Computer Vision - Assignment 9

1. What exactly are region proposals ?

Region Proposals simply means regions with **Convolutional Neural Networks (CNN)**. It tries to pick just a few regions of the image that make sense to run our **ConvNet classifier**. So, rather than running our **sliding windows** on every single window we instead select just a few windows and run our ConvNet classifier on those a few windows only. The way that R-CNN performs the region proposals is to run an algorithm called as **Segmentation Algorithm.**

2. What exactly do you mean when you say "Non-Maximum Suppression (NMS)" ?

Whenever in an image, for an object if we are getting multiple detections of that same object then in that scenario we use **Non-Maximum Suppression** which is a way to make sure that our algorithm detects each object only once.

When we run our object detection algorithm on every grid cells of an image then there is always a possibility that many of them will encounter the object component in them. So, we might end up with multiple detections for each object present in that image. Here in this case **Non-Maximum Suppression** will clean up these multiple detections and in the end the algorithm will detect only one detection per class rather than multiple detections per class. Suppose in an image there is a car and our algorithm is detecting 3 bounding boxes for that car associated with probabilities or confidence score of 0.75, 0.8, & 0.9 then our algorithm will take into consideration the largest probability which is 0.9. Now, **Non-Maximum Suppression** comes into picture and looks at all the remaining bounding boxes those who are having maximum overlap or high **Intersection Over Union (IOU)** with 0.9 probability and those will get suppressed in this case 0.75 & 0.8 will get suppressed. So, this is **Non-Maximum Suppression** in which the object detection algorithm will only output the maximal probabilities classifications & localizations of multiple objects present in an image by suppressing the close ones that are non-maximum hence the name **Non-Maximum Suppression.** If there are multiple classes in our object detection algorithm then you have to independently carry out **Non-Maximum Suppression** one on each of the output classes.

3. What is mAP, exactly ?

As the name directly suggest **mAP (Mean Average Precision)** is the mean or average of average precision. **Average Precision (AP)** is the **area** under the **precision recall curve**. For getting mAP first calculate AP for each of the class and then average with all the other class. mAP is the most commonly used evaluation metrics in object detection. Accuracy of the model is based on mAP. Higher the mAP better the object detection model is. mAP is always between 45 to 90 but never 100.

4. What is the definition of a Frame Per Second (FPS) ?

Frames per second (FPS) is a term or metric used in capturing the live or static video and also during video playback. So, FPS is a unit that measures display device performance in video captures and playback. **FPS** is used to measure **frame rate**. One frame is a single image and the number of images consecutively displayed each second is the video or FPS. If FPS is small then the video generated will have small size and vice-versa. The greater the FPS, the smoother the video motion appears. Full-motion video is usually 24 FPS or greater. Different video formats have different FPS rates.

5. What exactly do you mean when you say IOU (Intersection Over Union) ?

For **Evaluating Object Localization** which is also called as **Evaluating Confidence Score** the concept of **IOU (Intersection Over Union)** also known as **Jaccard Index** is used. When we do annotations for our dataset at that time we generate the coordinates of bounding boxes of our each class for all the images present in our dataset. Those bounding boxes are nothing but the ground truth bounding boxes. When we run our object detection algorithm the predicted bounding boxes are plotted over the objects. Now there will be so many predicted BB for each object in an image but our algorithm should pick only the best one. So, **IOU (Intersection Over Union)** is used. The higher the IOU the more accurate is the BB. In lot of computer vision task the IOU threshold is set to 0.5 but if we want very robust object detection we could vary or increase the IOU threshold to 0.75 or even more depending upon the use case.

$$IOU = \frac{\textit{Size of the Intersection of predicted BB over ground truth BB}}{\textit{Size of the Union of predicted BB over ground truth BB}}$$

So, this is one way to map localization to accuracy where we just count up the number of times an algorithm correctly detects and localizes an object. Moreover, generally IOU is a measure of the overlap between the two bounding boxes.

6. Describe the Curve Of Precision-Recall (PR Curve) ?

Precision-Recall Curve is a way to interpret or evaluate the performance of binary (two-class) classification predictive models. PR curve with an area under the curve scores can directly be used to compare classification models.

The metrics that make up the precision-recall curve is defined in terms of the cells in the confusion matrix. The confusion matrix provides more insight into not only the performance of a predictive model, but also which classes are being predicted correctly, which incorrectly, and what type of errors are being made. The simplest confusion matrix is for a two-class classification problem, with negative (class 0) and positive (class 1) classes.

<table>
<tr><td></td><td>|</td><td>Positive Prediction (1)</td><td>|</td><td>Negative Prediction (0)</td></tr>
<tr><td>Positive Class (1)</td><td>|</td><td>True Positive (TP)</td><td>|</td><td>False Negative (FN)</td></tr>
<tr><td>Negative Class (0)</td><td>|</td><td>False Positive (FP)</td><td>|</td><td>True Negative (TN)</td></tr>
</table>

Precision is a metric that quantifies the number of correct positive predictions made. The ability of the model to tell how much accurate the model is. Precision is between 0 (no precision) to 1 (full or perfect precision).

$$\textbf{Precision} = \frac{\textbf{True Positives}}{(\textbf{True Positives} + \textbf{False Positives})}$$

Recall is a metric that quantifies the number of correct positive predictions made out of all positive predictions that could have been made. Recall is also between 0 (no recall) to 1 (full or perfect recall).

$$\textbf{Recall} = \frac{\textbf{True Positives}}{(\textbf{True Positives} + \textbf{True Negatives})}$$

A precision-recall curve (or PR Curve) is a plot of the precision (y-axis) and the recall (x-axis) for different probability thresholds.

A model with perfect skill is depicted as a point at a coordinate of (1,1). A skillful model is represented by a curve that bows towards a coordinate of (1,1). A no-skill classifier will be a horizontal line on the plot with a precision that is proportional to the number of positive examples in the dataset. For a balanced dataset this will be 0.5. The focus of the PR curve on the minority class (positive) makes it an effective diagnostic for imbalanced binary classification models. A precision-recall curve can be calculated in scikit-learn using the *precision_recall_curve() function* that takes the class labels and predicted probabilities for the minority class and returns the precision, recall, and thresholds.

The Precision-Recall AUC is just like the ROC AUC, in that it summarizes the curve with a range of threshold values as a single score. The score can then be used as a point of comparison between different models on a binary classification problem where a score of 1.0 represents a model with perfect skill.

7. What does "Selected Search" means ?

Selective Search is an algorithm used for Object Detection in R-CNN family.

The problem of object localization is the most difficult part of object detection. One approach is that we use **sliding window** of different size to locate objects in the image. This approach is called **Exhaustive Search**. This approach is computationally very expensive as we need to search for object in thousands of windows even for small image size. Some optimization has been done such as taking window sizes in different ratios (instead of increasing it by some pixels). But even after this due to number of windows it is not very efficient. **Selective Search** algorithm uses both Exhaustive Search and Segmentation. Segmentation is a method to separate objects of different shapes in the image by assigning them different colors.

Algorithm Of Selective Search :

1. Generate initial sub-segmentation of input image using the method of “Efficient Graph-Based Image Segmentation “.

2. Recursively combines the smaller similar regions into larger ones. We use Greedy algorithm to combine similar regions to make larger regions.

- Greedy Algorithm :
  1. From set of regions, choose two that are most similar.
  2. Combine them into a single, larger region.
  3. Repeat the above steps for multiple iterations.

3. Use the segmented region proposals to generate candidate object locations.

- **Selective Search In Object Recognition :**

Using this algorithm for object detection and training a model using by giving ground truth examples and sample hypothesis that overlaps 20-50% with ground truth (as negative example) into SVM classifier and train it to identify false positive.

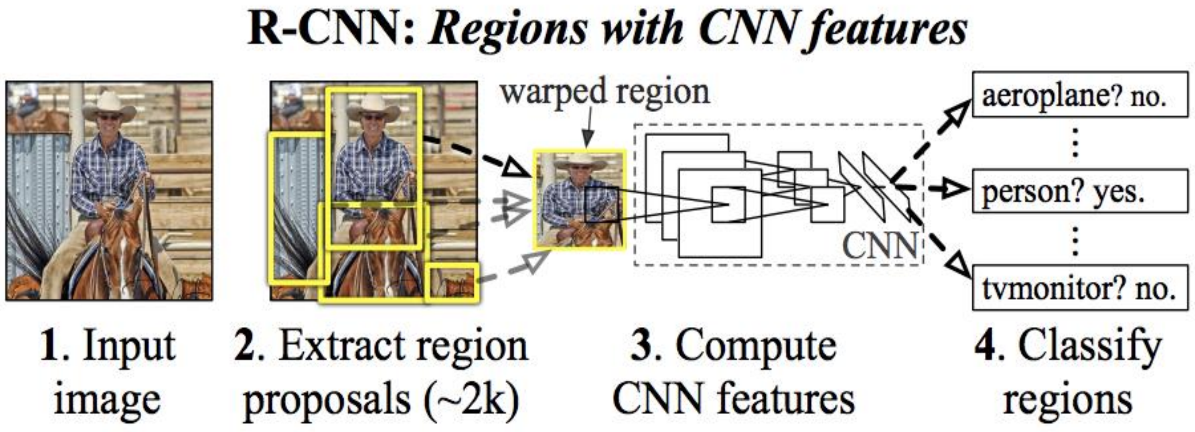
- **Applications & Drawbacks:**

Selective Search is widely used in early SOTA such as R-CNN, Fast R-CNN etc. However, Due to number of windows it processed, it takes anywhere from 1.8 to 3.7 seconds (Selective Search Fast) to generate region proposal which is not good enough for a real-time object detection system.

8. Describe the four components of the R-CNN model ?

R-CNN: To bypass the problem of selecting a huge number of regions, R-CNN method was adopted where we use selective search to extract just 2000 regions from the image and those regions were called as region proposals. Therefore, now, instead of trying to classify a huge number of regions, you can just work with 2000 regions. These 2000 region proposals are generated using the selective search algorithm.

Selective Search:
1. Generate initial sub-segmentation, we generate many candidate regions.
2. Use greedy algorithm to recursively combine similar regions into larger ones.
3. Use the generated regions to produce the final candidate region proposals.



Four Components of the R-CNN model

These 2000 candidate region proposals are warped into a square and fed into a Convolutional Neural Network that produces a 4096-dimensional feature vector as output. The CNN acts as a feature extractor and the output dense layer consists of the features extracted from the image and the extracted features are fed into an SVM to classify the presence of the object within that candidate region proposal. In addition to predicting the presence of an object within the region proposals, the algorithm also predicts four values which are offset values to increase the precision of the bounding box. For example, given a region proposal, the algorithm would have predicted the presence of a person but the face of that person within that region proposal could've been cut in half. Therefore, the offset values help in adjusting the bounding box of the region proposal.

9. What is the Localization Module, exactly ?

Object Detection & Recognition = Image Classification + Object Localization (Bounding Box) with Labels

Object localization refers to identifying the location of one or more objects in an image and drawing abounding box around their extent. Object detection combines these two tasks and localizes and classifies one or more objects in an image. For instance, In ADAS we need to detect not only just other vehicles but also may be pedestrians, trees, traffic signs & signals and may be even other objects of different categories. In object detection the training set contains probability of class present (0 or 1), 4 additional numbers ((bx, by) as midpoint of object & (bh, bw) height and width of the bounding box) giving the bounding box, and object class labels (suppose 3 classes 1 to 3). Therefore, in this case the output vector 'y' is 8-d which contains 8 components.

10. What are the disadvantages of R-CNN ?

Disadvantages with R-CNN :

- It still takes a huge amount of time to train the network as you would have to classify 2000 region proposals per image. R-CNN is very slow algorithm.
- It cannot be implemented real time as it takes around 47 seconds for each test image.
- The selective search algorithm is a fixed algorithm. Therefore, no learning is happening at that stage. This could lead to the generation of bad candidate region proposals.