International Journal of Advance Research in Science and Engineering
Volume No. 11, Issue No. 05, May 2022
www.ijarse.com
IJARSE
ISSN 2319 - 8354

VISUAL OBJECT DETECTION AND CLASSIFICATION USING YOLO ALGORITHM

Dr. B. Gnana Priya

Assistant Professor, Department of Computer Science and Engineering,

Faculty of Engineering and Technology, Annamalai University

Abstract

Visual object tracking is one of the most important areas of computer vision. Object detection has received a great deal of coverage in the past decades. With the introduction of Convolutional Neural Network (CNN) in 2012, different network architectures are made available for object detection. You Only Look Once (YOLO) is popular and simple approach nowadays due to its speed and accuracy in real-time object identification. In this paper, object detection using YOLO for custom image dataset that contains six specific classes are considered for detection.

Keywords: Object Detection, YOLO, Computer Vision, Image Classification

1. INTRODUCTION

Object detection is a much studied and widely applied technology which detects the semantic objects in a digital image or video pertaining to a variety of classes in which the objects belongs to. Object Localization is applied extensively for locating the objects in the image. In real-time systems, the scene contains many object of interest and it is often essential to locate more than one object. There exists variety of techniques for object detection and can be broadly classified into two categories. The first category is region based neural network classifications and examples being CNN and RNN. The interested regions from the image are selected and are classified using Convolutional Neural Network. The region based method is seldom slow, since prediction need to be made for every selected region. Also, all object detection algorithms before the introduction of YOLO used regions to localize the object within the image. They never looked at the complete image, but the attention was only on parts of the image that had a high chance of containing any given object.

YOLO (You Only Look Once), the second category of object detection algorithm is based on regression. A single Convolutional Neural Network is applied to predict the bounding boxes and class probabilities directly from full images in YOLO. Rather than selecting the interested regions from the image, the classes and bounding boxes are predicted. The algorithm uses only a single run to detect multiple objects. In real-time object detection applications YOLO proves to efficient and fastest due to gain achieved by changing the architecture to a single neural network. Also for classification problems YOLO proves to be the fastest when compared to other algorithms. YOLO predicts the class of the object along with bounding box specifying object location. Four descriptor are used for describing each bounding boxes. They are Center of the box, Width, Height,

corresponding class of an object. It also finds the probability (real number) that there is an object in the bounding box. YOLO algorithm makes localization errors but predicts less false positives in the background. YOLO splits the image into a grid, rather than searching for interesting regions in the image. Each one of the grid is responsible for predicting its own bounding boxes.

2. RELATED WORK

Different strategies have been proposed to solve the problem of object identification throughout the years. These techniques focus on the solution through multiple stages. Namely, these core stages include recognition, classification, localization, and object detection. Object Detection [1][2] is modeled as a classification problem where at all possible locations we take gaps of fixed sizes from the input object to feed these patches into an image classifier. Every window is fed to the classifier which determines the object's class in the window. Along with the technological progression over the years, these techniques have been facing challenges such as output accuracy, resource cost, processing speed and complexity issues. With the invention of the first Convolutional Neural Network (CNN) algorithm in the 1990s inspired by the Neocognitron by Yann LeCun et al. [3] and significant inventions like AlexNet [4], which won the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) in 2012 (thus later referred to as ImageNet) CNN algorithms have been capable of providing solutions for the object detection problem in various approaches. With the purpose of improving accuracy and speed of recognition, optimization focused algorithms such as VGGNet [5], GoogLeNet [6] and Deep Residual Learning (ResNet) [7] have been invented over the years. You Only Look Once: Unified, Real-Time Object Detection, by Joseph Redmon. Their prior work is on detecting objects using a regression algorithm. To get high accuracy and good predictions they have proposed YOLO algorithm in this paper [8]. Understanding of Object Detection Based on CNN Family and YOLO, by Juan Du. In this paper, they generally explained about the object detection families like CNN, R-CNN and compared their efficiency and introduced YOLO algorithm to increase the efficiency [9]. Learning to Localize Objects with Structured Output Regression by M. B.Blaschko is a paper about Object Localization. In this, they used the Bounding box method for localization of the objects to overcome the drawbacks of the sliding window method [10].

3. OBJECT DETECTION USING YOLO

First, an image is taken and YOLO algorithm is applied. In our example, the image is divided as grids of 3x3 matrixes. We can divide the image into any number grids, depending on the complexity of the image. Once the image is divided, each grid undergoes classification and localization of the object. The objectness or the confidence score of each grid is found. If there is no proper object found in the grid, then the objectness and bounding box value of the grid will be zero or if there found an object in the grid then the objectness will be 1 and the bounding box value will be its corresponding bounding values of the found object. The bounding box prediction is explained as follows. Also, Anchor boxes are used to increase the accuracy of object detection which also explained below in detail.

3.1 Residual blocks

Initially the image is divided into various grids. Each grid has a dimension of S x S. The grid cells are such that they have an equal dimension. Every grid cell will detect objects that appear within them. If an object centre appears within a certain grid cell, then that cell will be responsible for detecting it.

3.2 Bounding box predictions

An outline to man object in an image is given by bounding box. The bounding box generally consists of the following attributes: (i) Width (bw), (ii) Height (bh), (iii) Class (for example, Human ,bus, car, traffic light in the problem). This is represented by the letter c. (iv) Bounding box center (bx,by). Both image classification and object localization techniques are applied for each grid of the image and each grid is assigned with a label. Then the algorithm checks each grid separately and marks the label which has an object in it and also marks its bounding boxes. The labels of the gird without object are marked as zero.

3.3 Anchor Box

By using Bounding boxes for object detection, only one object can be identified by a grid. So, for detecting more than one object we go for Anchor box. Any number of anchor boxes can be used for a single image to detect multiple objects.

3.4 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.

3.5 Non-max suppression

This will lead to the algorithm detecting the object multiple times instead of just once. This is where Non-max suppression comes in where it ensures the algorithm detects each object only once. As was previously stated, each cell outputs y = (P_c, b_x , b_y , b_w, b_h , c) with P_c being the probability there is an object. Non-max suppression takes the bounding box with the highest P_c, discards any bounding boxes with $P_c \leq 0.6$.

4. EXPERIMENTS

The task of object detection is to find the presence of objects in images and classify them to their relevant classes to which they belong. In this work, we take into account six classes for classification: car, bus bicycle, human, traffic light and truck. The numbers of objects are taken such that they are evenly distributed. They are then annotated with bounding boxes. Each object we want to detect is put into a box and is labelled with the object classes to which they belong. The pre-trained weights from YOLOV3 model is used here. YOLO first takes an input image: Fig 1.The framework then divides the input image into grids (Example: 3 X 3 grid). For each grid image classification and localization are applied. The algorithm then predicts the bounding boxes and their corresponding class probabilities for each object. Non-Max Suppression is used to ensure that every object is detected only once.

International Journal of Advance Research in Science and Engineering
Volume No. 11, Issue No. 05, May 2022
www.ijarse.com
IJARSE
ISSN 2319 - 8354

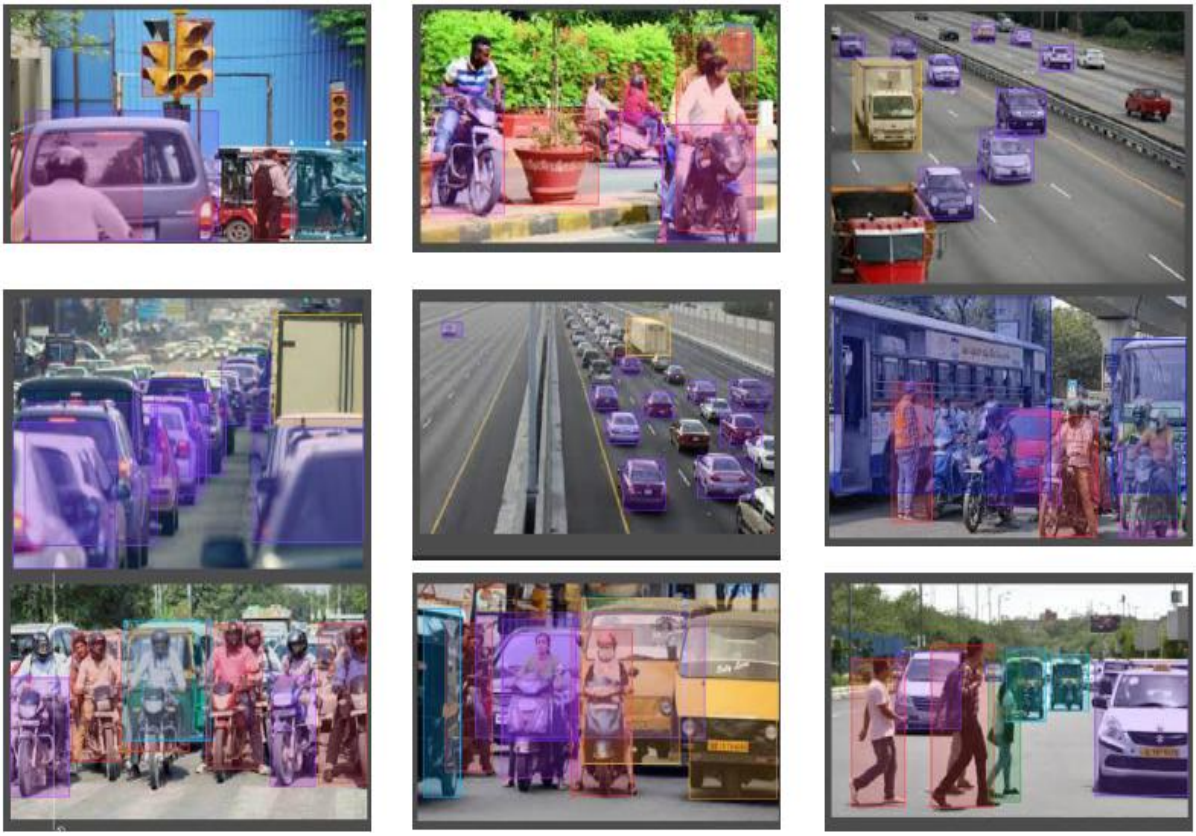

Fig. 1 : Input Images

5. RESULTS

YOLO algorithm uses a convolution neural network which consists of 24 convolutional layers and 2 fully connected layers. Each layer has its own functionality and significance. The convolutional layers are followed by pooling layers and then two fully connected layers are connected. It is pre-trained on the ImageNet dataset that consists of 1000 classes for classification. The pre-training for classification is performed on the dataset with the image resolution of 224 x 224×3. For object detection, in the end, the last 4 convolutional layers followed by 2 fully connected layers are added to train the network. Then the final layer predicts the class probabilities and bounding boxes. All the other convolutional layers use leaky ReLU activation whereas the final layer uses a linear activation. The output is the class prediction of the detected object enclosed in the bounding box. By using two fully connected layers it performs a linear regression to create a 7x7x2 bounding box prediction. A prediction is made by considering the high confidence score of a box. For a single grid cell, the algorithm predicts multiple bounding boxes. To calculate the loss function we use only one bounding box for object responsibility. For selecting one among the bounding boxes we use the high IoU value. The box with high IoU will be responsible for the object.

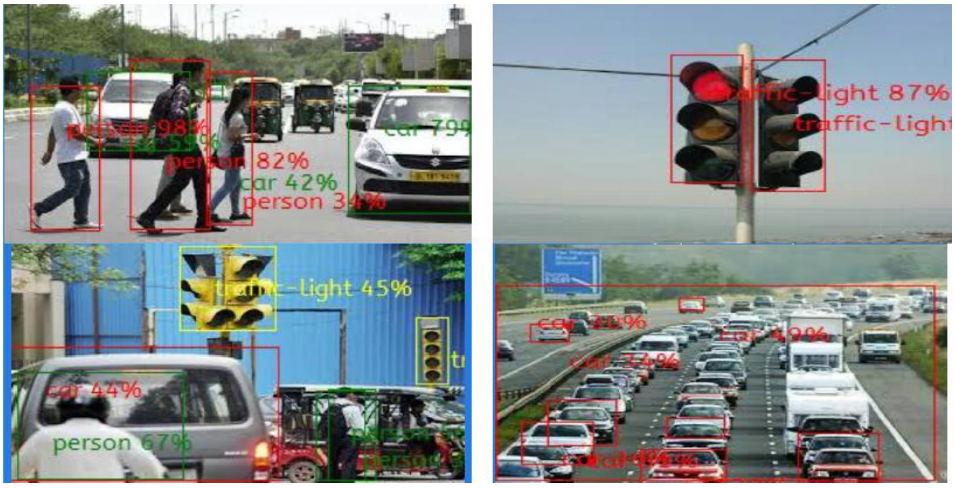

Fig 2 : Object Detection Using YOLO

6. CONCLUSION

In this paper, object detection is done on images by training detector for custom dataset consisting of 600 images for 6 specific classes. The object detection is done using YOLOV3 detector. The model can be tried for more epochs and model can be fine tuned for improved accuracy and precision. In future, the model can be extended to train more classes. It can also be extended for videos in different domains and different objects can be detected and tracked. Different types of object detectors like R-CNN, SSD can also be used and the results can be compared against each others.

REFERENCES

1. D J. Redmon and A. Farhadi, "Yolov3: An incremental improvement," arXiv preprint arXiv:1804.02767, 2018.
2. Mekonnen and F. Lerasle, "Comparative Evaluations of Selected Tracking-by-Detection Approaches," IEEE Transactions on Circuits and Systems for Video Technology, vol. 29, no. 4, pp. 996-1010, 2019.
3. Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-Based Lerning Applied to Document Recognition," proc. IEEE, 1998.
4. T. F. Gonzalez, "Handbook of approximation algorithms and metaheuristics," Handbook Approx. Algorithms Metaheuristics, pp. 1–1432, 2007, doi: 10.1201/9781420010749.
5. K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 3rd International Conference on Learning Representation ICLR 2015 - pp. 1–14, 2015.
6. C. Szegedy , "Going Deeper with Convolutions," 2015, doi: 10.1002/jctb.4820.
7. K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," doi: 10.1002/chin.200650130.
8. J.Redmon, S.Divvala and R.Girshick, "You Only Look Once: Unified, Real-Time Object Detection", The IEEE Conference on Computer Vision and Pattern Recognition , 2016.

9. J.Du1,"Understanding of Object Detection Based on CNN Family", New Research, and Development Center of Hisense.
10. B.Matthew, Blaschko, H.Christoph and Lampert, "Learning to Localize Objects with Structured Output Regression", Published in Computer Vision – ECCV, 2008.
11. D.Erhan, C.Szegedy and A.Toshev, "Scalable Object Detection using Deep Neural Networks", The IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 2147-2154.
12. S.Ren, Kaiming He, R.Girshick and J.Sun, "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks", Published in Advances in Neural Information Processing Systems 28 2015.
13. J.Redmon, A.Farhadi," YOLO9000: Better, Faster, Stronger", The IEEE Conference on Computer Vision and Pattern Recognition , pp. 7263-7271,2017.
14. J.Dai, Yi Li, Kaiming He, Jian Sun, "R-FCN: Object Detection via Region-based Fully Convolutional Networks", published in: Advances in Neural Information Processing Systems 29, 2016.