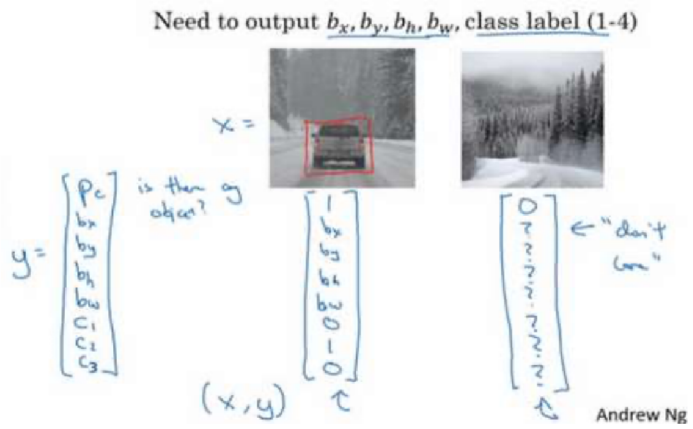
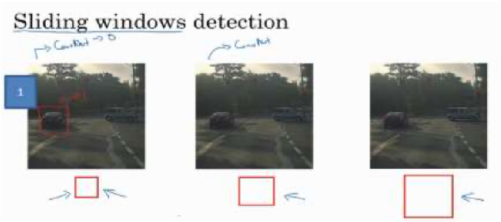
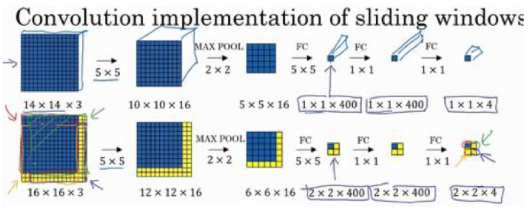
Course 4 - Week 3

1. Distinguish image classification, localization, and object detection?
    a. Image classification: classify which object is in the image (single objects)
    b. Localization: not only classify but also put a bounding box (single objects)
    c. Object detection: for multiple objects
2. In this course, describe the desired output notation for a bounding box prediction?
    a. Output vector for one box: `y = [pc, bx, by, bh, bw, c1, c2, ....cn]`
    b. Explanation
        i. pc: probability of whether there is any desired object (1), or is it just background (0)
        ii. bx, by, bh, bw: suppose the width and height of the image range from 0 to 1, therefore these 4 values are relative/proportion coordinates instead of absolute position coordinates
        iii. c1...cn: class probability
    c. What if pc = 0?
        i. Other values are “don’t cares”



3. What is sliding window approach? Disadvantages? Improvements?
    a. Slide a small window across the image and classify the object in each crop
    b. Disadvantages
        i. Computational cost is very expensive
        ii. When sliding the window (especially with stride), it is possible that some objects may not be completely captured. In other words, the bounding boxes may not be accurate enough.
    c. Improvements: Instead of repeatedly apply the same CNN to every crop, now we can only apply the CNN once and receive all the results in one go. See the right picture for intuition. This improvement solves the problem of computational cost, but the second challenge still exists.





4. Describe YOLO algorithm in both training and testing phase (This Github project helps a lot in understanding dimensions and procedures!)
    a. Training:
        i. **Set up output/label placeholder**: a 19 x 19 x num_anchor_boxes x (5 + num_classess) map. We call each square in the map a grid. Refers to question 2 to understand all fields in the vector of length (5 + num_classess). Since one grid may contain multiple objects (very rare case though), we need more than one bounding box in this grid. Therefore, the concept of anchor box is introduced to detect multiple objects in the same grid.
        ii. **Label each object / Update placeholders**: pick the centre of each object and see which grid it falls in. Each object then belongs to one grid. Update the (5 + num_classess) vector. If a grid has unused anchor boxes, the value of those vectors are set to “don’t cares”.
        iii. **Train the neural network**: regression loss and classification loss
    b. Inference:
        i. Forward pass the image through the network and get a list of bounding boxes vectors of length (5 + num_classess)
        ii. Filter out some bounding boxes based a threshold on pc
        iii. Apply NMS algorithm: for bounding boxes in **each class**, start with the one with highest pc value, filter out those that have a high IOU with it; then start over with the second largest pc value and iterate the process...
5. For RCNN-based networks, we will discuss more in my CS231N notes