Implementation of YOLO V1 and comparison of YOLO V1 with Faster RCNN*

1^{st} Sachin Palahalli Chandrakumar
Artificial Intelligence
Northeastern University
Boston, United States
chandrakumar.s@northeastern.edu

***Abstract*—In this paper, we will discuss two object detection algorithms Yolo V1 and Faster RCNN. Faster RCNN is a two stage/proposal detection model and Yolo V1 is a single stage/proposal free detection model. Both the detection models have advantages and disadvantages and we will compare both the algorithms in this paper and delve deeper into the implementation of YOLO V1.**

***Index Terms*—object detection, faster rcnn, Yolo, real time, non maxima suppression**

I. Introduction

Object detection has some of the most prominent applications. How does self driving cars be able to detect whats in front and able to understand what to do based on what is present ahead. It is able to achieve its goals with the help of object detection. Object detection is a technique use to identify and locate objects in images or videos. The two main tasks of Object detection is Localization and Classification. Localization helps in determining where the objects are present in the given frame and classification helps in identifying which is the object present. Object detection plays a major role in improving safety and improving efficiency in industries , advancements in health care etc. The main tasks of object detection are Data Collection, Feature Extraction, Detecting the bounding boxes, predicting the labels, model evaluation. Object detection can be differentiated into Traditional Methods and Deep Learning Methods. The Deep Learning methods can further be differentiated into Region Proposal Based approaches, Single Stage neural and transformer based neural network. RCNN, Fast RCNN, Faster RCNN are Two stage methods, Yolo, SSD are single stage algorithms.

R-CNN achieves precise object localization and classification. It accuratley localizes and classifies objects in images. R-CNN was proposed in 2015 and Fast RCNN and FasterRCNN was proposed in 2015.

Faster RCNN is an improved version of RCNN and Fast RCNN. It optimizes RPN and Fat RCN stages in end to end approach. It is faster than both RCNN and Fast RCNN. The Short comings of RCNN based approaches and overcome using YOLO approach. RCNN approaches have lot of training time and is complex because of its two stage nature. This is overcome using single stage object detector using YOLO. Two stage methods are more accurate than single stage methods but is very slow and not feasible for real time object detection. YOLO is a single stage object detector with real time end to end system approach. It predicts both bounding boxes and class probabilities in one iteration.

In he following sections, we will analyze the architectures of RCNN, Faster RCNN, YOLO and analyze each other.

II. Related Work

There are many works done in the field of object detection. Object detection can be considered as either a regression or a classification model where regression model is used to predict bounding box values and classification model is used to determine the class of the objects. Object detection can be classified into Traditional methods and Deep Learning methods. One of the traditional based object detection approaches uses SIFT based approach. Deep learning methods can further be differentiated as Region Based approaches, Single stage neutral network and Transformer based. RCNN, Fast RCNN and all belong to Region based. YOLO, SSD belongs to Single stage neural network approach and DERT belongs to Transformer based approaches.

TABLE I
RObject Detection Approaches

Two Stage	One Stage
RCNN	YOLO
FAST RCNN	SSD
FASTER RCNN	
MASK RCNN	

III. Methods

A. R-CNN

R-CNN consists of two stages. Stage 1 is region proposal and stage 2 is Feature extraction.
Stage 1- Region Proposal:
It uses selective search to generate region proposals and segments images to potential regions. Each region is considered as separate image and is classified individually.
Stage 2- Feature Extraction, Classification and Localization:
Once regions are separated, features are extracted and object classification and bounding box regression is performed. Then SVM is applied using these features and regions are classified. It predicts the class probability and adjusts bounding box

coordinates for localization.
It uses cross-entropy loss for classify objects and L1 loss to predict the bounding box coordinates. The advantages of RCNN is that, its very accurate and robust. However, it is computationally expensive and not end to end. The improvement to this version was Fast RCNN and eventually Faster RCNN was implemented.

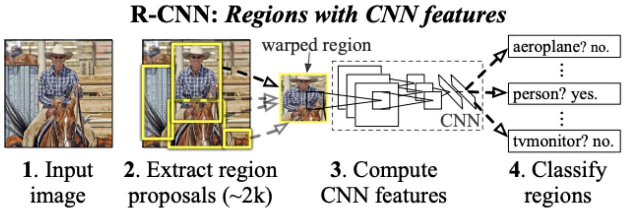

Fig. 1. RCNN Object Detection.

B. Faster R-CNN

Faster R-CNN introduces Region Proposal Network(RPN). This will avoid selective search. It uses different neural network and uses convolutions layers to generate region proposals. The features and results of RPN networks is input into Region of Interest pooling layer of Fast RCNN approach. The Loss function of Faster RCNN is calculated using two RPN loss and two Fast RCNN loss. RPN loss is calculated using Binary Classification loss and Bounding box Regression Loss. Fast RCNN loss is computed using Cross- entropy loss and L1 loss between predicted and ground truth boxes. All these losses are then linearly combined to get Faster R-CNN loss function.

The advantages of Faster RCNN is that it implements End to End training. It simultaneously optimizes RPN and Fast RPN stages end to end. It is highly accurate, faster and is independent of external models.

C. YOLO V1

Disadvantages of Using RCNN approach for object detection
It uses multi stage architecture and is very highly computationally expensive and takes very high time for training the models and is uses very complex architecture. The main reason for these disadvantages is because of the use of two stages used for object detection. They are more accurate but takes lot of time for prediction and hence might not be applicable where real time detection is necessary.
You Only Look Once(YOLO) was a new implementation for object detection which only uses single stage. It is fast and hence real time and performs end to end process of object detection on an input image. In one iteration, it determines the bounding boxes and class probabilities.

It splits the image into S*S grid and for each cell, it predicts B boxes and each bounding box 5 values: x-coordinate, y-coordinate, width, height and the object probability. Then, each grid predicts confidence of each category in that grid. Yolo algorithm image width, height and number of channels as input for training and outputs S*S*(B*5 + C) size tensor, where S*S is the total number of grids, B is the Bounding boxes defined per cell and C is the total number of Classes.

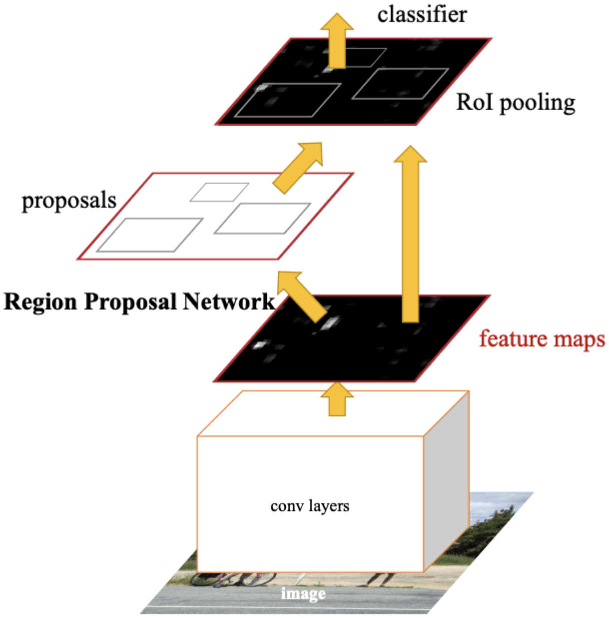

Fig. 2. Faster R-CNN is a single, unified network for object detection. The RPN module serves as the 'attention' of this unified network.

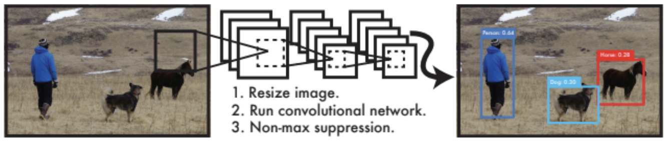

Fig. 3. The YOLO Detection System. (1)Yolo resizes the input image to 448 × 448, (2) runs a single convolutional network on the image, and (3) thresholds the resulting detections by the model's confidence.

Architecture of YOLO: Yolo architecture has 24 convolution networks which is followed by two fully connected layers Our detection network has 24 convolution layers followed by 2 fully connected layers. The images are pre-trained on Imagenet classification at half the resolution(224 *224) and detection is applied for double the resolution.

Yolo Detection approach:
The output obtained from the architecture is used to determine the final predictions. We first filter out bounding boxes with low probabilities og objects and the argmax over the distribution of probability of class given the object is considered to determine the object present in the bounding box considered. YOLO uses Anchor boxes to finalize the bounding boxes. Its a predefined bounding boxes of different aspect ratios and

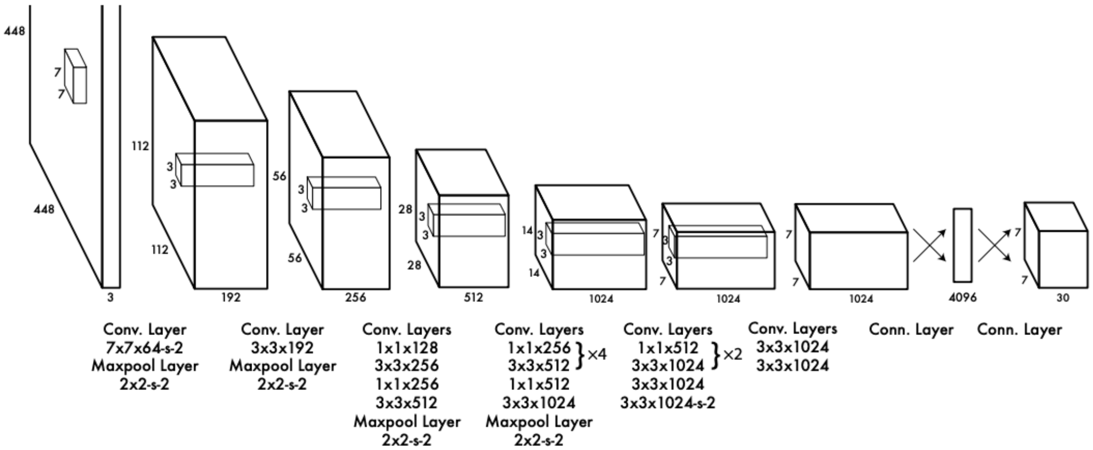

Fig. 4. Yolo detection network has 24 convolutional layers followed by 2 fully connected layers. Alternating 1 × 1 convolutional layers reduce the features space from preceding layers. It pretrains the convolutional layers on the ImageNet classification task at half the resolution (224 × 224 input image) and then double the resolution for detection.

scales. For predicting bounding boxes, uses combined anchor box and predicted bounding box offsets to determine bounding box values.

YOLO uses a new loss function which helps in detecting the objects accurately. It combines localization loss, Confidence loss, class loss and total loss f=to get the complete loss for the model.

- Localization loss analyzes accuracy of predicted bounding box coordinates by calculating the MSE of predicted and ground truth coordinate values.
- Confidence loss analyzes the confidence score of predicted bounding boxes. It analyzes the Intersection Over Union(IOU) region of the predicted and ground truth bounding box coordinate values.
- Class loss is used to determine the accuracy of classification of objects. It uses cross entropy loss to determine the class loss.
- Total loss combines all the above losses linearly with weight factors. This total loss helps in predicting the bounding boxes and objects accurately.

The model is evaluated using concepts like:

- Intersection over Union which measures the overlap between the predicted values and the ground truth values. It is given by Intersection area/ Union area.
- Precision-Recall curve which varies Intersection over Union for precision and recall values. Precision value is given by true positive/predicted positives and recall value is given by true positive/ ground truth positive.
- Average precision is given by the area under precision recall curve. It analyzes performances of precision and recall values
- mean Average Precision averages Average precision over all classes.

IV. RCNN VS YOLO

- YOLO uses single stage approach where it predicts and detects in one single pass where as RCNN uses two stages to process and detect the objects.
- Yolo uses grid based approach to select bounding box where as Faster RCNN uses Region proposal network to determine the bounding boxes.
- Yolo is a simple architecture but Faster RCNN has a complex architecture.
- YOLO is faster because of its architecture and single stage detection where as RCNN is slower because of two stage approach.
- YOLO is less accurate as it tries to be more faster and RCNN approaches are highly accurate
- YOLO can be used real time applications like self driving etc where as RCNN cane be used where high accuracy is critical like in medical fields.

TABLE II
RCNN VS YOLO

YOLO	RCNN
Single stage process	Multi Stage Process
Grid based Bounding box prediction	RPN based bounding box prediction
Simple approach	Complex approach
Faster	Slower
Less accurate	More accurate
Real time applications	High accuracy tasks

V. EXPERIMENTS AND RESULTS

For this project, I have implemented Yolo V1 architecture and custom loss function used for object prediction in real time. I have considered PascalVOC dataset. It contains 20 classes with classes like dogs, cars , chairs etc. Each image is annotated with bounding box values of objects present and the class label.

Custom loss function is implemented by calculating localization loss by calculating the MSE of predicted and ground truth of bounding box coordinate values. Confidence loss is calculated by analyzing the confidence score of predicted bounding boxes. The final total loss is calculated by combines all the above losses linearly with weight factors $lambdacoord$ and $lambdanoobj$

VI. Summary

We can clearly see that both the models are used for different use cases and when there is high accuracy needed , we can go for RCNN based approaches, where as real time detection is critical, we can use YOLO approach. RCNN uses two stages to detect the objects and bounding box coordinates whereas YOLO uses one stage for both detection of bounding boxes and classification of objects. YOLO uses a specific loss function which is critical for detection process and its performance. There is another method called Fast YOLO, which is used for faster object detection. YOLO is also ideal for detection new domains and is state of the art in real time object detection. There have been multiple upgrades to YOLO model to make it more efficient, faster and accurate. The latest version is Yolov9 which was released in February 2024.

References

I have referred [1] to implement the Yolo architecture and the custom loss function. I have also utilized some of the util function code of the actual Yolo implementation in my implementation as they were standard way of calculating the values. I have also refereed to [2] and [3] to analyze and understand RCNN methods. The images reference are taken by the original research papers published which are [1] [2] [3]

References

[1] Joseph Redmon, Santosh Divvala, Ross Girshick, Ali Farhadi, ”You Only Look Once: Unified, Real-Time Object Detection”

[2] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun, ”Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks”

[3] Ross Girshick, Jeff Donahue, Trevor Darrell, Jitendra Malik, “Rich feature hierarchies for accurate object detection and semantic segmentation,”