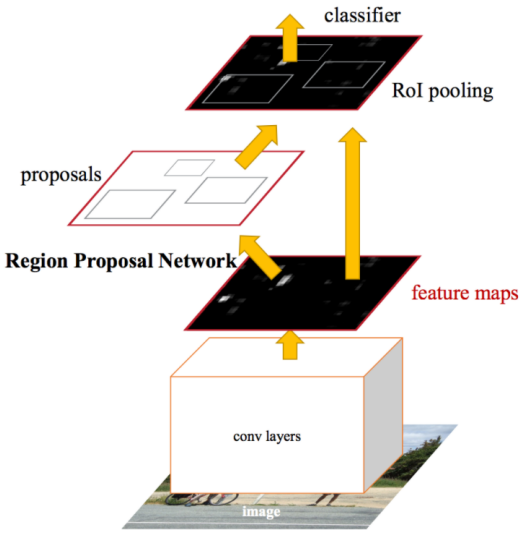
**3 Main Detectors in Open CV**

1. Object Detectors that we will be considering
    a. Faster R-CNN
        i. Two Modules
            1. Region Proposal Network (RPN)
                a. Convolutional Neural Network for proposing regions and the type of object to consider in the region
            2. Fast R-CNN
                a. Convolutional Neural Network for extracting features from the proposed regions and outputting the bounding box and class labels
        ii. Faster R-CNN Architecture:
        iii. 

        iv. RPN takes the output of a pre-trained deep CNN, such as VGG-16, and passes a small network over the feature map and outputting multiple region proposals and a class prediction for each
        v. Region Proposals are bounding boxes, known as anchor boxes
            1. Positive samples have IoU (Intersection Over Union) > 0.7
            2. Negative samples have IoU < 0.3
            3. Slide a small n x n spatial window over the conv feature map of the entire image.

4. At the center of each sliding window, we predict multiple regions of various scales and ratios simultaneously. An anchor is a combination of (sliding window center, scale, ratio). For example, 3 scales + 3 ratios => k=9 anchors at each sliding position.

vi. Class prediction
1. Another neural network takes proposed regions from the first stage and assign them to several specific areas of a feature map level, scans these areas, and generates objects classes

vii. With that, the Output would be
1. Bounding Box of Object
2. Object Class (Name)

b. YOLO (You Only Look Once)
i. Single Neural Network that predicts bounding boxes and class probabilities directly from full images in one evaluation
ii. Constructs a blob from the input image, putting it through an object detector and gives the bounding boxes of a segment of the image that it thinks is an object and their probabilities
iii. Blob: A group of images processed to be the same size, shape, depth
iv. Uses the convolutional neural network
1. Consists of many layers which are made of 1 convolutional layer and 1 pooling layer
2. Treats images as a matrix of pixels
3. The convolutional layer runs the whole image through a filter to extract high-level features, such as edges, colours, and gradient orientation
4. Pooling layer reduces the input into an even smaller matrix to reduce computational power, getting dominant features, suppressing noise in the image

c. SSD
i. Two components
1. Backbone Model

        a. A pre-trained image classification network as a feature extractor
        b. Typically a network like ResNet trained on ImageNet
    2. SSD Head
ii. Workflow
    1. Divides the image using a grid and have each grid cell be responsible for detecting objects in that region of the image (SSDs do not use a sliding window)
    2. Detection of object - Prediction of class and location of an object within that region
    3. Generates Anchor boxes - Responsible for a size and shape within a grid cell
    4. Matching phase - Anchor box with the highest degree of overlap with an object is responsible for predicting that object's class and location

**Detector of Choice**

1. Detector of Choice - YOLO
2. Why YOLO
    a. Designed for speed and real-time use
    b. Very suitable for this project which helps blind people as speed is a very important consideration factor
3. Why not Faster-RCNN
    a. Training data is too long
    b. Training happens in multiple phases
    c. Network is too slow at inference time
4. Why not SSD
    a. Has problems detecting small objects
    b. Hence it is unsuitable for our project as our project scans physical papers for blind paper, and images on the paper may appear small