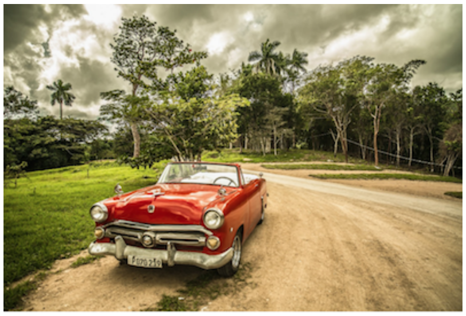
Week 3: Object Detection Algorithms

1. You are building a 3-class object classification and localization algorithm. The classes are: pedestrian (1), car (2), motorcycle (3). What would be a label for car image? Recall $y = [p_c, b_x, b_y, b_h, b_w, c_1, c_2, c_3]$

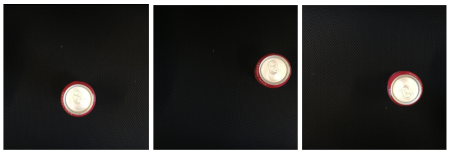


Ans: $y = [1, 0.3, 0.7, 0.3, 0.3, 0, 1, 0]$

2. Continuing from the previous problem, what should y be for the image with no label present? Remember that ? means `Don't Care` condition. As before, $y = [p_c, b_x, b_y, b_h, b_w, c_1, c_2, c_3]$

Ans: $y = [0, ?, ?, ?, ?, ?, ?]$

3. Your factory automation system will see a can of soft-drink coming down a conveyor belt, and you want it to decide whether (i) there is a soft-drink can in image, and if so (ii) its bounding box. Since the soft-drink can is round, the bounding box is always square, There is at most one soft drink can in each image. What is the most appropriate set of output units for your NN?

   Ans: Logistic unit, b_x, b_y

4. If you build a neural network that inputs a picture of a person's face and outputs N landmarks, how many output units will the network have?

   Ans: 2N, considering X-Y parameters.

5. When training object detection systems, you need a training set that contains many pictures of the object(s). However, bounding boxes do not need to be provided in the training set, since the algorithm can learn to detect the objects by itself.

   Ans: False

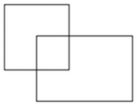
6. Suppose you are applying a sliding windows classifier (non-convolutional implementation). Increasing the stride would tend to increase accuracy, but decrease computational cost.

   Ans: False

7. In the YOLO algorithm, at training time, only one cell (the one containing the center/midpoint of an object) is responsible for detecting this object.

   Ans: True

8. What is the IoU between the two boxes, where the boxes are 2x2, and 2x3. with overlap region at 1x1.

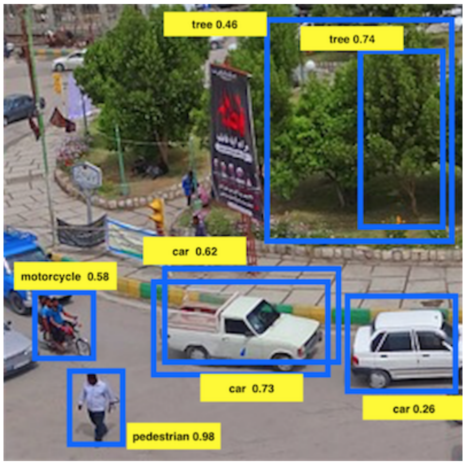
   

Ans: Interseciton area = 1 unit, Union area = 9 units (A+B-I), Hence IOU=1/9

9. Consider YOLO on a 19×19 grid, on a detection problem with 20 classes, and 5 anchor boxes. During training, for each image you will need to construct an output volume y as the target value. What is the dimension needed?

   Ans: $19 \times 19 \times 5 \times 25$. The number of outputs for each anchor box is 5 attributes + number of classes

10. Suppose you run non-max suppression on the image given. Boxes with probability $p \leq 0.4$ are discarded and IOU set at 0.5. How many boxes will remain?



Ans: 5