Object Detection, Classification and Event Detection Using YOLO

Vinod Kumar M
Department of I. S. E,
Bangalore Institute of Technology
K. R Road, Bangalore

Anupama K C
Department of I. S. E,
Bangalore Institute of Technology
K. R Road, Bangalore

Dr.R.Nagaraja
Department of I. S. E,
Bangalore Institute of Technology
K. R Road, Bangalore

***Abstract*— Object Detection is a important application in Computer Vision technology, which is characterized by the extraction of features from the given input image. With the advances in deep learning technology it has gained a drastic speed to it due to its accuracy and performance enhancement. The proposed work aims at achieving the accuracy and classification of the objects in a given image. The work also gives the features of usages with the detected classes by predicting the event or idea about the area surrounding the image. The major problem of object detection is due to the object detection model depend on the other computer vision technologies. The model is developed with completely the using deep learning approach. The model is trained using publically available dataset COCO, which is consisting of common daily images. The resulting system detects the objects of various class and predicts the event based on the objects class present. The available models use deep learning for object detection but the accuracy decreases with comparatively.**

Keywords*— *COCO, Deep Learning, system, feature selection, machine learning.

I. INTRODUCTION

Object detection and classification is a method of detecting the various types of objects present in image. In a single it may be consisting of different numbers of different types of object present a model developed to detect these kind of objects. Some time the image consisting of these objects may be blurred algorithm or model developed should be designed to predict all objects to its maximum. The objection detection has many times gone wrong because of multiple objects and multiple images present so proper model should be designed to predict the images or objects with all types or different class.

In practice there are different types of different number of models have been developed for object detection. Usage of proper object detection algorithms have various ways of implementations like driverless automatic cars highly depend on object detection in real time decision making so a real time object detection plays a vital role in making decisions. most of the object detection uses Convolution neural network. here a single unified detection is used where a single convolution neural network is run is applied on the image to get bounding boxes then to predict the class object.

Then after making use from class object obtained future usage of detected class object done by getting count of number of object present in a image this is usefull in getting total number of present people in a seminar hall or in classroom there by a single image does a job of counting number of people.

The proposed idea is simple that it takes an input divides it in to grid cells them for all the the divided grid cell the CNN is applied in one run. Then by feature extraction of each grid cell object class is obtained this the part of YOLO algorithm. Then the object class labels are taken with along dataset dataset can be COCO, Imagenet, WIDER, VOC or any other object detection dataset which are available in the open source for access. Imagenet is consisting of class names labels of 1000 categories(like glass, train, car, dog, person..etc)

The solution obtained is with bounding box which contains the class labels for the detected object then in our solution extended to take the count of various objects present here making use of class label person for calculating the number of persons present in the image. After obtaining the each number of person or other object class extend the work to get the information about the type of event in the image by making use of class label or area where the object is present place information simple strategy is that for example if a cake and balloons are present in a image it can predict it as a birthday event. Likewise if it is raining umbrellas can be detected as the object.

Object Detection algorithms can be divided based on usage or followed or working procedure they follow. The algorithms can use any of these methods for detecting the objects present in image.

i. Classification Based Algorithms: These algorithms work on the basis of selecting important regions in a given input region. They work in two steps. First step is to select on the important regions where the objects are present in the image. In next step classifying the interested image using convolution neural networks. Some of the known algorithms based on this method are Region-based Convolution Neural Networks (RCNN), Fast RCNN, Faster RCNN. Due to the slow solution these usages are reducing.

ii. Regression Based Algorithm: Here instead of selecting regions and running every time neural network for each selected region in regression method the algorithm is applied on the whole image and run once. It gives the output with bounding boxes with selected class of object naming. The major example that uses regression method is You Look Only Once(YOLO) which is made use in the proposed system due to is fast solution.

The problem statement is that object detection in an image. All the objects should be identified in the image is an major problem because the minute objects in the image may not be identified because of space problem in an image, the detected objects should be classified into correct category, sometimes because of similar features it may predicted wrong.

II. RELATED WORK

A. *Literature Survey*

Object detection models are not a new concepts early years there are many models that have been proposed and each have their own usages and also they their own drawbacks so that new models have been proposed to address the drawbacks or to address them.

P. F. Felzenszwalb, R. B. Girshick, D. McAllester and D. Ramanan [1] proposed a model for object detection based on regions called Deformable parts model called DPM. This model is based on high resolution areas or regions in the image are taken in to consideration and drawn feature extractions. Then model uses a pipeline to extract the static features from that region then drawing the bounding box for that predicted regions of high scoring region. This approach is used in DPM to detect the objects in the image. The dis advantage of the DPM is they are taking high scoring regions for the object detection other regions are left out without taking into considerations. Without embedding other regions may have chances of available of objects.

K. E. A. van de Sande[2] proposed a approach called selective search for detection of objects in the image by using various algorithms to search the locations in the image where the object is present instead of depending on only one search approach selective search used various algorithms for localization called selective search, segmentation to locate possible region where the object present and variety of partitioning the image to get more possible way predicting presence of image. Then this selective search applies a bounding box , then convolution neural network extracts the feature from the bounding box then applies the non-max suppression to eliminate duplicate detections. Yolo is similar in making bounding box and then using feature extractions but this selective search is limited to predicting particular locations in the image.

J. Redmon and A. Angelova[3] ,presented Fast YOLO, a new framework for the purpose of real-time embedded object detection in video. Although YOLOv2 is considered as a state-of the-art framework with real- time inference on powerful GPUs, it is not possible to use it on embedded devices in real-time. Here, we take advantage of the evolutionary deep intelligence framework to produce an optimized network architecture based on YOLOv2. The optimized network architecture is utilized within a motion-adaptive inference framework to speed up the detection process as well as reduce the energy consumption of the embedded device. Experimental results showed that the proposed Fast YOLO framework can achieve an average run-time that is _3.3X faster compared to original YOLOv2, can reduce the number of deep inferences by an average of 38.13%, and possesses a network architecture that is _2.8X more compact.

Sermanet, P., Eigen[4] proposed a model Deep multibox which uses Convolution neural network to process to image and it takes the input of area of instrest where the object is present. Deep multibox meaning to locating single bounding box for each detected object in the image.That take place is single neural network is applied on the interested region and then their confidence score is calculated to assign to a particular class label. This model varies in a position where it selects area of interest otherwise neural network application are same to this model output image consists of bounding box with reasonable confidence score values.

Overfeet[5] which uses Convolution neural networks to classification, localization and object detection. here the model used sliding window and multiscale both used to in Convolution neural network. The model trains neural network to get localization and apply the localizer to get the object detection or to perform object detection. The model uses the Imagenet dataset which consisting of over 1000classes to detection of objects. The main drawback is that it uses optimize for localization and does not concentrate on performance. Only concentrating on local information while making prediction .

S. Rahman, S. H. Khan and F[6] proposed a model, Object detection is considered as one of the most challenging problems in the object detection computer vision, since it requires correct prediction of both classes and locations of objects in images. In this study, we define a more difficult scenario, namely zero-shot object detection (ZSD) where no visual training data is available for some of the target object classes. Model yolo present a novel approach to tackle this ZSD problem, where a convex combination of embeddings are used in conjunction with a detection framework. For evaluation of ZSD methods, here proposed a simple dataset constructed from Fashion-MNIST images and also the custom zero-shot split for the Pascal VOC detection challenge. The experimental results suggested that our method yields better results for ZSD.

III. PROPOSED WORK

The overview of the proposed work is as shown in the flowchart Figure 1.

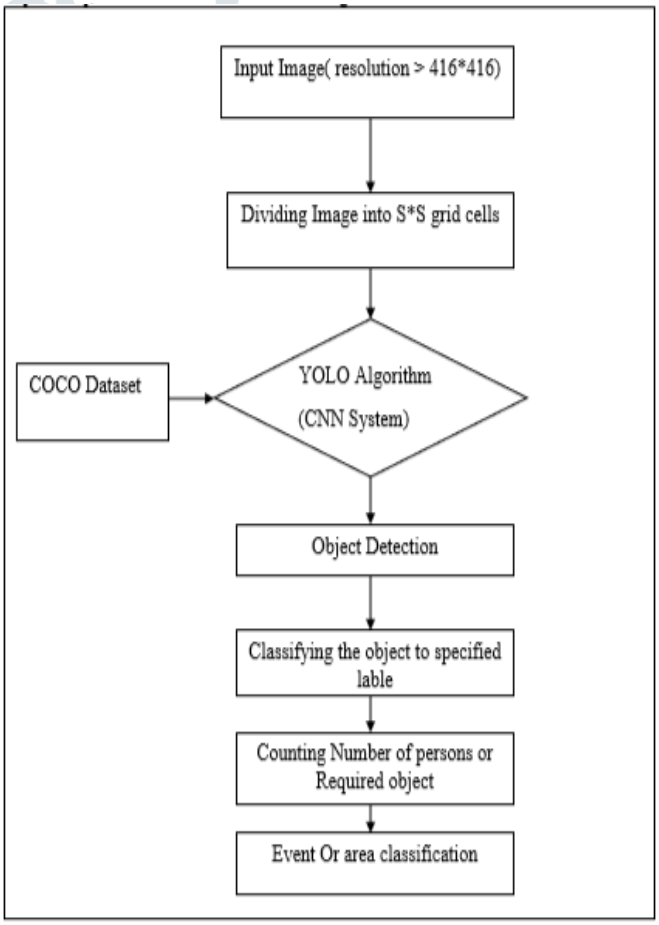

Figure 2. Flow Chart of the Proposed Work

The proposed work is implemented in python.This section gives how yolo is implemented and how it works with our model to detect the objects. Some of the features are unique with respect yolo and also some are taken from other models to yolo(shared).

One of the unique feature which represents the yolo in its own way is splitting the image into number of grid cells for easy understanding here a image is divided into 3*3 grid. The Figure 2 shows how the image is splitted or divided into 9 grid cells.

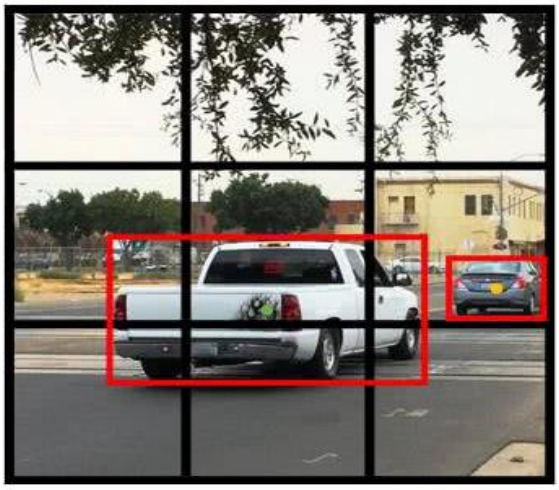

Figure 2 Splitting Image into Grid Cells.

Each Box will have the specific parameters that leads to predict the image. So, specification for each grid cell can be represented by,

$$y = \begin{bmatrix} p_c \\ b_x \\ b_y \\ b_h \\ b_w \\ c_1 \\ c_2 \\ c_3 \end{bmatrix}$$

Where, Pc is the value for detecting object have value 1 if object detected otherwise 0.If Pc is zero all other values are not taken the whole grid value becomes Zero.
bx is the x co-ordinate value for the upper left corner.
by is the y co-ordinate value for the upper left corner.
bh and bw are the width and height of the particular grid cell.
C1,C2,C3... etc are the class labels for the object detection.

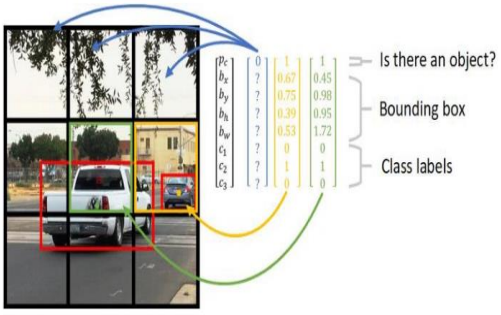

Figure 3 Output labels of the each Grid Cell.
Output of after labeling each grid cell of taken is shown in Figure 3 here for empty grid cells Pc value is zero so, that all other parameter values are taken taken into consideration. bx and by values are the x and y co-ordinate values which is in between 1 and 0. Bw and Bh are the values of height and weight which may be greater than 1.

Here for the class variable values 3 classes are taken it can be Person, Car, Dog. C2 variable is taken 1 since it detects the class object of car. C1, C3 values are taken zero since no objects with respect these class are not taken. In our project total number of 80 class objects are taken so value of output for the variable Y is five variables along with 80 different classes ie, 5+80 labels for each grid cell is taken. This for each grid cell containing one object if the grid cell multiple object then the challenge arises to calculate multiple object in each grid cell. So, yolo gives a way to predict multiple objects in the same grid cell by using Anchor Box. Objective is to predict multiple objects in same image so here is the challenge to predict the multiple object in a single grid cell or in a single image.
Anchor Boxes are very important in objects due to presence of the multiple objects overlapping in a single image that can be seen in the Figure 4 which gives a view where a person and a car object are in the single and in a single grid where it need to detect in the same grid cell 2 objects i e, car and a person.

Figure 4 Detecting multiple objects in single Grid cell.

The idea of anchor box adds a another dimension to object detection by adding a 2 values for the same grid cell. In the figure 5.4 both car and person objects are present in single grid cell so it has to create 2 anchor boxes for same grid cell keeping same centre point. So the value for the output y increases the label by taking 2 object values in single y which is represented by below expression of y.

$$y = \begin{bmatrix} p_c \\ b_x \\ b_y \\ b_h \\ b_w \\ c_1 \\ c_2 \\ c_3 \\ p_c \\ b_x \\ b_y \\ b_h \\ b_w \\ c_1 \\ c_2 \\ c_3 \end{bmatrix} = \begin{bmatrix} 1 \\ b_x \\ b_y \\ b_h \\ b_w \\ 1 \\ 0 \\ 0 \\ 1 \\ b_x \\ b_y \\ b_h \\ b_w \\ 0 \\ 1 \\ 0 \end{bmatrix}$$

Anchor box 1
Pedestrian

Anchor box 2
Car

So from the expression one anchor box for the pedestrian and another anchor box for the car in the same grid cell is taken in single y output. Here C1 is 1 for the pedestrian and C2 is 1 for the car with referring to their class labels.

IV IMPLEMENTATION

After the implementation of yolo algorithm with prediction main intention the flow for prediction is given in figure 5 which starts with Processed image taking 600*600 feeding that image into Deep convolution network. Then for each grid cell the Anchor Bounding boxes are drawn with each bounding box parameters. After drawing the filter boxes if the cell didn't had any object class they are filtered with filter score. The final output of yolo algorithm is the predicted class with having Bounding Box for each found object.

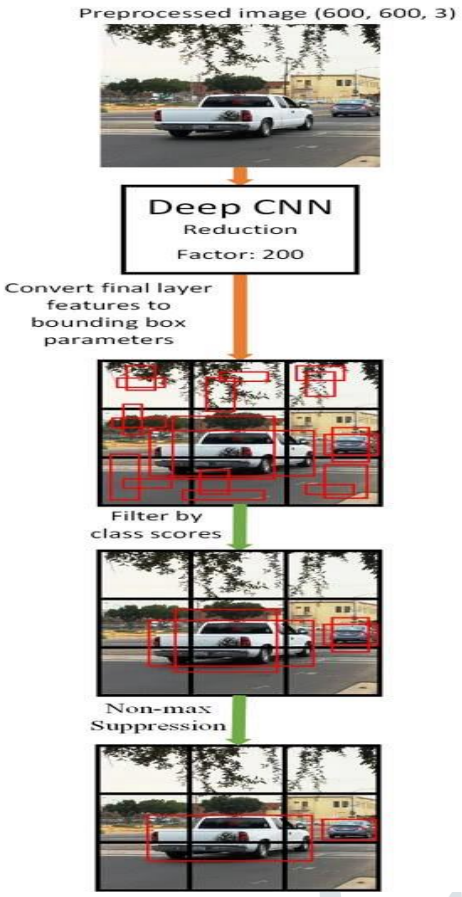

Figure 5 Prediction Flow with Yolo

Implementation is the process of putting a decision or plan into execution can be seen from Figure 5 and is the realization of an application, or execution of a model, design, specification, standard, algorithm or policy. As such implementation is the action that must follow any preliminary thinking in order for something to actually happen.

Implementing an object detection system includes:

i. Dataset collection and preprocessing of data.
ii. Darknet installation.
iii. Training the dataset using the darknet.
iv. Yolo Implementation.
v. Object detection for the given input image.
vi. Event detection of the image.

i Dataset collection

In this project the coco dataset is used. The coco dataset contains about 80 different classes. The dataset is ready for the training in the darknet with all it's weights and the fully connected networks with all its class names.

ii Darknet installation

Darknet is an open source neural network framework which is written in C. Using this darknet the objects can be detected in an image. The darkent is downloaded through opencv in the program which helps in easy computation of the program.

iii Training the dataset using darknet

The dataset used for training is the coco dataset. Give the dataset for the training by giving the images, convolution network, class names of the images and weights. These are the input for training in the darknet. In this training the features of the objects in the image are extracted.

iv Object detection for the given input

The object detection is nothing but identifying the extracted features into an object and build the bounded box around the object and classify it into its category or class. The COCO names file stores all the class variable names referring to particular objects these are taken by the yolo algorithm on matching of ClassID. So, each object name is having particular unique ClassID.

v Event detection for the image

The events of the image is categorized in this step based on the objects it is classified. For example, if there is a baseball, baseball bat, glove are identified as the objects in the image then the event is categorized as sports.

From the explained steps the object detection and event detection takes place, the above steps are the generalized steps of the implementation for the object detection and event detection in an image using yolo.

In this project the output of yolo is a convolution feature map that contains the bounding box attribute with the depth of feature map. The bounding box attributes predicted by cell are stacked one by one each other. So, the second bounding cell should be access at (5,6) the the index will be map[5,6 (5+C) : 2* (5+C)]. The threshold of object confidence, adding the grid offsets value to the centre of the bounding box and applying anchors are very inconvenient for the output processing.

The another problem is the detection will happen in three scales, so the dimensions of the prediction feature maps will be different. Even though the dimension of three feature maps are different, the processing of the output are done in similar.

V RESULTS DISCUSSION

Results or output of the experiment or the entire involves the output image with bounding boxes for the objects detected in the in the image Figure 6 shows the Inference speed FPS and Accuracy precision in mAP of YOLO models which shows tiny-YOLO with maximum Inference speed and Accuracy.

This shows the maximum accuracy and speed further the detection of objects and events in the image can be shown where the output is taken and represented in figures with bounding boxes in the image. Further the total count of the persons in the image can also be shown.

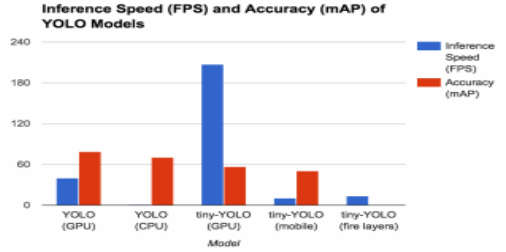

Figure 6 Accuracy and Inference speed of various versions of YOLO.

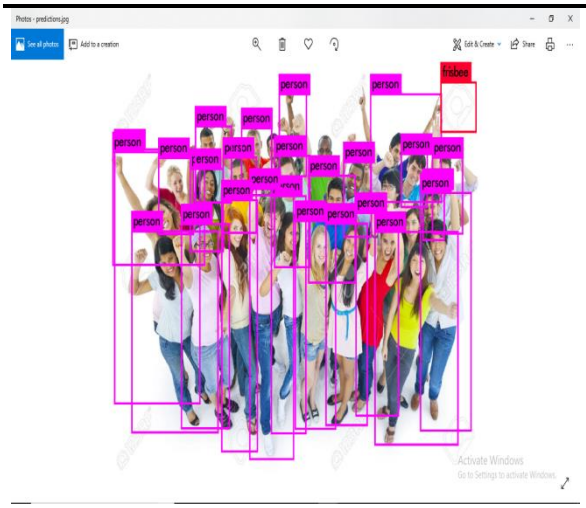

Figure 7 Output image for counting number of persons.

The figure 7 it gives the bounding boxes for images with persons that are many in number for this the number of persons has to be counted this is made easy be getting the class Id of the object.

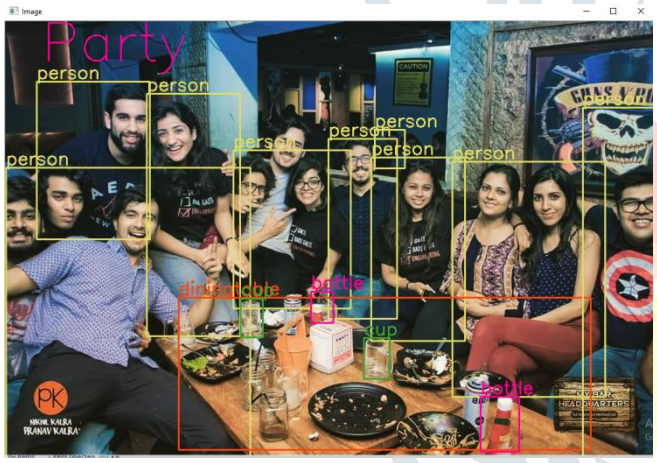

Figure 8 Output of Event detection

The Figure 8 gives the prediction of event in the image which is labelled on the top left corner of the image. By recognizing the object class wineglass, bottles , cup we can predict the images including these class of objects is Party.

VI CONCLUSION AND FUTURE WORK

The way towards the object detection and classification is a important area in the study of computer vision technology. The model is simple to construct and can be trained for images of various type. The images can have different class objects all the class of objects are identified upto the level of accuracy also based on the class of specific class which are unique a prediction of location or the event taking place in the place is also identified upto a level. Trained images which consists of different 80 classes of objects which are unique these classes give a way to identify the event or location in the images.

The model has a tremendous future enhancement where the expressions of the human face can be detected and classify the situation and also detection of exact location by processing the type of people and action of the people present. In Future can also used to classify the images based on the all the objects by taking or improving the number of classes in the model can also give more addition to the results. In live streaming applications on live telecasting instant classification and identification can be done and give results instantly to the viewers.

REFERENCES

[1] P. F. Felzenszwalb, R. B. Girshick, D. McAllester and D. Ramanan, "Object Detection in Discriminatively Trained Part-Based Models," in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 9, pp. 1627-1645, Sept. 2016.

[2] K. E. A. van de Sande, J. R. R. Uijlings, T. Gevers and A. W. M. Smeulders, "Segmentation a selective search for the object recognition," *2011 International Conference on Computer Vision*, Barcelona, 2011, pp. 1879-1886.

[3] R. Girshick, "Fast R-CNN," *IEEE International Conference on Computer Vision (ICCV)*, Santiago, 2015, pp. 1440-1448.

[4] D. Erhan, C. Szegedy, A. Toshev and D. Anguelov, "Scalable Object Detection Using Deep Neural Networks," *2014 IEEE Conference on Computer Vision and Pattern Recognition*, Columbus, OH, 2014, pp. 2155-2162.

[5] Sermanet, P., Eigen, D., Zhang, X., Mathieu, M., Fergus, R., & Lecun, Y. (2014). Overfeat: Integrated recognition, localization and detection using convolutional networks. In *International Conference on Learning Representations (ICLR2014), CBLS, April 2014* .

[6] S. Rahman, S. H. Khan and F. Porikli, "Zero-Shot Object Detection: Learning to Simultaneously Recognize and Localize Novel Concepts," Asian Conference on Computer Vision (ACCV), Perth, 2018.

[7] J. Redmon and A. Angelova, "Real-time grasp detection using convolutional neural networks," *IEEE International Conference on Robotics and Automation (ICRA)*, Seattle, WA, 2015, pp. 1316-1322.

[8] R. Girshick, "Fast R-CNN," *IEEE International Conference on Computer Vision (ICCV)*, Santiago, 2015, pp. 1440-1448.

[9] J. Redmon, S. Divvala, R. Girshick and A. Farhadi, "You Only Look Once: Unified, Real-Time Object Detection," *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, 2016, pp. 779-788.

[10] J. Redmon and A. Farhadi, "YOLO9000: Better, Faster, Stronger," *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI, 2017, pp. 6517-6525.

[11] Lin Ty . et al. Microsoft COCO: Common Objects in Context. In : Fleet D ., Pajdla T, Schiele ., Tuytelaara T.(eds) computer vision – ECCV 2014. ECCV 2014.