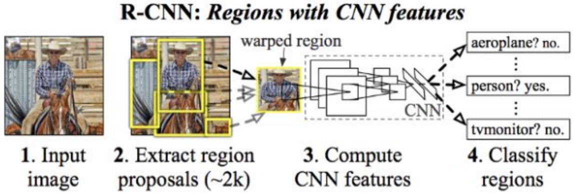
Object Detection Algorithm Comparisons

R-CNN: Region-based Convolutional Neural Network

Description: Selective Search is utilized to extract 2000 regions from an image (region proposals). The regions are the input of a CNN to create 4096-dimensional feature vector and extracts features in the process. The extracted features serve as the input for a Support Vector Machine to classify an object in the region proposal. The algorithm also predicts four offset values to improve the precision of the bounding box.

Faults:

- Long period to train the network (classifying 2000 region proposals for each image)
- Cannot be implemented real time
- Selective search algorithm is a fixed algorithm (does not learn)
- Takes two shots
  - 1. Generating regional proposals
  - 2. Detecting object of each proposal)



SSD: Single Shot (Multibox) Detector

Description:

Through convolution feature extractions, a feature layer of a certain size (m x n) with # channels (p) is created. For each location, a # of bounding boxes (k) of varying size and aspect ratios are created. Each bounding box will be computed a class score and 4 offsets based on the default bounding box shape [(c+4)*kmn].

YOLO: You Only Look Once

Description: A single convolutional network predicts bounding boxes and class probabilities of the boxes. The image is divided into a grid. For each grid, m bounding boxes are extracted. Each bounding box has their class probability ranked and the ones above a threshold value are chosen to locate the object in the image. YOLO is quite fast as it operates at around 45 frames per second

Faults:

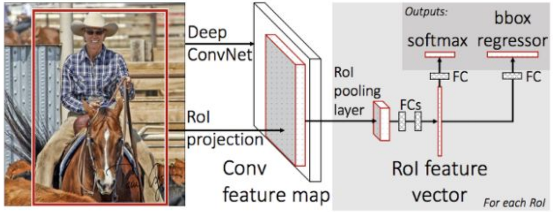
- Struggles with small objects in an image

Fast R-CNN

Description: Different than R-CNN in that the image is the input to the CNN instead of the region proposals. Based on the convolutional feature map, the region of proposals are identified and converted into squares. Using a RoI pooling layer, it is then reshaped into a fixed size as input to a fully connected layer. From RoI feature vector, a softmax layer is used to predict a class of a proposed region and offset values from the bounding box

[improved & faster because convolution operation is done once per image and feature map is generated compared to applying 2000 region proposals to the CNN]

Fast R-CNN

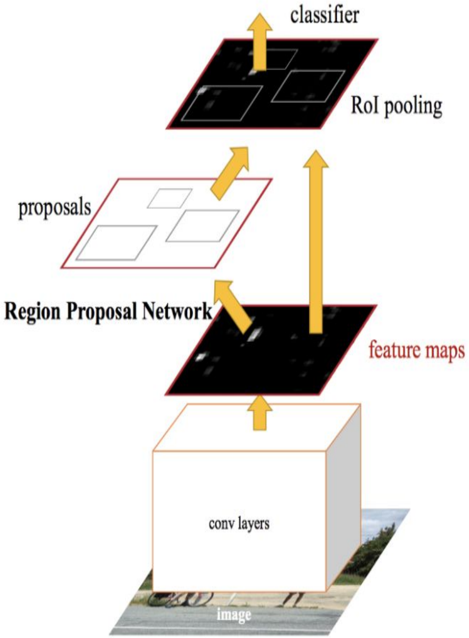


Fast R-CNN

Faster R-CNN

Description: Similar to R-CNN and Fast R-CNN except it does not utilize selective search. The image is sent into the convolutional network which creates a convolutional feature map. A separate network is used to observe the feature map to find region proposals instead of utilizing the selective search algorithm. The predicted region proposals that were found from the network are then reshaped using a RoI pooling layer. These are used to classify the image in the proposed region and predict offset values for the bounding boxes.

Faster R-CNN



Faster R-CNN

Miscellaneous Facts/Information

Source for information and images about the algorithms mentioned:
https://towardsdatascience.com/

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

SSD300 achieves 74.3% mAP at 59 FPS while **SSD500 achieves 76.9% mAP at 22 FPS**, which outperforms Faster R-CNN (73.2% mAP at 7 FPS) and YOLOv1 (63.4% mAP at 45 FPS)