

- 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 $y = [p_c, b_x, b_y, b_h, b_w, c_1, c_2, c_3]$.



<https://www.pexels.com/es-es/foto/mujer-vestida-con-falda-azul-y-blanca-caminando-cerca-de-la-hierba-verde-durante-el-dia-144474/>

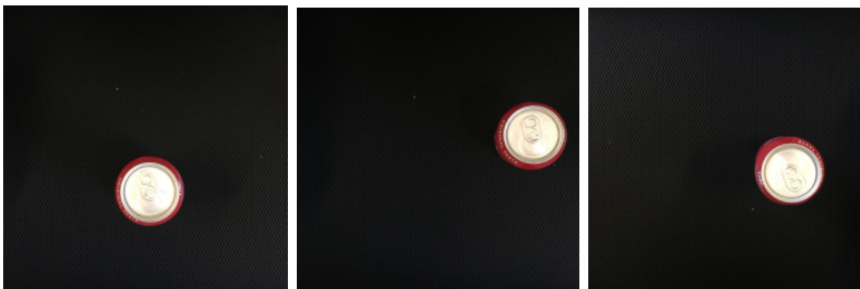
- ☒ $y = [1, 0.66, 0.5, 0.16, 0.75, 1, 0, 0]$
- ☐ $y = [1, 0.66, 0.5, 0.75, 0.16, 0, 0, 0]$
- ☐ $y = [1, ?, ?, ?, ?, 1, ?, ?]$
- ☐ $y = [1, 0.66, 0.5, 0.75, 0.16, 1, 0, 0]$

[Expand](#)

✗ **Incorrect**

Notice that here $b_w > b_h$, and that doesn't correspond to the proportions of the bounding box for the pedestrian.

- 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:



What are the most appropriate (lowest number of) output units for your neural network?

- ☒ Logistic unit, b_x and b_y
- ☐ Logistic unit, b_x, b_y, b_h (since $b_w = b_h$)
- ☐ Logistic unit, b_x, b_y, b_h, b_w
- ☐ Logistic unit (for classifying if there is a soft-drink can in the image)

 Expand

 Correct

3. If you build a neural network that inputs a picture of a person's face and outputs N landmarks on the face (assume the input image always contains exactly one face), how many output units will the network have?

1 / 1 point

- ☐ $3N$
- ☐ N^2
- ☒ $2N$
- ☐ N

 Expand


 Correct
Correct

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.

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

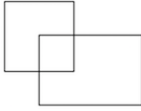
 Expand

 Incorrect

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



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

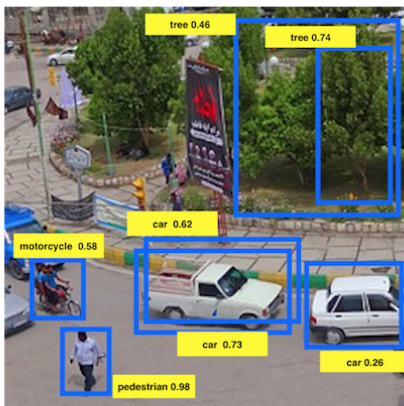
[Expand](#)

✓ 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 below. The parameters you use for non-max suppression are that boxes with probability ≤ 0.4 are discarded, and the IoU threshold for deciding if two boxes overlap is 0.5.

1 / 1 point



Notice that there are three bounding boxes for cars. After running non-max suppression, only the bounding box of the car with 0.73 is kept from the three bounding boxes for cars. True/False? Choose the best answer.

- ☐ False. All the cars are eliminated since there is a pedestrian with a higher score of 0.98.
- ☐ False. Two bounding boxes corresponding to cars are left since their IoU is zero.
- ☒ True. The non-maximum suppression eliminates the bounding boxes with scores lower than the ones of the maximum.

[Expand](#)

✓ Correct

Correct. The bounding box for the car on the right is eliminated because its probability is less than 0.4. Of the two bounding boxes in the middle, one is eliminated because their IoU is higher than 0.5. So, only one bounding box remains.

7. Suppose you are using YOLO on a 19×19 grid, on a detection problem with 20 classes, and with 5 anchor boxes. During training, for each image you will need to construct an output volume y as the target value for the neural network; this corresponds to the last layer of the neural network. (y may include some "?", or "don't cares"). What is the dimension of this output volume?

- ☐ $19 \times 19 \times (20 \times 25)$
- ☒ $19 \times 19 \times (5 \times 25)$
- ☐ $19 \times 19 \times (5 \times 20)$
- ☐ $19 \times 19 \times (25 \times 20)$

 Expand

 **Correct**


Correct, you get a 19×19 grid where each cell encodes information about 5 boxes and each box is defined by a confidence probability (p_c), 4 coordinates (b_x, b_y, b_w, b_h) and classes (c_1, \dots, c_{20}).

8. We are trying to build a system that assigns a value of 1 to each pixel that is part of a tumor from a medical image taken from a patient.

This is a problem of localization? True/False

- ☐ False
- ☒ True

 Expand


 **Incorrect**

This is a problem of semantic segmentation since we need to classify each pixel from the image.

10. Suppose your input to a U-Net architecture is $h \times w \times 3$, where 3 denotes your number of channels (RGB). What will be the dimension of your output?

- ☒ $h \times w \times n$ where n = number of output channels
- ☐ $h \times w \times n$ where n = number of input channels
- ☐ $h \times w \times n$ where n = number of output classes
- ☐ $h \times w \times n$ where n = number of filters used in the algorithm

 Expand

 **Incorrect**

To revise, watch the lecture .



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

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

Expand

✓ Correct

Correct. $p_c = 1$ since there is a motorcycle in the picture. We can also see that b_x, b_y as percentages of the image are adequate. They look approximately correct as well as b_h, b_w , and the value of $c_3 = 1$ for the motorcycle.

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

- ☐ False
- ☒ True

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.

When training one of the object detection systems described in the lectures, each image must have zero or exactly one bounding box. True/False?

1 / 1 point

☒ False

☐ True

[Expand](#)

✓ **Correct**

Correct. In a single image, there might be more than only one instance of the object we are trying to localize, so it must have several bounding boxes.

7. Suppose you are using YOLO on a 19x19 grid, on a detection problem with 20 classes, and with 5 anchor boxes. During training, for each image you will need to construct an output volume y as the target value for the neural network; this corresponds to the last layer of the neural network. (y may include some "?", or "don't cares"). What is the dimension of this output volume?

1 / 1 point

☐ 19x19x(25x20)

☐ 19x19x(20x25)

☒ 19x19x(5x25)

☐ 19x19x(5x20)

[Expand](#)

✓ **Correct**

Correct, you get a 19x19 grid where each cell encodes information about 5 boxes and each box is defined by a confidence probability (p_c), 4 coordinates (b_x, b_y, b_h, b_w) and classes (c_1, \dots, c_{20}).

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?

☐ True

☒ False

[↗ Expand](#)



Correct

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

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 $h \times w \times c$. True/False?

☐ 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. The number of channels doesn't have to match between input and output.