Convolutional Neural Networks

Week 3: Detection Algorithms

1 Version 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? (motorcycle picture)

Answer: 1,0.22,0.5,0.2,0.3,0,0,1

Comment: $p_c = 1$ 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.

1 Version 2: You are building a 3-class object classification and localization algorithm. (car picture)

Answer: 1,0.3,0.7,0.3,0.3,0,1,0

2 Version 1: 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: 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?)

Answer: False, we do not need b_h, b_w since the cans are all the same size.

Comment: 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.

2 Version 2: Continuing from the previous problem, 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.

Answer: $y = [0, ?, ?, ?, ?, ?, ?, ?]$.

3 Version 1: 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), which is true about $\hat{y}^{(i)}$?

Answer: $\hat{y}^{(i)}$ has shape (2N,1).

Comment: Since we have two coordinates (x,y) for each landmark we have N of them.

3 Version 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 appears as the same size in the image. There is at most one soft drink can in each image. Here're some typical images in your training set: What is the most appropriate set of output units for your neural network?

Answer: Logistic unit, b_x, b_y.

4 Version 1: When training one of the object detection systems described in the lectures, each image must have zero or exactly one bounding box. True/False?

Answer: False.

Comment: 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.

4 Version 2: 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?

Answer: 2N.

5 Version 1: What is the IoU between the red box (4 by 5) and the blue box (4 by 5) in the following figure? Assume that all the squares have the same measurements.

Answer: 3/7.

Comment: IoU is calculated as the quotient of the area of the intersection 16 over the area of the union 28.

5 Version 2: When training one of the object detection systems described in lecture, you need a training set that contains many pictures of the object(s) you wish to detect. However, bounding boxes do not need to be provided in the training set, since the algorithm can learn to detect the objects by itself.

Answer: False.

6 Version 1: 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. How many boxes will remain after non-max suppression?

Answer: 5.

6 Version 2: Suppose you are applying a sliding windows classifier (non-convolutional implementation). Increasing the stride would tend to increase accuracy, but decrease computational cost.

Answer: False.

7 Version 1: Suppose you are using YOLO on a 19 by 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?

Answer: 19 by 19 by (5 by 25).

Comment: You get a 19 by 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_h, b_w) and classes ($c_1, \ldots, c_{20}$).

7 Version 2: 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.

Answer: True.

8 What is Semantic Segmentation?

Answer: Locating objects in an image by predicting each pixel as to which class it belongs to.

9 What is the IoU between these two boxes? The upper-left box is 2 by 2, and the lower-right box is 2 by 3. The overlapping region is 1 by 1.

Answer: 1/9.

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

Answer: h by w by n, where n is the number of output classes.