Task : 08

YOLO-

You only look once -- V2

YOLO is one of the best models in object recognition, able to recognize objects and process frames at the rate up to *150 FPS* for small networks. At *67 FPS*, YOLOv2 gives mAP of 76.8% and at *67 FPS* it gives an mAP of *78.6%* on VOC 2007 dataset better than the models like *Faster R-CNN* and *SSD*. YOLO 9000 used YOLO v2 architecture but was able to detect more than 9000 classes.

Architecture Changes vs YOLOv1:

The previous YOLO architecture has a lot of problems when compared to the state-of-the-art method like Fast R-CNN. It made a lot of localization errors and has a low recall. There are some incremental improvements that are made in basic YOLO these changes below:

1. Batch Normalization:

   By adding batch normalization to the architecture can increase the convergence of the model that leads to faster training. And this removes Dropout without overfitting. Also Increases in mAP by *2%* as compared to basic YOLO.

2. High Resolution Classifier:

   After training by 224×224 images, YOLOv2 also uses 448×448 images for fine-tuning the classification network for 10 epochs on ImageNet.

   4% increase in mAP.

3. Use Anchor Boxes For Bounding Boxes:

   YOLOv2 removes all fully connected layers and uses anchor boxes to predict bounding boxes

4. Convolutions with Anchor Boxes:
    a. YOLOv2 removes all fully connected layers and uses anchor boxes to predict bounding boxes.
    b. One pooling layer is removed to increase the resolution of output.
    c. And 416×416 images are used for training the detection network now.
    d. And 13×13 feature map output is obtained, i.e. 32× downsampled.
    e. Without anchor boxes, the intermediate model got 69.5% mAP and recall of 81%.
    f. With anchor boxes, 69.2% mAP and recall of 88% were obtained. Though mAP is dropped a little, recall is increased by a large margin.

5. Dimension Clusters
    a. The sizes and scales of Anchor boxes were pre-defined without getting any prior information, just like the one in Faster R-CNN.
    b. Using standard Euclidean distance-based k-means clustering is not good enough because larger boxes generate more error than smaller boxes
    c. YOLOv2 uses k-means clustering which leads to good IOU scores.
    d. k = 5 is the best value with a good tradeoff between model complexity and high recall. Direct Location Prediction
    e. YOLOv1 does not have constraints on location prediction which makes the model unstable at early iterations. The predicted bounding box can be far from the original grid location.
    f. YOLOv2 bounds the location using logistic activation σ, which makes the value fall between 0 to 1

6. Fine-Grained Features
    a. The 13×13 feature map output is sufficient for detecting large objects.
    b. To detect small objects well, the 26×26×512 feature maps from the earlier layer are mapped into 13×13×2048 feature maps, then concatenated with the original 13×13 feature maps for detection.
    c. 1% increase in mAP is achieved.

7. Multi-Scale Training
    a. For every 10 batches, new image dimensions are randomly chosen.
    b. The image dimensions are {320, 352, …, 608}.
    c. The network is resized and continues training.

Training:

The YOLOv2 is trained for two purposes:

1. For classification tasks the model is trained on ImageNet-1000 classification task for 160 epochs with a starting learning rate 0.1, weight decay of 0.0005 and momentum of 0.9 using Darknet-19 architecture. There are some standard Data augmentation techniques applied for this training.
2. For detection there are some modifications made in the Darknet-19 architecture which we discussed above. The model is trained for 160 epochs on starting learning rate 10-3 , weight decay of 0.0005 and momentum of 0.9. The same strategy is used for training the model on both COCO and VOC.