YOLO3

This paper briefs about the changes made to YOLO to make it more effective. Here, a bigger and more accurate neural network is trained which also has increased speed. They have even trained a better classifier.

In YOLO500, the bounding box coordinates, tx, ty, tw, th are outputted and the ground value truth for some co-ordinate is taken as t^\* and the actual prediction is taken as t\*. And hence the gradient descent is calculated as t^\*-t\*.

In YOLO-v3, each bounding box output is done by a logistic classifier which gives value 1 only if the bounding box covers ground truth object more than prior bounding boxes. If there was no prior bounding box, there is no loss. Each box contains the class the bounding box is containing. Softmax classification is not used here because softmax would not work when the bounding box has overlapping labels because softmax assumes each box to have only one class. Hence multi-level approach is seen fit for this.

YOLOv3 predicts boxes of 3 scales. The feature map is taken from previous two layers of neural network and also from earlier part of neural network and is combined with unsampled features. Few more layers of neural network are added to process the resultant feature map. The same design is used for another scaling. K-means clustering is used for prior bounding box prediction where 9 clusters are used.

A new network is taken which is a mix of YOLOv2 and DarkNet-19. It has successive 3X3 and 1X1 convolutional layers. It has around 53 layers and hence is called DarkNet-53. Dark-net is more efficient and faster than ResNet-101. It even utilizes the GPU better. Its almost as efficient as Retina-Net. It has a very good accuracy in detection and is one of the strongest neural network for detection. During training, data augmentation, batch normalization, multi-scale training is used. Its seen that when the threshold for IOU increases, the YOLOv3 performance decreases. In the past YOLO has had difficulties in detecting small objects whereas in YOLOv3 the machine struggles with big images.

A few of the approaches were tried in this research which did not yield good results. Some of them included using linear activation for prediction of the bounding box co-ordinates x and y instead of logistic activation. Another approach which failed was predicting the offset co ordinates as multiples of height or width of bounding box. In faster CNNs, two IOU thresholds are considered. If the overlapping with ground truth is above 0.7, it is considered positive prediction and if it is less than 0.3 , it is considered negative prediction. The author tried implementing this approach but it did not work out in their favour.

At the end, the author expresses their concern over the use of the CV technology. They fear the technology might be used for bad cause and wants the researches to take responsibility of avoiding this.