Module -5 Region Based CNN - Syllabus

- Encoder-Decoder Models,
- Attention approaches,
- **RCNN,**
- Yolo and its versions
- Data Collection,
- Image labeling and Training.
- Build Custom models,
- Comparative analysis.
- Various Applications

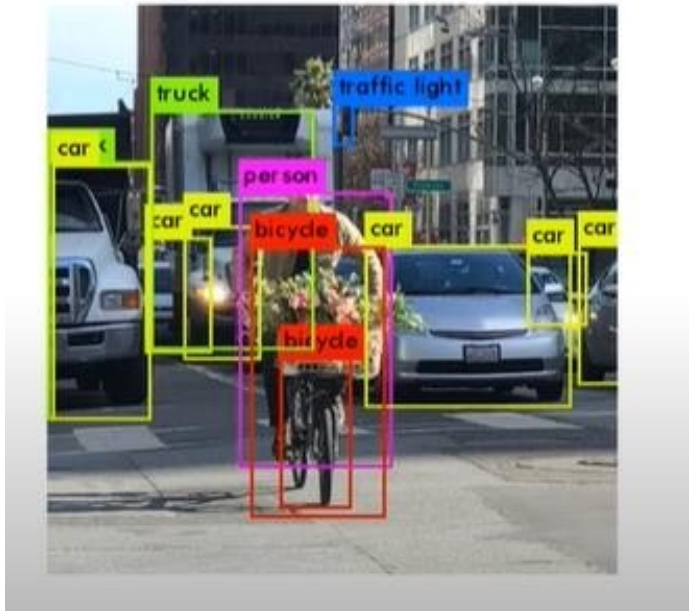
Module 5

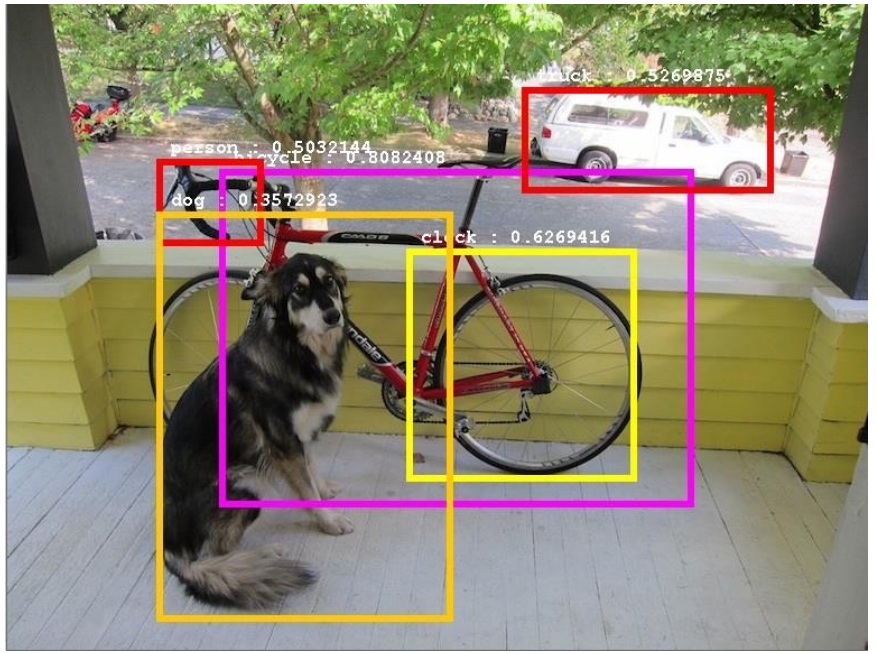
Region-based CNNs

Lecture-2

Region Based CNN

- Deep learning models are only able to tell whether an image contains an object or no.
- Suppose we want to work on models that could also tell where that object is, in an image?
- ***Object Detection:*** Predicting bounding boxes of multiple objects of multiple classes. May contain many objects belonging to the same class as well.





- **<u>Image Localization</u>**:
- Predicting bounding box of only a single object, of a single class, in an image
- In classification algorithms, the final layer gives a probability value ranging from 0 to 1.
- In contrast, localization algorithms give an output in four real numbers, as localization is a regression problem.

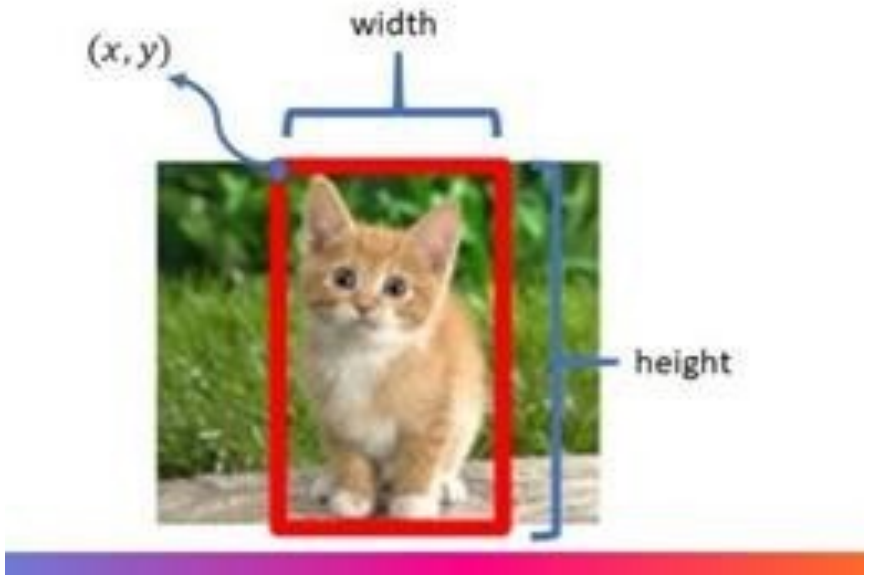
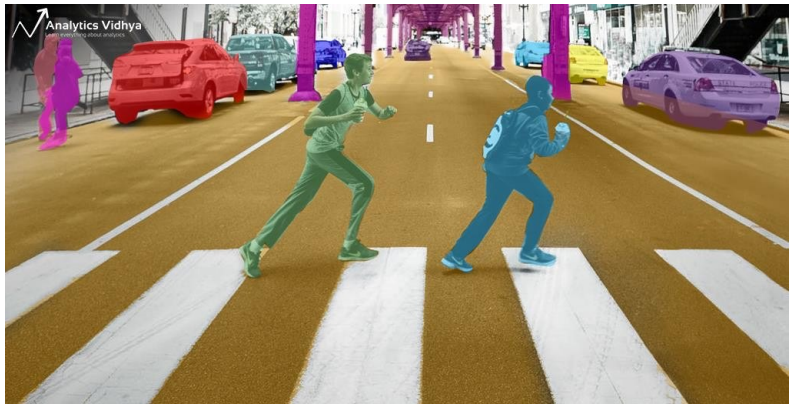


Image Segmentation:

- To create a pixel-wise mask for each of the objects in an image.



Sliding window approach

- Many of the sub-section of images selected by the window won't be detecting any objects
- This will not work for a class of objects
- Unknown position/scale/aspects of objects could not be traced Can a hint be given on the places where there are objects and tell that whether the object is there in that specific location?

Solution : R-CNN

R-CNN

- The goal of R-CNN is to take in an image, and correctly identify where the main objects (via a bounding box) in the image.
- **Inputs**: Image
- **Outputs**: Bounding boxes + labels for each object in the image.
- R-CNN creates these bounding boxes, or region proposals, using a process called Selective Search

What is Region based CNN?

•**Region-based Convolutional Neural Networks (R-CNN)** are a family of machine learning models for computer vision and specifically object detection.

How region based CNN works?

Region based CNN consists of three modules —

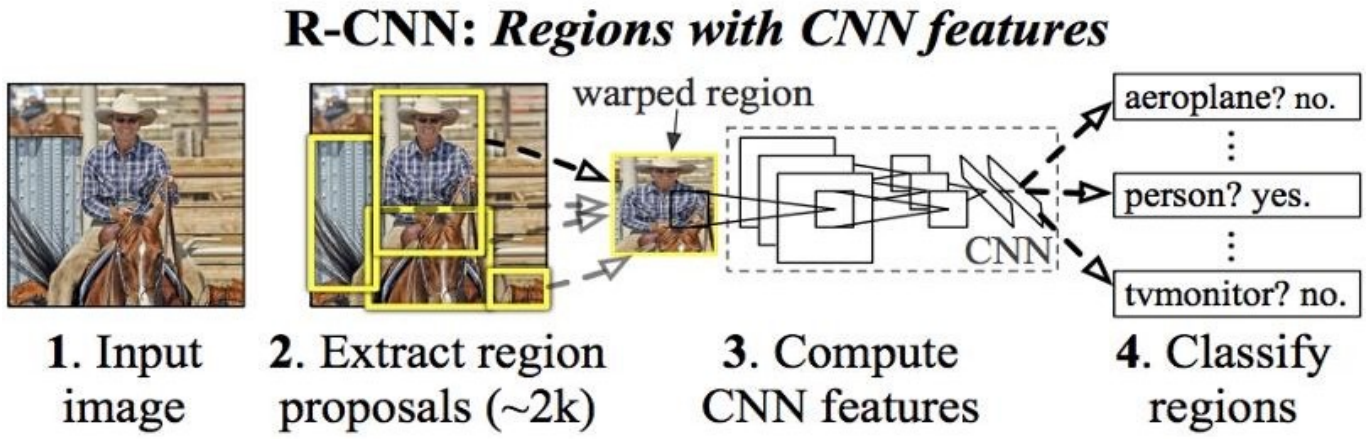
- ❖ Region Proposal
- ❖Feature Extractor
- ❖ Classifier

Region Proposal : **When an input image is given region proposal tries to detect different regions (~2000) in different sizes and aspect ratios**. In other words, it draws multiple bounding boxes in input image as shown below.

Why Region based CNN are used in Object Detection?????

Region-based Convolutional Neural Networks (R-CNN)

Object detection consists of two separate tasks that are classification and localization. R-CNN stands for Region-based Convolutional Neural Network. The key concept behind the R-CNN series is region proposals. Region proposals are used **to localize objects within an image**

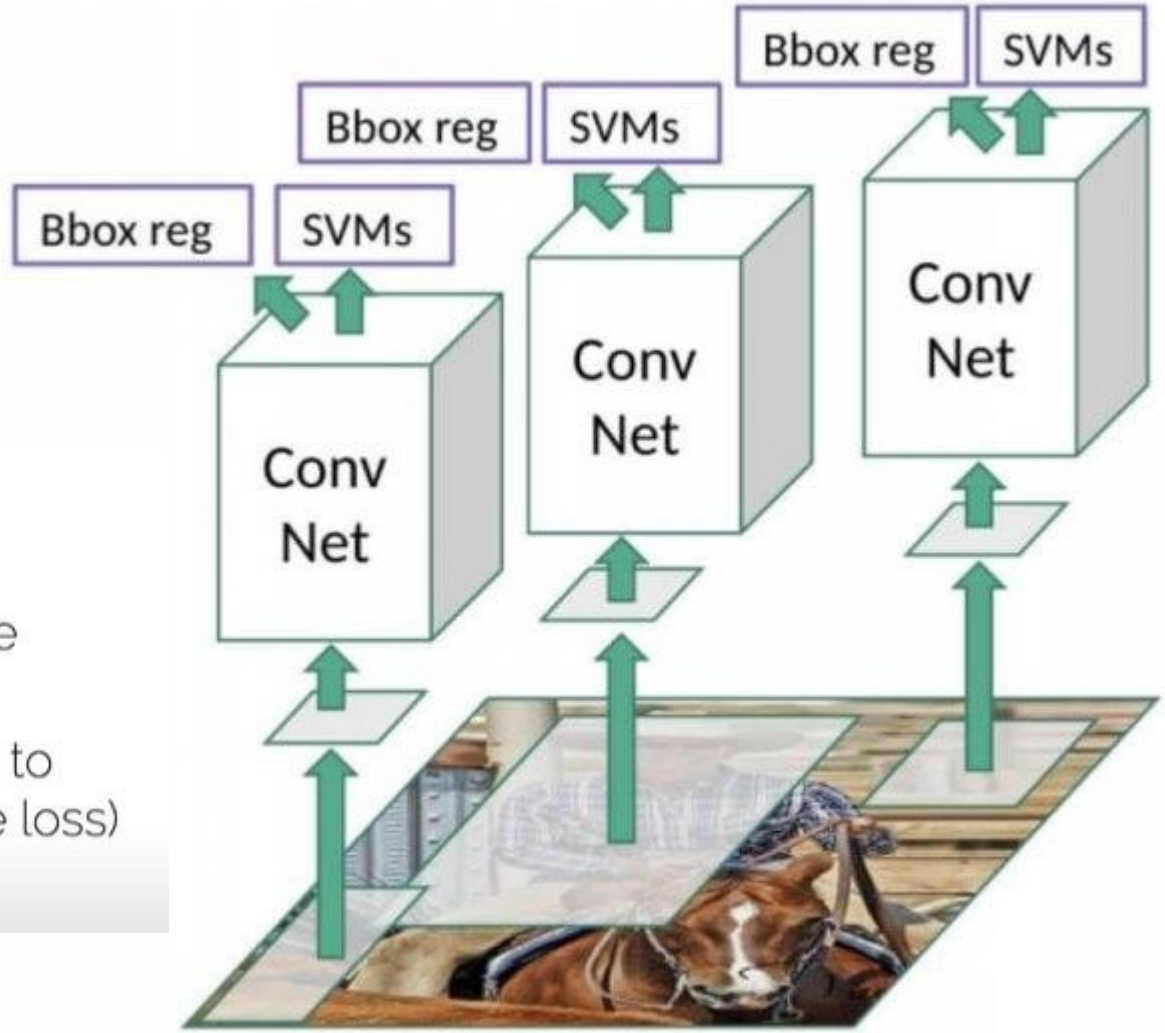


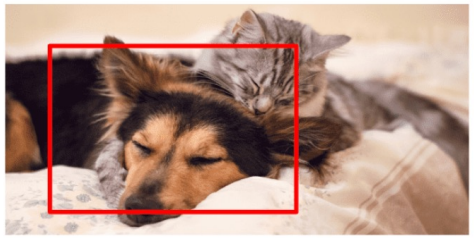
Steps: RCNN

1.Perform *selective search* to extract multiple high-quality region proposals on the input image (Uijlings *et al.*, 2013). These proposed regions are usually selected at multiple scales with different shapes and sizes. Each region proposal will be labeled with a class and a ground-truth bounding box.
2.Choose a pretrained CNN and truncate it before the output layer. Resize each region proposal to the input size required by the network, and output the extracted features for the region proposal through forward propagation.
3.Take the extracted features and labeled class of each region proposal as an example. Train multiple support vector machines to classify objects, where each support vector machine individually determines whether the example contains a specific class.
4.Take the extracted features and labeled bounding box of each region proposal as an example. Train a linear regression model to predict the ground-truth bounding box.

R-CNN ARCHITECTURE

- Training scheme:
  - 1. Pre-train the CNN on ImageNet
  - 2. Finetune the CNN on the number of classes the detector is aiming to classify (softmax loss)
  - 3. Train a linear Support Vector Machine classifier to classify image regions. One SVM per class! (hinge loss)
  - 4. Train the bounding box regressor (L2 loss)





Bounding Box Regressor [Class = *Dog*]

Convolutional Network

4096-dimensional feature vector

SVM [Class = *Bird*] — 0.17

SVM [Class = *Dog*] — 0.87

...

SVM [Class = *Cat*] — 0.42

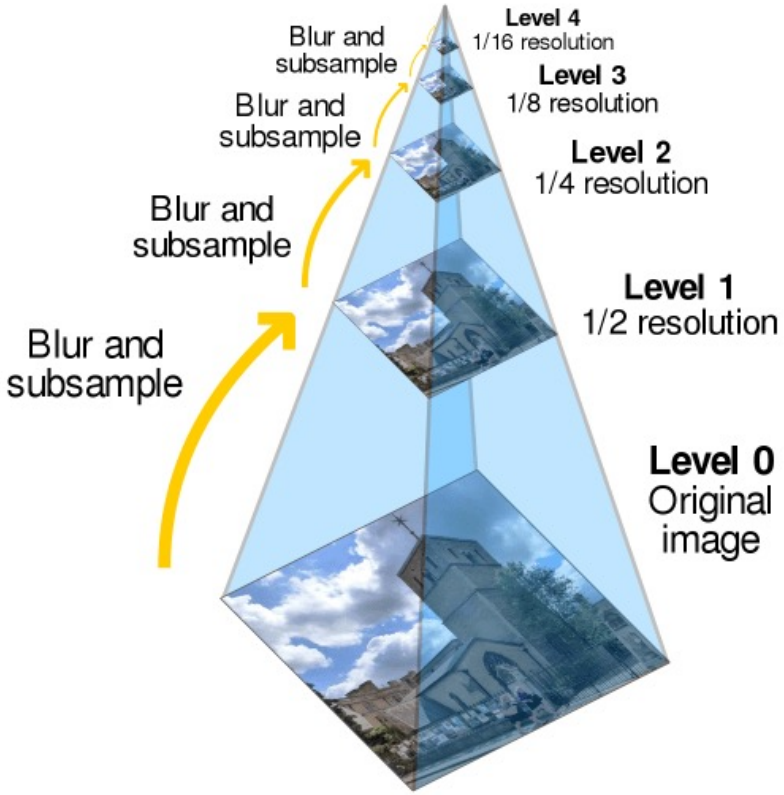
Region Proposals- approaches

- Sliding Windows
- Colour Contrast
- Edge Boxes
- Super Pixel Straddling
- Selective Search, etc.
- Extracting Region Proposals is the process of sampling various cropped regions of the image with an arbitrary size which may or may not have the possibility of the object being inside the cropped region.

- At a high level, Selective Search looks at the image through windows of different sizes, and for each size tries to group together adjacent pixels by texture, color, or intensity to identify objects.
- Once the proposals are created, R-CNN warps the region to a standard square size and passes it through to a modified version of AlexNet (the winning submission to ImageNet 2012 that inspired R-CNN), as shown in the previous slide

Image pyramid-example

- The concept of image pyramids has also been employed in deep learning for feature extractions.
- The idea behind this is that features that may go undetected at one resolution can be easily detected at some other resolution.
- For instance, if the region of interest is large in size, a low-resolution image or coarse view is sufficient.
- While for small objects, it's beneficial to examine them at high resolution.
- Now, if both large and small objects are present in an image, analyzing the image at several resolutions can prove beneficial.



Limitations

- Extremely slow
- Different values might lead to different results

Selective Search-Algorithm

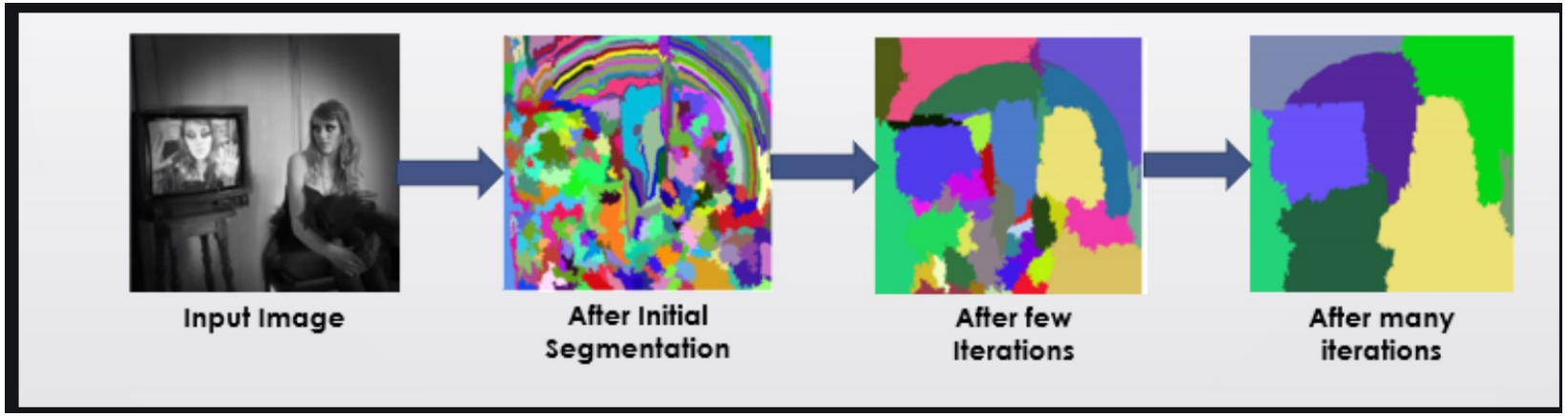
1.Generate initial sub-segmentation of input image using the method describe by *Felzenszwalb et al* in his paper “Efficient Graph-Based Image Segmentation

Selective Search-Algorithm

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

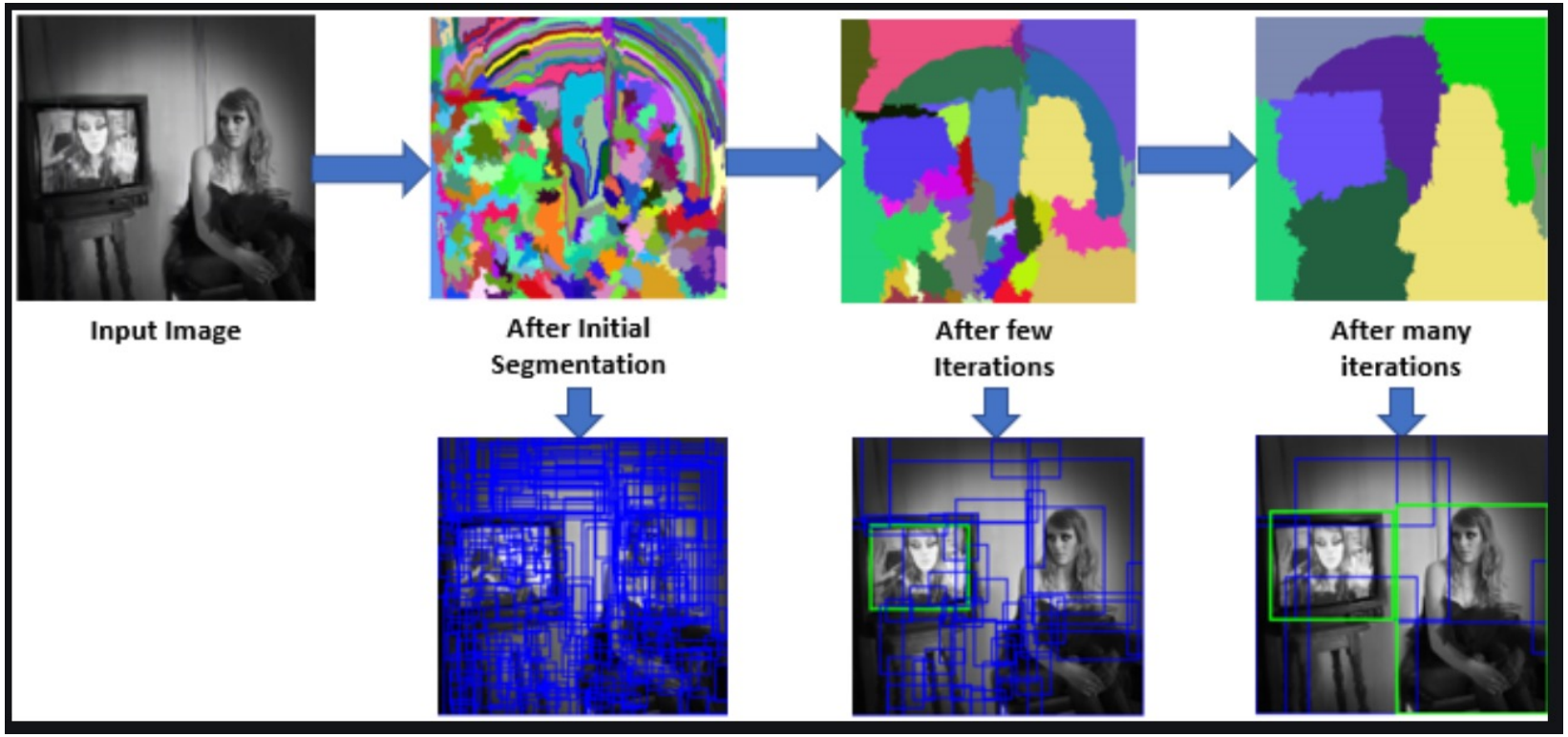
Greedy Algorithm :

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.



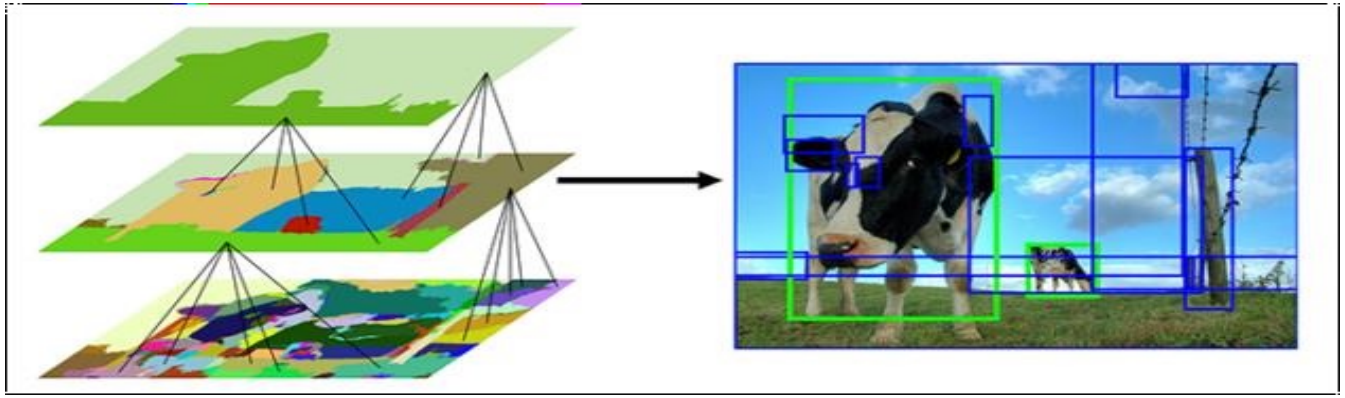
Selective Search-Algorithm

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



The selective search paper considers four types of similarity when combining the initial small segmentation into larger ones.

- Selective Search merges superpixels in a hierarchical fashion based on five key similarity measures:
- Color
- Texture
- Size
- Shape similarity
- A final meta-similarity

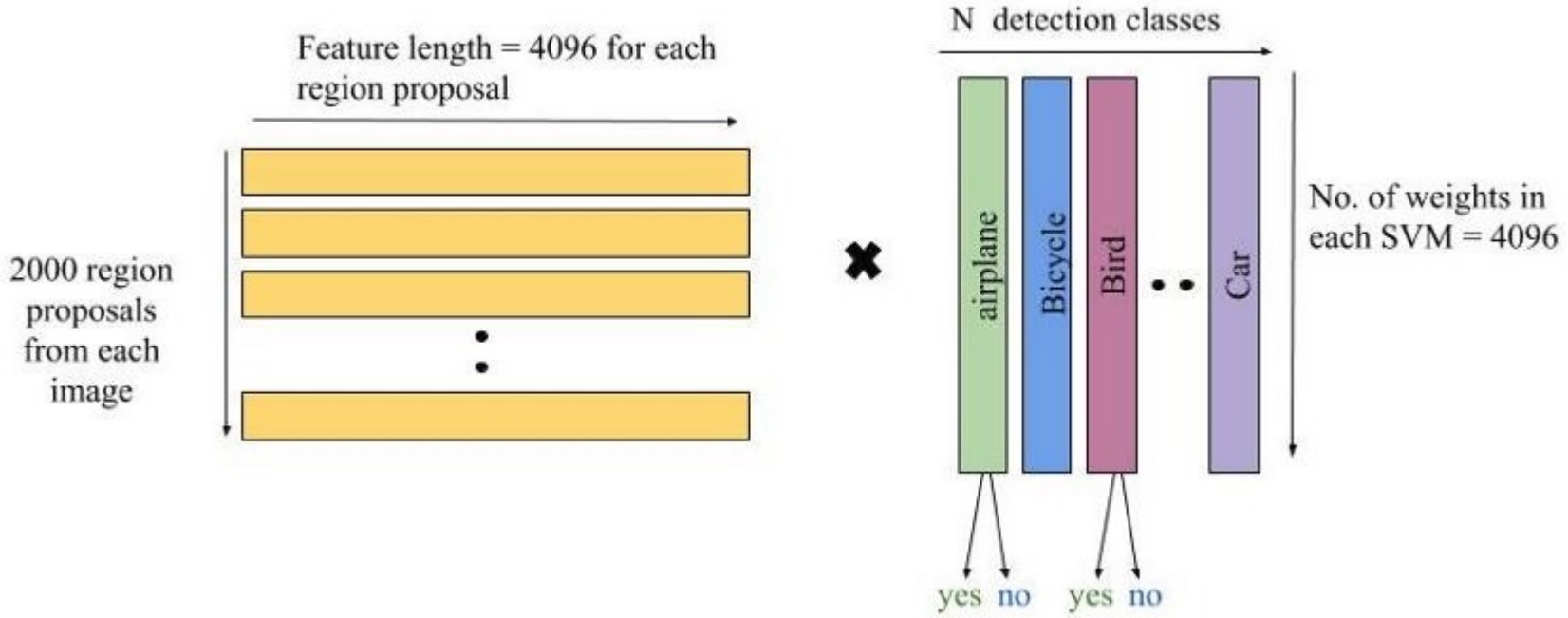


1.Color similarity (color histograms)
2.Texture similarity (gradients)
3.Size similarity (favors smaller regions merging first)
4.Shape Compatibility (how well regions fit together when combined)

Feature Extraction phase

- AlexNet was used as a pretrained network (which was popular then) to generate a 4096-dimensional vector output for each of the 2000 region proposals.
- we can use the Pre-trained AlexNet by removing the last softmax layer for generating the feature vectors, and then fine- tune the CNN for our distorted images and the specific target classes.
- The labels used are the ground-truths with the maximum IoU (Intersection over Union) which are in the positive category, the rest others are negative labels (for all classes).
- So, the output from this Feature Extraction Phase is a 4096-dimensional feature vector.

- Final layer SVM is used for each object class
- For each feature vector we have n outputs where n is the total number of classes we are considering and the actual output is the confidence score.
- Based on the highest confidence score we can make the inference of Object Class(es) given in a particular image.

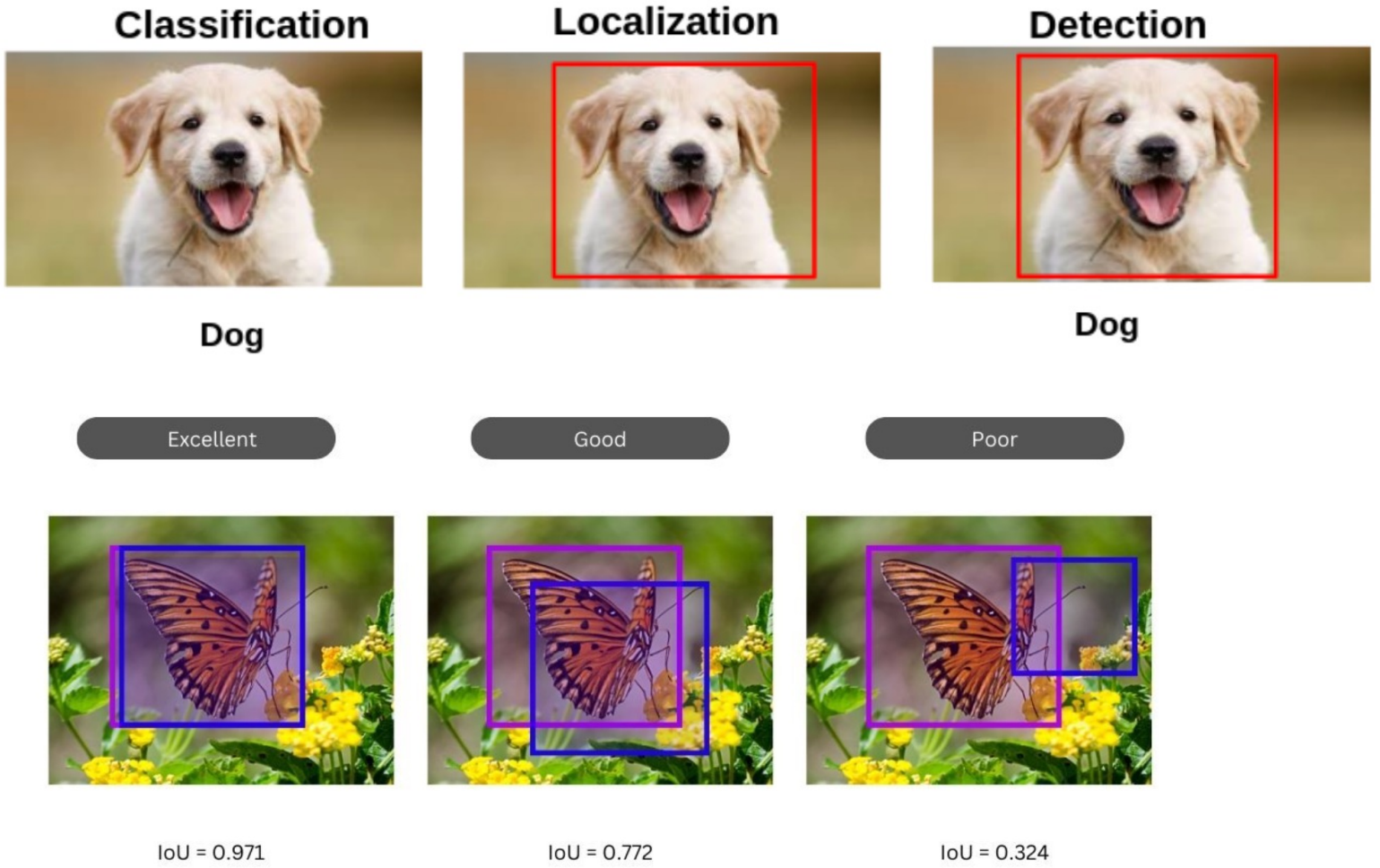


Object Detection

There are various algorithms for object detection tasks.

To improve the performance further, and capture objects of different shapes and sizes, the algorithms predict multiple bounding boxes, of different sizes and aspect ratios.

how to select the most appropriate and accurate bounding?



Non-Max Suppression(NMS)

suppress the ones that are not maximum (score)

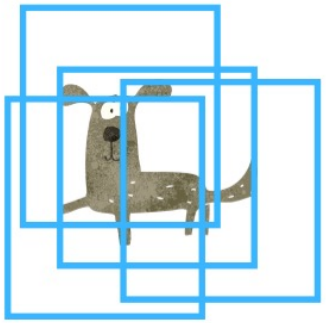
Since each bounding box has a score, we can sort them by descending order. Suppose the red bounding box has the highest score.

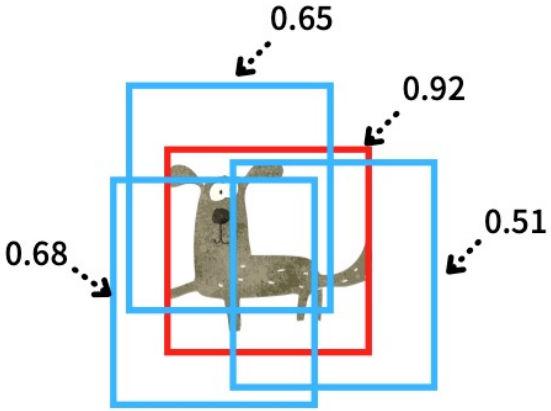
For each of the blue bounding boxes (with a lower score than the red one), we calculate the IoU (Intersection over Union) with the red one.

An IoU value represents an overlap between two boxes, ranging from 0 (no overlap) to 1 (maximum overlap).

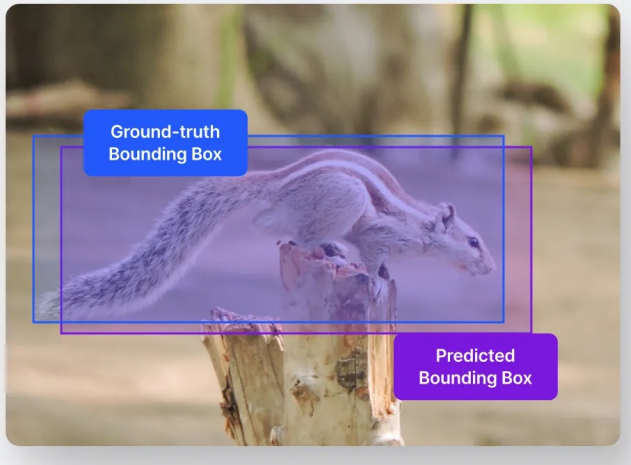
Compare this bounding box's IoU to every other predicted bounding box of the same class, and if the IoU exceeds the user-defined IoU threshold, discard it as a duplicate detection.

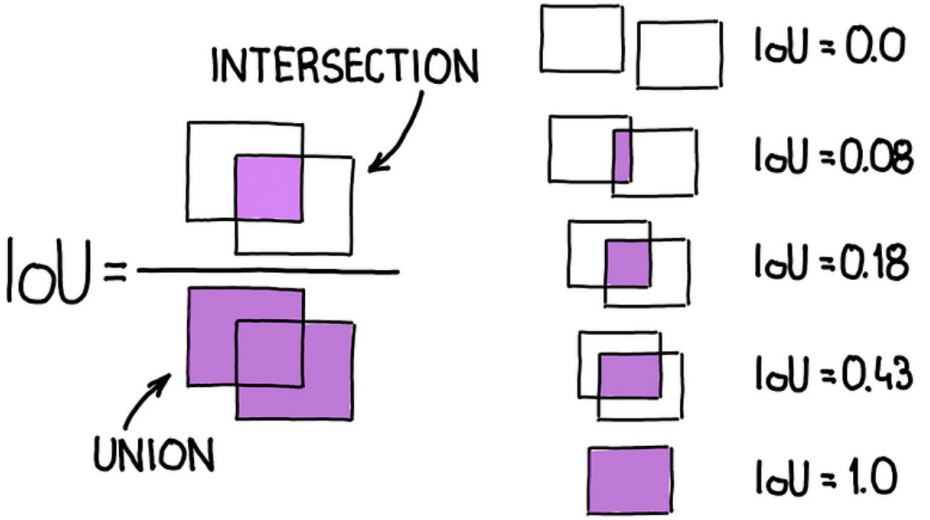
Remove the predicted bounding box from the list of bounding boxes.



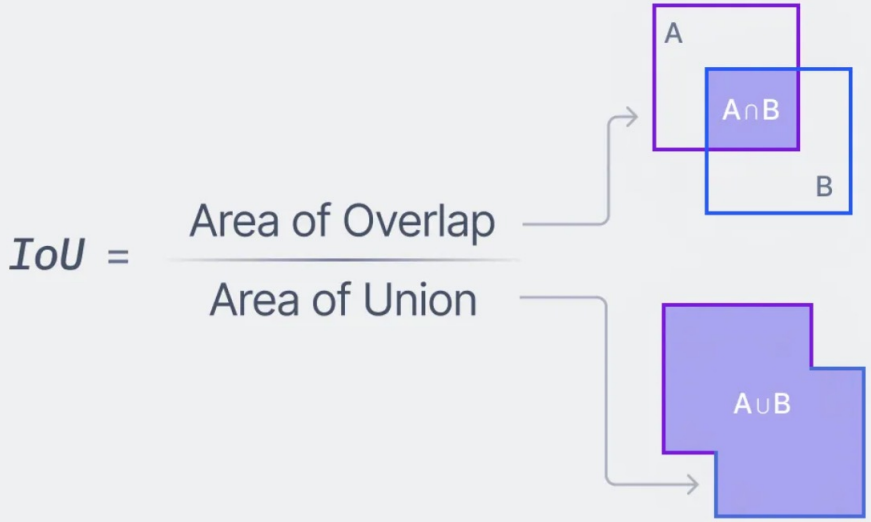


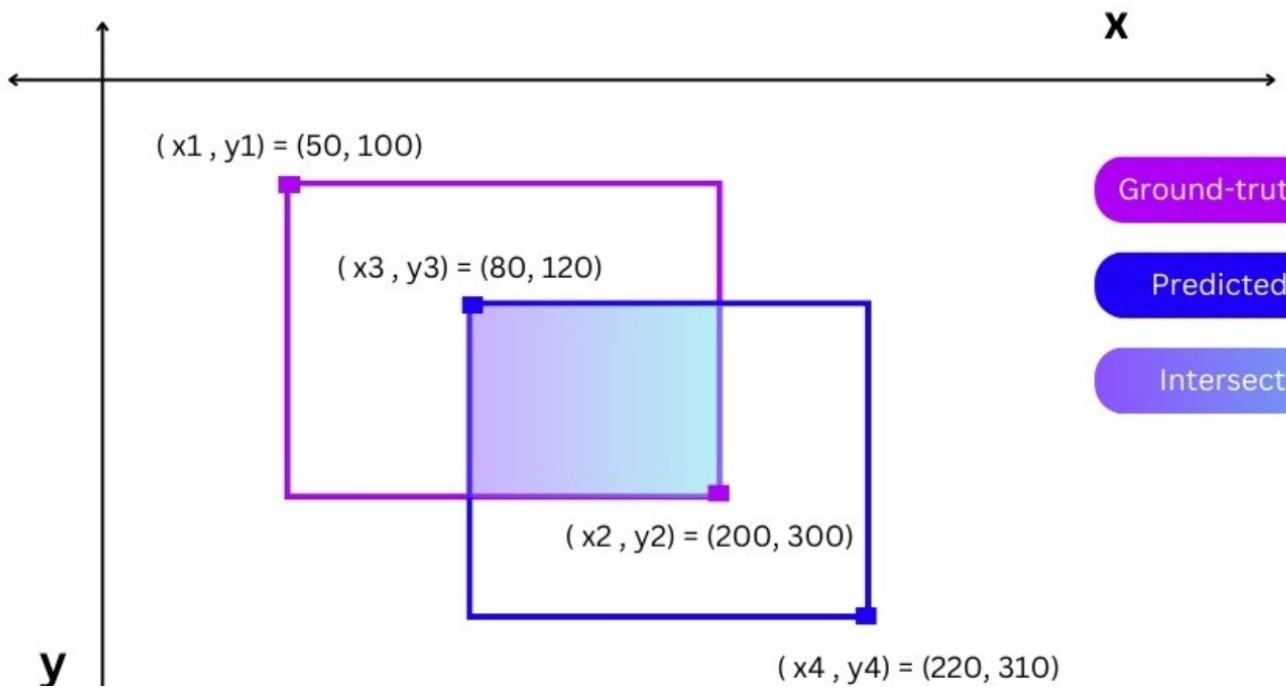
IoU





Intersection over Union





Ground-truth Bounding Box

Predicted Bounding Box

Intersection / Overlap

$$: \frac{21,600}{35,000}$$

$$= 0.618$$

$$Area\ of\ Intersection = |A \cap B| = (x_2 - x_3).(y_2 - y_3)$$

$$|A \cap B| = (200 - 80).(300 - 120)$$

$$|A \cap B| = (120).(180)$$

$$|A \cap B| = 21,600$$

$$Area\ of\ Union = |A| \cup |B| = |A| + |B| - |A \cap B|$$

$$|A| \cup |B| = [(x_2 - x_1).(y_2 - y_1)] + [(x_4 - x_3).(y_4 - y_3)] - |A \cap B|$$

$$|A| \cup |B| = [(200 - 50).(300 - 100)] + [(220 - 80).(310 - 120)] - 21,600$$

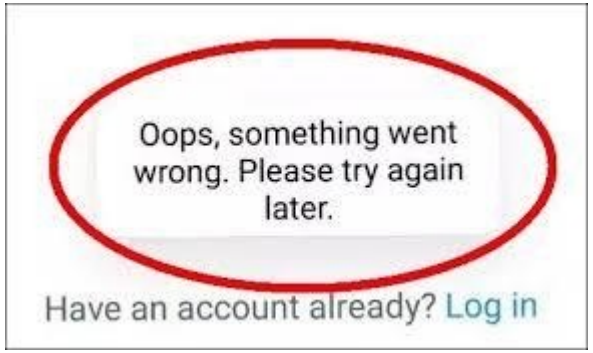
$$|A| \cup |B| = 30,000 + 26,600 - 21,600$$

$$|A| \cup |B| = 35,000$$

Localization aspect of Object Detection

- A regression model with an L1/L2 Loss Function is attached to predict the bounding boxes coordinates.
- This Bounding Box Regression is optional and was added later to the Original R-CNN implementation to increase the localization accuracy.
- The reason is of adding accuracy-→ need to input the ground truth bounding box coordinates also while training this stage.

- Solution is as usual !!!!!
  - fine-tuned the network by training using an n-layer softmax output layer. This increased the accuracy by 10%.
  - Again there is an issue--→>>



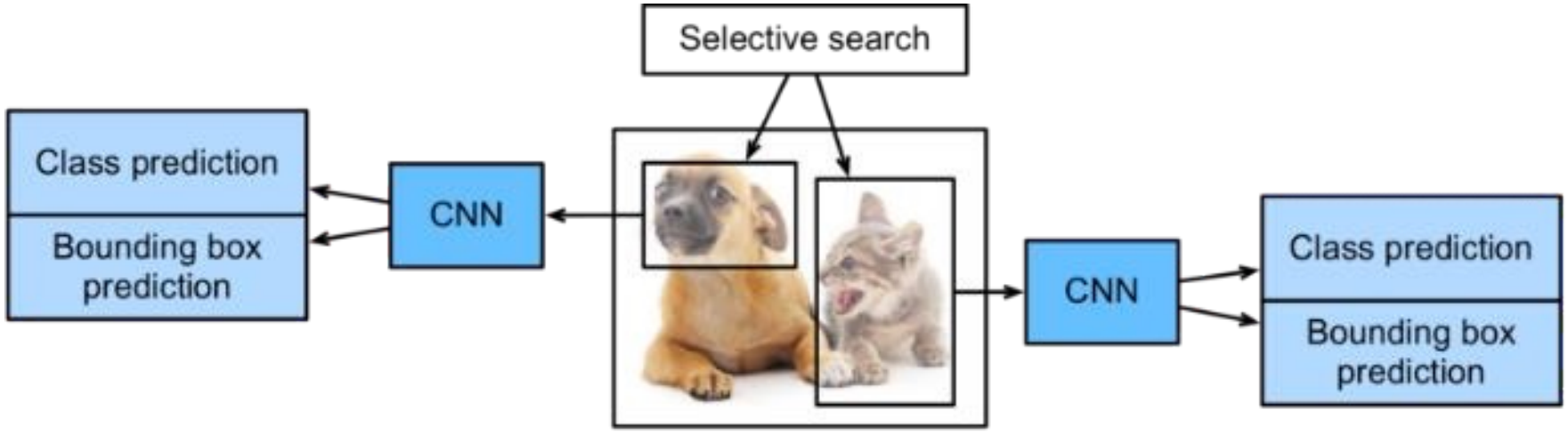
  - Model might predict multiple bounding boxes for a single image, say around 5 considering a single object is in the image.
  - **Greedy approach** of iteratively sorting and selecting the boxes with the IoU confidence scores helps overcome the overlaps of multiple boxes and the single best bounding box coordinates are predicted.

- A Bounding Box Regressor helped to get the predicted bounding boxes closer to the ground truth coordinates.
- This led to a jump of at least 10% in accuracy and later when the VGG network was used in place of AlexNet, the accuracy reported was close to 66 %.
- The R-CNN achieved a mAP of 54% on the PASCAL VOC 2010 and 31% on the ImageNet datasets.

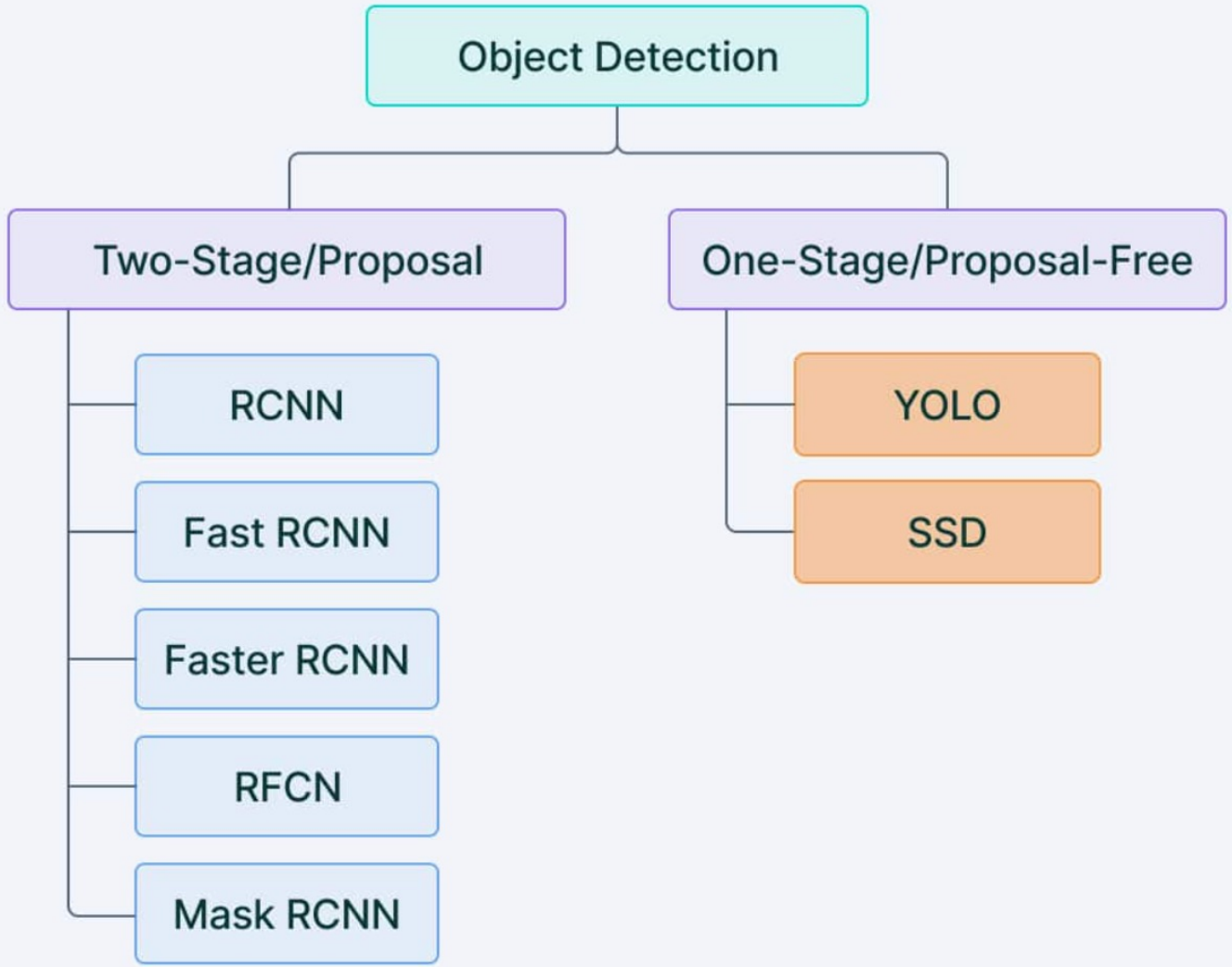
R-CNN -pros

- It takes a huge amount of time to train the network as you would have to classify 2000 region proposals per image.
- 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.

R-CNN Architecture



One and two stage detectors



RNN VARIANTS