

# YOLO: A Unified Real-Time Object Detection (Review)

Anudeep Reddy Puthalapattu, Department of Computer Science, Michiagn Technological University, Houghton, Michigan, aputhala@mtu.edu

Joseph Redmon, Santosh Divvala, Ross Girshick, Ali Farhadi. You Only Look Once: Unified, Real-Time Object Detection. <https://arxiv.org/abs/1506.02640>

**Abstract**— The objective of the paper is to review the author proposed unified network model alias YOLO algorithm and the contribution of the work and discuss the strengths and drawbacks of the new model. In the end, discussed the experimental results of the author's model, and the demo model is built to detect real-time objects. The author achieved the mean average precision of 75 with the combination of Fast R-CNN and YOLO models. The demo model achieved the confidence interval of 0.9999 and detected the three images with the bounding boxes and assigned the associated class to the object.

**Keywords**—object detection, YOLO model, Fast R-CNN, unified network.

## I. INTRODUCTION

The human eye detects multiple objects simultaneously and the YOLO model detects the multiple objects in an image, unlike traditional models. The review of the unified network model with low latency and the proposed architecture to improve the performance with a combination of Fast R-CNN and YOLO implemented by the author is reviewed. The proposed model establishes the bounding of the boxes around the objects with the class probabilities. At once it assigns one class and predicts two boxes. The author demo project page is provided as a reference <http://pjreddie.com/yolo/>. The implemented models are open source and pre-trained weights are available to download.

## II. SUMMARY

The author presents a new approach to YOLO for object detection rather than traditional methods like Deformable Parts Models (DPM), R-CNN, and its variants (Fast R-CNN and Faster R-CNN), Deep MultiBox, OverFeat, and MultiGrasp.

The DPM approach for object detection uses a sliding window to classify the regions, extract the static features and assign the high scoring regions to respective bounding boxes. The approach by the author replaces the entire system with a single convolutional network, performs more accurately, and provides better results than the DPM model.

The CNN and its variants act as a feature extractor and the extracted features in the output dense layer are fed into the SVM to identify the object within that proposal region. The bounding box precision increases by predicting the four offset values in CNN models. The model takes more time as it requires tuning independently and is very slow for each test image. Even though YOLO shares a similar methodology, unlike CNN it put

restrictions on bounding boxes i.e., 98 as compared to 2000 in CNN models.

The Fast R-CNN and Faster R-CNN offer speed as compared to traditional R-CNN with higher accuracy, but they still failed to show desired output in real-time performance. Instead of selective search, it uses a neural network to propose the region. Unlike R-CNN variants, YOLO detects multiple objects simultaneously without using the pipeline.

The Multibox is similar to YOLO but draws multiple boxes around the ground truth to improve the accuracy instead of a selective search methodology. Still, it's just a part of the detection pipeline, unlike YOLO which is a fully developed detection system.

The Overfeat is a kind of image localization where the task is to predict the main object along with the boundaries in an image, unlike traditional methods where it predicts all objects in an image similar to Deformable Parts Models.

Multigroup is one of the grasp detection techniques where it detects only one object in an image and doesn't provide or classify other located objects and their boundaries. But YOLO detects multiple objects and their boundaries along with class probabilities

The author runs multiple experiments with real-time detectors and provided the results of each methodology. In the end, YOLO was tested in real-time, by connecting to a webcam.

## III. BACKGROUND AND MOTIVATION

The author's objective is to detect the objects simultaneously with their class probabilities using bounding boxes rather than seeing it just as a regression problem. The architecture of YOLO has a single convolutional neural network, trains on all images, and optimizes the performance of the detection with the associated class probabilities. YOLO doesn't require a complex pipeline to predict the test image and it has multiple benefits unlike traditional models which are having complex architecture, multiple pipelines and layers, having high latency in prediction, etc., For example, the author mentioned YOLO took less than 25 ms of latency in processing the video in real-time and achieved twice the mean average precision. The test image prediction is identified by simply multiplying the individual box predictions with the conditional probabilities:

$$\Pr(\text{Class}_i|\text{Object}) * \Pr(\text{Object}) * IOU_{pred}^{truth} = \Pr(\text{Class}_i) * IOU_{pred}^{truth}$$

In the end, the model was tested on multiple academic datasets, Picasso dataset, and People-Art Dataset. The author wants to introduce a unified model approach that was trained directly on the complete set of images in object detection with high precision rather than traditional classifier-based approaches.

#### IV. MAIN CONTRIBUTION

The major contribution of the author is specifying the individual components into a single neural network for simultaneous object detection in an image with low latency in real-time and evaluating using the dataset PASCAL VOC detection to predict both the coordinates and associated probabilities for the test images. Even though the network architecture is designed by the GoogLeNet model for image classification, the model mentioned in the paper used 1 X 1 reduction layers followed by  $3 \times 3$  convolutional layers.

In the model, the image was pre-trained for half the resolution and then the resolution for detection was doubled. The model is trained on the ImageNet 1000-class competition dataset, for approximately a week and achieved a top-5 accuracy of 88% on the ImageNet 2012 validation dataset, comparable to the GoogLeNet models in Caffe's Model Zoo. To avoid the model instability, considered the boxes that do not contain objects to decrease the loss from confidence predictions and increase it from bounding box coordinate predictions.

The experimental results were provided while compared to other models. In the end, to eliminate the fewer background mistakes by the YOLO model, Fast R-CNN was used in combination which boosts the performance of the final model. The results of individual and combined models are shown in Table 1.

	mAP	Combined	Gain
<b>Fast R-CNN</b>	71.8	-	-
<b>Fast R-CNN (2007 data)</b>	66.9	72.4	0.6
<b>Fast R-CNN (VGG-M)</b>	59.2	72.4	0.6
<b>Fast R-CNN (CaffeNet)</b>	57.1	72.1	0.3
<b>YOLO</b>	63.4	75.0	3.2

Table 1: Experiment results of authors proposed model

#### V. STRENGTHS

The authors proposed unified network has the following strengths:

- The YOLO model is a unified network i.e., build on a single convolutional neural network and considers the entire image for prediction.
- The proposed architecture by the author is simple and easy to implement.
- The test image has bounding boxes with associated probabilities.
- The model has low latency as compared to other traditional models.
- The YOLO model is implemented in real-time to detect multiple objects in a single image.
- The paper implemented a combined model (Fast R-CNN + YOLO) which boosts the performance.

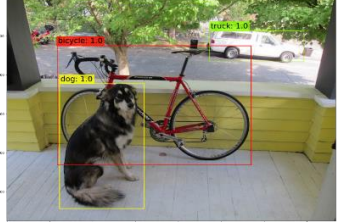
#### VI. DRAWBACKS

The authors proposed unified network has the following drawbacks:

- The model struggles when the objects are small in the image as the predictions considered only two boxes and assign one class.
- The architecture has multiple down sampling layers.
- The errors of small and large bounding boxes are the same which is unusual i.e., incorrect localizations.
- The Single Fast R-CNN achieved 66.9 mAP as compared to the YOLO model which has 63.4 mAP.

#### VII. DEMO MODEL

The demo model [2] is implemented for object detection using the YOLO algorithm, to detect images of the cat, city scene, dog, dog2, eagle, food, giraffe, horses, motorbike, person, surf, and wine. The pre-trained weights are taken from the COCO database and an open-source neural network framework darknet is used. The images are resized, and the first layer network has the input size of  $416 \times 416 \times 3$ . Non-Maximal Suppression (NMS) is set to 0.6, which keeps the best bounding box. The threshold of Intersection Over Union (IOU) is set to 0.4 to select the bounding boxes having the highest prediction ability. The model has low latency and detects the object in 4.351 seconds and the total detected objects is 3, Table 2 and Figure 1.



Object	Confidence Level
Dog	0.9999
Truck	0.9916
Bicycle	0.9999

Table 2: Experimental Results and Figure 1: Predicted class labels of demo model

#### VIII. CONCLUSION

The contribution of the author, developing a unified network for object detection with low latency, comparison of other object detection models, strengths and drawbacks of the proposed model, and performance review is discussed. In the end, a demo YOLO model is developed to detect the objects in real-time achieving the confidence interval of 0.99.

#### ACKNOWLEDGMENT

Sincere thanks to Dr. Xiaoyong Yuan for providing frequent feedback and Garima Nishad for providing demo model in github.

#### REFERENCES

- [1] Joseph Redmon, Santosh Divvala, Ross Girshick, Ali Farhadi. You Only Look Once: Unified, Real-Time Object Detection. <https://arxiv.org/abs/1506.02640>
- [2] Garima13a. YOLO Object Detection <https://github.com/Garima13a/YOLO-Object-Detection.git>