A Study of Object Detection in Image Processing

Sufiyan Sayyed*, Sayali Patkar, Atharva Patil, Mahendra Patil

Department of Computer Engineering, Atharva College of Engineering, Mumbai University, Mumbai, India

Publication Info

Article history:
Received : 13 February 2020
Accepted : 22 May 2020

Keywords:
Digital Image Processing, Object Detection, Deep Learning, Image Classification, CNN, R-CNN, Fast R-CNN, YOLO.

Corresponding author:
Sufiyan Sayyed
e-mail: *sayyedsufiyan2121@gmail.com*

Abstract

Nowadays, we live in a world where data processing is very important. Power holds those who have the information. Making something out of all those data is becoming a challenge. Image is also one type of data by processing it we get to extract some useable information from it. In computer science, Image processing is the use of digital computers to process a digital image through an algorithm.

Digital image processing can be applied to wide range of fields such as pattern recognition, video processing, and for various business application, medical imagining and industrial automation, etc. Object Detection is a computer technology associated with processing of image and detecting instances of semantic objects. It allows us to understand the scene and to examine it in image or video, deep learning has been applied to object detection in recent years. So this paper discusses and analyzed different methods for image classification and optimized the best algorithm from them for object detection.

1. INTRODUCTION

Digital image processing is a technology field under computer science and engineering which is very prominent and increasing rapidly. This development contributes to technical advances in the fields of computer processing and digital imaging. A computer vision system detects images obtained from an electronic camera, which is like the human visual system where stimuli produced from the eye are interpreted by the brain [4]. Computer vision is a multidisciplinary field that has gained tremendous popularity in recent years. Object detection is a technique of computer vision, which identifies and locates objects in a provided image or video. It can count objects in a frame and determine and/or monitor their exact position in respective frame, all with accurately labeling each of them. Object detection is related to techniques in computer vision such as image recognition, classification and segmentation, allowing us to recognize and interpret a scene in images or videos. But major variations still exist. Image recognition refers a picture to a name. A dog's image earns a "dog" mark. An image consist of more than one cars will still receives the label “car”. In comparison, object detection draws a segmentation mark such as drawing rectangle or coloring the entire object with opaque color, which is car in this case and marks the object name "car" above it. Therefore identification of objects provides more knowledge about an image than recognition. This paper discusses many deep learning-based approaches for detecting objects and provides a review on these detection frameworks namely CNN, Fast R-CNN, R-CNN, and YOLO.

2. LITERATURE REVIEW

Object detection is a technique of computer vision that is use for detecting objects in given image(s). The entity may be usually defined either from the image or from video feeds. It locates an object's presence in an image and highlights it, either by drawing box around it or segmenting object by opaquely coloring it. Deep learning has been used in the classification of images in recent years [1].

2.1. Convolutional Neural Network (CNN)

Task involving computer vision are highly effective by Convolutional Neural Network (CNN) CNN has a network of inputs layer, hidden layers and output layers. The hidden layers typically consist of various different layers such as convolutional layers, ReLU layers, pooling layers, fully connected layers and Softmax or Logistic layer. Convolutional layers add a convolutional process to the data provided. It moves the data to the succeeding layer. Pooling in the next layer combines neuron cluster outputs into one single neuron. Fully connected layers are layers where each and every neuron present in a layer is connected to each and every neuron in the following layer. Softmax or logistics layer is mostly last layer of CNN which decides for class logistics layer is used when we require Yes/No answer (two class) where as softmax is used when we have to classify for more than two classes. CNN works by removing image attributes. This removed the need for the extraction of manual functionality. Hidden layers are key

of CNN's feature detection learning, these hidden layers can be 10 or 100 in number. Each new layer helps in learning more features. So basically a CNN begins with an image and create a feature map to it. Then it applies a ReLU function to increase non-linearity. After that each feature map is passed through pooling layer after passing with pooling the pooled images are flattens into one vector, then the vector is given to a fully connected layer that direct the features throughout the network. The fully connected layer that is situated at last is connected to Softmax or Logistic layer which decides for final classes that is being training for, it trains through forwarding propagation and back propagation for specified number of epochs. The process of training is repeated until we are left with a almost perfect and in some case perfect neural network with feature detectors and trained weights [2]. The limitation of the discussed method is it has a slow real-time prediction to implement because we have to run a prediction for every selected region [1].

2.2. R-CNN and Fast R-CNN

The R-CNN technique combines two key approaches: the application of high-capacity convolutional neural networks for proposal of the positioning and segmentation of objects in the bottom-up region, and supervised pre-training for auxiliary tasks. The object-detection system is divided into three modules each having their own functionalities. The first module computes a regional proposal which doesn't rely on any single category, defining the potential detection in image. Next module is responsible for extracting fixed-feature vector by using a broad convolutional neural network. The third module comprises of a class of support vector machines (SVM) [3]. The algorithm takes image as an input and extracts around 2000 bottom-up region proposals. The CNN is used to compute each feature for proposal. Linear support vector machines (SVMs) specific to each class are used to classify each region.

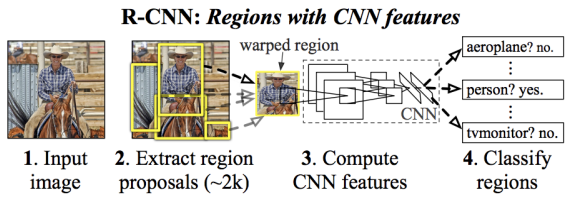

Figure 2.1: R-CNN Module

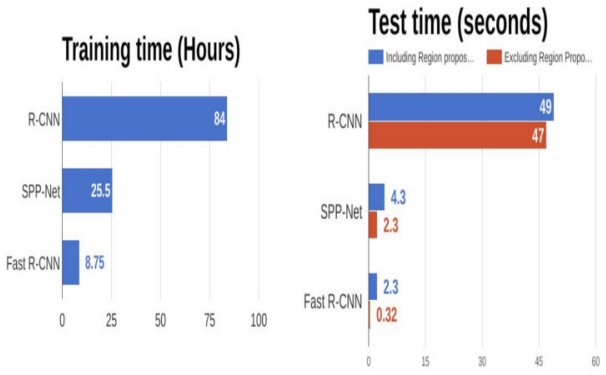

Figure 2.2: Training and test time of R-CNN and Fast R-CNN

R-CNN has some drawbacks:

- Multi-stage pipeline is used for training. Which consists of Modulating a convolutional neural network on object proposals, fitting SVMs to the CNN features, and at the end learning bounding box regressors [3].
- VGG16 (CNN Deep network model) take up huge amounts of space thus; training is expensive with respect to space and time [3].
- ConvNet forward pass for each object proposal makes object detection slow [3].

When we bring R-CNN and Fast R-CNN head to head, Fast R-CNN has some advantages such as higher mean average precision (mAP), no disk is required because of single-stage training, training updates all network layers and for feature caching disk storage is also not required [4].

In Fast R-CNN architecture it gives image as input to CNN instead of region proposal which is used by traditional R-CNN. Then the image is processed by algorithm to generate a convolutional feature map, using convolutional and max-pooling layers. For each region proposal, we extract fixed size vector from each feature map with the help of RoI pooling layer. The vectors are made of fixed size the reason being that it can be given as input to fully connected layer. These branch out into two output layers. With the help of RoI feature vector, softmax is used to estimates probability over many classes of objects. It also produces four values which are bounding box of image [3].

All the techniques like CNN, R-CNN and Fast R-CNN doesn't look at entire image instead use region within image with high probabilities of having an object.

2.3. You only look once (YOLO)

You only look once (YOLO) method is intended for real time processing. In YOLO the image data is distributed into the grid of S x S. Every cell of the grid is required to predict

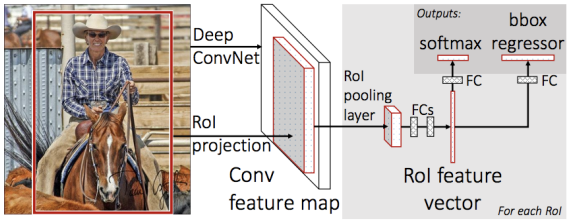

Figure 2.3: Working of Fast R-CNN

2010-2019 S-JPSET : Vol. 12, Issue Supplementary-1, ISSN : 2229-7111 (Print) and ISSN : 2454-5767 (Online)

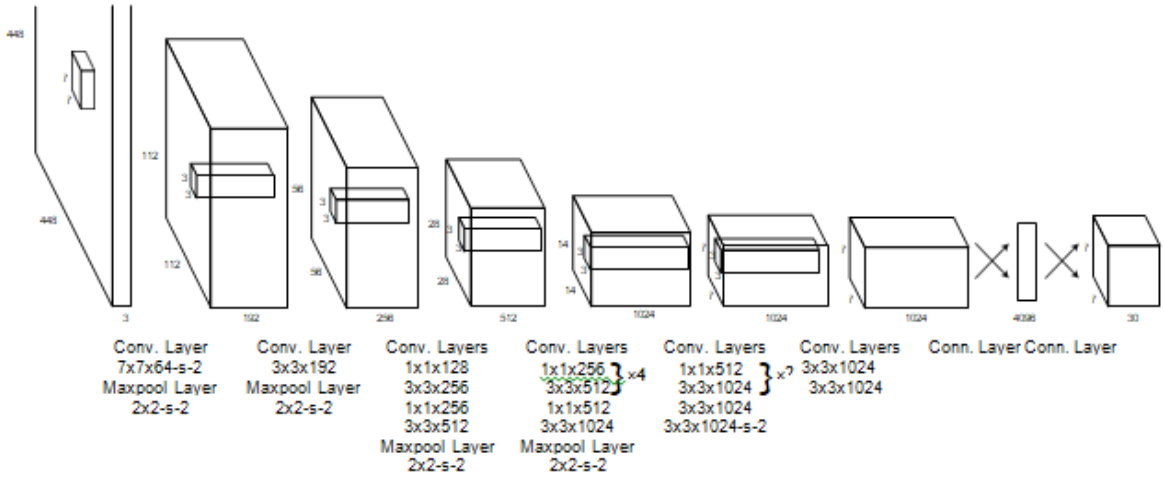

Figure 2.4: Architecture of YOLO model

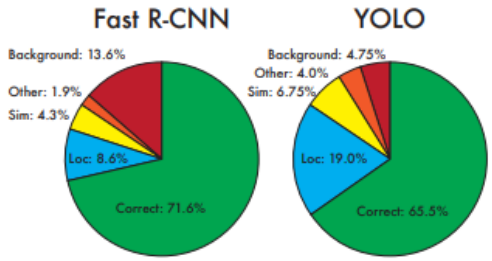

Figure 4: Error Analysis: Fast R-CNN vs. YOLO These charts show the percentage of localization and background errors in the top N detections for various categories (N = # objects in that category).

Figure 2.5: Error in detection Fast R-CNN vs YOLO

only one object. Out of all grid cells each cell forecast a specified number of bounding boxes. Every bounding box has five elements and they are x, y, w, h and a confidence value of a box. The confidence value indicates probable of an item in the box and precision of the bounding box. In the bounding box the parameters are ' w' that is width,' h' that is height,' x' and 'y' are cell offset. Thus x, y, w, and h will be having values 0 and 1. There are twenty conditional class probabilities for each cell. The conditional class probability states how probable an object being detected is from a certain class (for every cell one probability of each class). The prediction of YOLO therefore is of a form (S, S, B + C) = (7, 7, 2 x 5+ 20) = (7, 7, 30) For S = 7 B = 2 and C = 20[5].

YOLO's key idea is, creating a CNN network for prediction of a tensor (7, 7, 30). Using CNN it reduces the dimension to 7x7, with 1024 output channels at each position. Linear regression is performed by using two fully connected layers which makes boundary box prediction. We hold high box confidence values as our preliminary predictions for making final prediction. For every prediction box, the confidence value for the class is calculated as the box confidence value x conditional class probabilities. YOLO has 24 convolutional layers that are connected to two fully connected layers (FC).Alternatively, certain convolution layers use 1 x 1 reduction layers for reduction of depth in the feature map. Then at the last convolution layer output is of (7, 7, 1024) tensor. Thus, making the tensor flat. Two fully connected layers are used as linear regression, these layers outputs 7×7×30 parameters which is then reshaped to (7, 7, 30) vector, that it 2 boundary box predictions for each position [5].

3. CONCLUSION

In this paper, we discussed various object detection techniques and how these techniques work. Eventually, conclude that YOLO is much more effective and practical for real-time application. It's easy to construct and can be trained on a complete image directly. Thus YOLO is much more efficient and fastest to use when computing than other algorithms.

4. ACKNOWLEDGEMENT

We are grateful to our HOD of Computer Engineering Prof. Suvarna Pansambal for providing us the necessary help and encouragement.

5. REFERENCES

[1] Geethapriya.S, N. Duraimurugan, S.P. Chokkalingam,"Object Detection with YOLO" 2019 International Journal of Engineering and Advanced Technology.

[2] Tianmei Guo, Jiwen Dong, Henjian Li, Yunxing Gao, "Simple Convolutional Neural Network on Image Classification",2017 IEEE 2nd International Conference on Big Data Analysis. Beijing, China.

[3] Girshick Ross,"Fast R-CNN." 2015 Proceedings of the IEEE International Conference on Computer Vision.

[4] Burak Berber, "What is the difference between CNN and R-CNN?" ,Quora, (2019) December 7 Retrieved from https://www.quora.com/What-is-the-difference-between-CNN-and-R-CNN.

[5] Joseph Redmon, Santosh Divvala, Ross Girshick, Ali Farhadi," You Only Look Once: Unified, Real-Time Object Detection", 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas, NV, USA.

2010-2019 S-JPSET : Vol. 12, Issue Supplementary-1, ISSN : 2229-7111 (Print) and ISSN : 2454-5767 (Online)