1. You are building a 3-class object classification and localization algorithm. The classes are: pedestrian (c=1), car (c=2), motorcycle (c=3). What should y be for the image below? Remember that "?" means "don't care", which means that the neural network loss function won't care what the neural network gives for that component of the output. Recall $y = [p_c, b_x, b_y, b_h, b_w, c_1, c_2, c_3]$.

0 / 1 point

https://www.pexels.com/es-es/foto/fotografia-de-motocicleta-clasica-en-carretera-995487/

- $y = [1, 0.22, 0.5, 0.2, 0.3, 1, 1, 1]$
- $y = [1, 0.22, 0.5, 0.2, 0.3, 0, 0, 0]$
- $$y = [1, 0.22, 0.5, 0.2, 0.3, ?, ?, 1]$$

Loading [MathJax]/jax/output/CommonHTML/jax.js

- $y = [1, 0.22, 0.5, 0.2, 0.3, 0, 0, 1]$

Expand

Incorrect

Since $$p_c = 1$$, we need to fill all the other values.

2. You are working on a factory automation task. Your system will see a can of soft-drink coming down a conveyor belt, and you want it to take a picture and decide whether (i) there is a soft-drink can in the image, and if so (ii) its bounding box. Since the soft-drink can is round, the bounding box is always square, and the soft-drink can always appear the same size in the image. There is at most one soft-drink can in each image. Here are some typical images in your training set:

1 / 1 point

The most adequate output for a network to do the required task is $y = [p_c, b_x, b_y, b_h, b_w, c_1]$. (Which of the following do you agree with the most?)

- False, since we only need two values c_1 for no soft-drink can and c_2 for soft-drink can.
- True,

  p_c

  indicates the presence of an object of interest,

  b_x, b_y, b_h, b_w

  indicate the position of the object and its bounding box, and

  c_1

  indicates the probability of there being a can of soft-drink.

Expand

Correct
Correct. With the position $$b_x, b_y$$ we can completely characterize the position of the object if it is present. We should use only one additional logistic unit to indicate if the object is present or not.

3. When building a neural network that inputs a picture of a person's face and outputs N landmarks on the face (assume that the input image contains exactly one face), we need two coordinates for each landmark, thus we need 2N output units. True/False? **1 / 1 point**

- (•) True
- ( ) False

Expand

Correct
Correct. Recall that each landmark is a specific position in the face's image, thus we need to specify two coordinates for each landmark.

4. You are working to create an object detection system, like the ones described in the lectures, to locate cats in a room. To have more data with which to train, you search on the internet and find a large number of cat photos. **1 / 1 point**

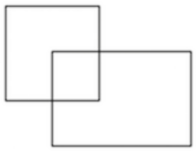
Which of the following is true about the system?

- We should use the internet images in the dev and test set since we don't have bounding boxes.
- We can't add the internet images unless they have bounding boxes.
- We should add the internet images (without the presence of bounding boxes in them) to the train set.
- We can't use internet images because it changes the distribution of the dataset.

Expand

Correct
Correct. As this is a localization model, we also need the coordinates of the bounding boxes, not just the images.

5. What is the IoU between these two boxes? The upper-left box is 2x2, and the lower-right box is 2x3. The overlapping region is 1x1. **1 / 1 point**



- $\frac{1}{10}$
- $\frac{1}{6}$
- None of the above
- $\frac{1}{9}$

Expand

Correct
Correct. The left box's area is 4 while the right box 's is 6. Their intersection's area is 1. So their union's area is 4 + 6 - 1 = 9 which leads to an intersection over union of 1/9.

6. Suppose you run non-max suppression on the predicted boxes nelow. The parameters you use for non-max suppression are that boxes with probability ≤ 0.7 are discarded, and the IoU threshold for deciding if two boxes overlap is 0.5. **1 / 1 point**

After non-max suppression, only three boxes remain. True/False?

- (●) True
- ( ) False

Expand

Correct
Correct. After eliminating the boxes with a score less than 0.7 only three boxes remain, and they don't intersect. Thus three boxes are left.

7. Which of the following do you agree with about the use of anchor boxes in YOLO? Check all that apply. **1 / 1 point**

- [x] Each object is assigned to the grid cell that contains that object's midpoint.

  Correct
  Correct. This is the way we choose the corresponding cell.

- [ ] They prevent the bounding box from suffering from drifting.
- [x] Each object is assigned to an anchor box with the highest IoU inside the assigned cell.

  Correct
  Correct. This is the way we choose the corresponding anchor box.

- [ ] Each object is assigned to any anchor box that contains that object's midpoint.

Expand

Correct
Great, you got all the right answers.

8. Semantic segmentation can only be applied to classify pixels of images in a binary way as 1 or 0, according to whether they belong to a certain class or not. True/False? **1 / 1 point**

- ( ) True
- (●) False

Expand

Correct
Correct. The same ideas used for multi-class classification can be applied to semantic segmentation.

9. Using the concept of Transpose Convolution, fill in the values of **X**, **Y** and **Z** below. **1 / 1 point**

(*padding = 1, stride = 2*)

Input: 2x2

1	2
3	4

Filter: 3x3

1	0	-1
1	0	-1
1	0	-1

Result: 6x6

	0	1	0	-2	
	0	**X**	0	**Y**	
	0	1	0	**Z**	
	0	1	0	-4	

- ○ X = 2, Y = -6, Z = 4
- ○ X = -2, Y = -6, Z = -4
- ◉ X = 2, Y = -6, Z = -4
- ○ X = 2, Y = 6, Z = 4

Expand

Correct

10. When using the U-Net architecture with an input $h \times w \times c$, where c denotes the number of channels, the output will always have the shape $h \times w$. True/False? **1 / 1 point**

- ○ True
- ◉ False

Expand

Correct

Correct. The output of the U-Net architecture can be $$h \times w \times k$$ where $$k$$ is the number of classes.