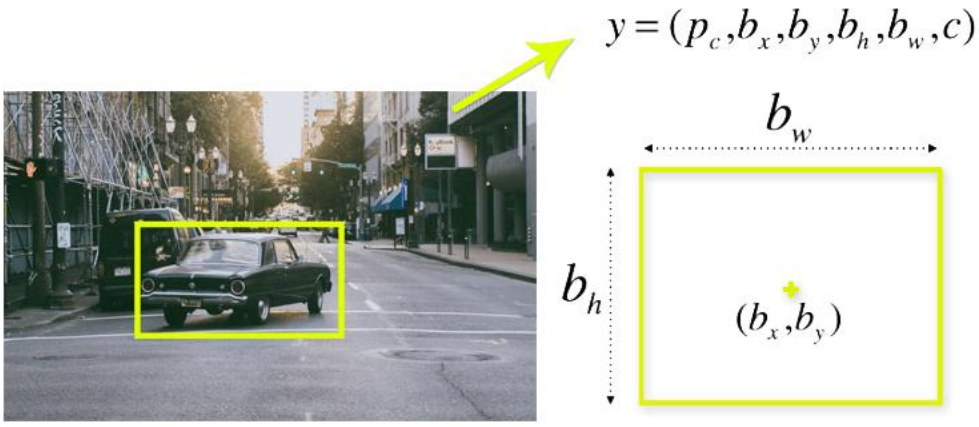

Figure 4: Bounding Box Regression

YOLO uses a single bounding box regression to predict the height, width, center, and class of objects. In the image above, represents the probability of an object appearing in the bounding box.

C. Intersection Over Union (IOU)

Intersection over union (IOU) is a phenomenon in object detection that describes how boxes overlap. YOLO uses IOU to provide an output box that surrounds the objects perfectly. Each grid cell is responsible for predicting the bounding boxes and their confidence scores. The IOU is equal to 1 if the predicted bounding box is the same as the real box. This mechanism eliminates bounding boxes that are not equal to the real box. The following image provides a simple example of how IOU work

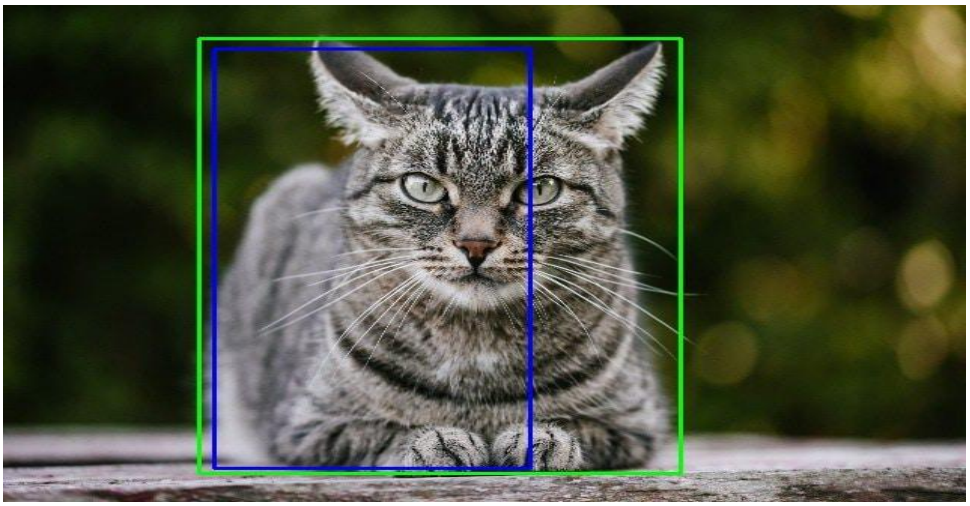

Figure 5: Intersection over union

In the image above, there are two bounding boxes, one in green and the other one in blue. The blue box is the predicted box while the green box is the real box. YOLO ensures that the two bounding boxes are equal.

D. Combination of the Three Techniques

The following image shows how the three techniques are applied to produce the final detection results.

First, the image is divided into grid cells. Each grid cell forecasts B bounding boxes and provides their confidence scores. The cells predict the class probabilities to establish the class of each object.

For example, we can notice at least three classes of objects: a car, a dog, and a bicycle. All the predictions are made simultaneously using a single convolutional neural network.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 4, May 2022
Impact Factor: 6.252

Intersection over union ensures that the predicted bounding boxes are equal to the real boxes of the objects. This phenomenon eliminates unnecessary bounding boxes that do not meet the characteristics of the objects (like height and width). The final detection will consist of unique bounding boxes that fit the objects perfectly. For example, the car is surrounded by the pink bounding box while the bicycle is surrounded by the yellow bounding box. The dog has been highlighted using the blue bounding box.

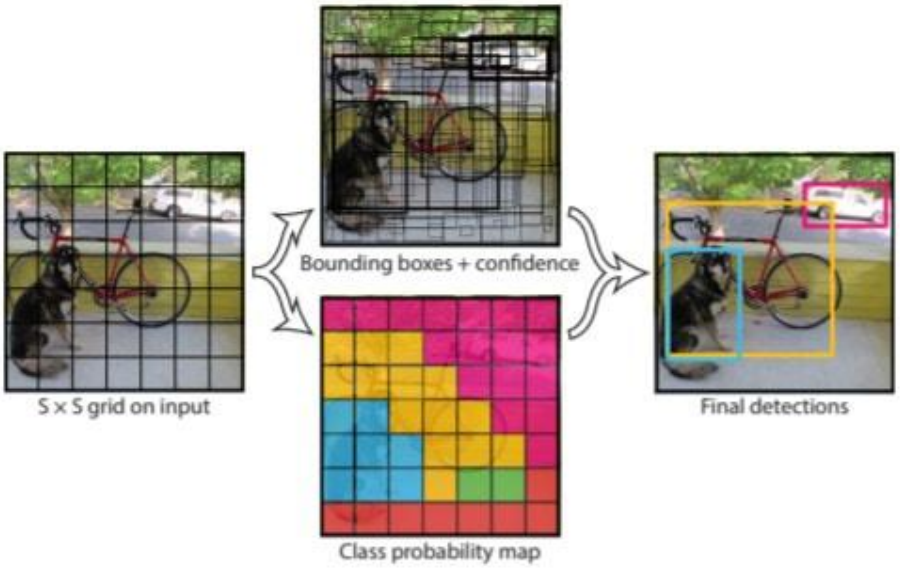

Figure 6: All combine

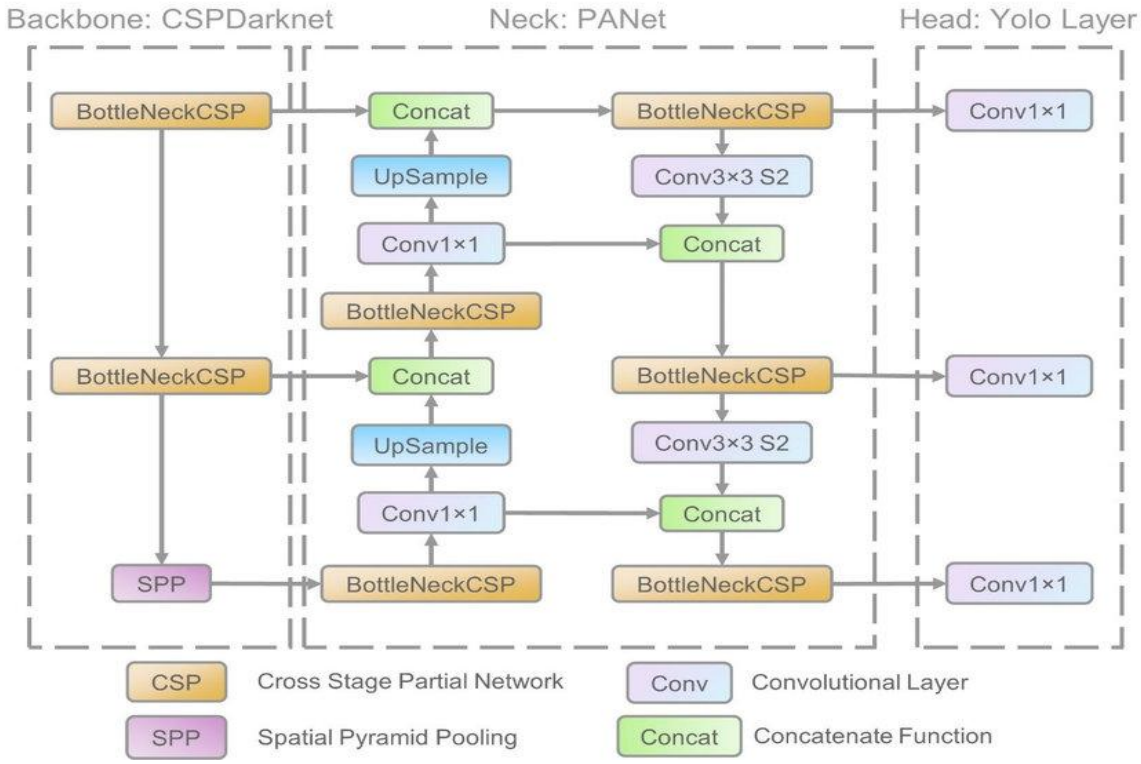

Figure 7: YOLO architecture

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 4, May 2022
Impact Factor: 6.252

III. APPLICATIONS

A. Image Classification

In this depending upon the probability of its presence visual object class labeling can be done [9]. In deep learning most of the methods involve bags of visual words through which initially we get a histogram of quantized visual words then we proceed for classification. Most of the time sparse coding is used torecover the information loss.

B. Object Identification

Object identification is somewhat related to image classification. In object identification process, it takes image as an input and then object estimation can be done depending upon the class and positions. The PASCAL VOC with 20 classes is often used for object identification. YOLO is a advanced object identification after CNN and Faster CNN. Now days YOLO is a popular algorithm for object identification.

IV. CONCLUSION

In this paper, we have given a survey on Deep learning for Computer vision. The analysis of prevailing deep learning architectures is done by developing a categorical layout. We have gone through the main reasons why the YOLO algorithm is important. . The YOLO algorithm and how it is used in object detection. This technique provides improved detection results compared to other object detection techniques such as Fast R-CNN and Retina-Net. We have learned how the YOLO algorithm works. We have also gained an understanding of the main techniques used by YOLO to detect objects.

ACKNOWLEDGMENT

The completion of our paper based on our project study with it a sense of satisfaction, but it is never complete without those people who made it possible and whose constant support has crowned our efforts with success. One cannot even imagine our completion of the literature paper without guidance and neither can we succeed without acknowledging it. It is a great pleasure that we acknowledge the enormous assistance and excellent co-operation to us by the respected personalities.

REFERENCES

[1]. Encyclopedia of Information System 2003. By Mehdi Khosrow-Pour, D.B.
[2]. Renjie Xu, Haifeng lin, Kangjie Lu, Yunfei Liu: A Forest Fire Detection System Based on Ensemble Learning 2021
[3]. Krizhevsky A, Sutskever I, Hinton G E. Imagenet classification with deep convolutional neural networks, in: NIPS, 2012
[4]. LeCun Y, Bottou L, Bengio Y, et al. Gradient-based learning applied to document recognition, Proceedings of the IEEE, 1998, 86(11): 2278- 2324.
[5]. M. Zeiler. Hierarchical convolutional deep learning in computervision.
[6]. Szegedy C, Liu W, Jia Y, et al. Going deeper with convolutions,in: CVPR, 2015.
[7]. Oquab M, Bottou L, Laptev I, et al. Is object localization for free? Weakly-supervised learning with convolutional neu- ral networks, in: CVPR, 2015
[8]. Boureau Y L, Ponce J, LeCun Y. A theoretical analysis offeature pooling in visual recognition, in: ICML, 2010
[9]. MASTER S. Large scale object detection. Department of Cybernetics Faculty of Electrical Engineering, Czech Tech- nical University, 2014.