Solutions

Q1 What is the purpose of using multiple anchors per feature map cell?
Solution 1:
Multiple anchors are used per feature map to predict multiple objects of different sizes per input image. For each anchor box, you can find Intersection Over Union (IOU). If IOU is greater than the threshold, detect the object in the boundary box. You can define several anchor boxes, each for different object sizes.
A traditional way of using multiple anchors for the object detector in the image is as follows:

1. The first step is to take an input image and output region where an object may present. This location is also known as region proposal or region of interest. We can find these regions using a selective search or by a neural network.
2. The second step is to take the output from the first step and classifies it into one of the target object classes. After identifying the region proposal, that part of the input image is cropped and fed to the next neural network for classifying the target classes. This network predicts what the target object is present in that location.

There are multiple deep learning algorithms used for object detection such as SSD, YOLO, and Faster RCNN. MaskRCNN is also used for object detection with instance segmentation.

Q2. Does this problem require multiple anchors? Please justify your answer.
Solution 2:
No, This problem does not require multiple anchors because of all the products are of the same size. Product detection and boundary box across the products are done using a single network to perform N = 10 object detection. Such a network is called Single Shot Object Detector. ResNet CNN is used as a neural network for both steps. In the first step, I am finding the Region Proposal Network (RPN) is used for identifying the object in the input image using foreground and background detection. After that in second step dense layer is used to predict the class probability. SSD is logically the same as RPN but different in architecture.

Details:
Given an input image through a convolutional neural network and at end add a fully connected layer that converts the final output vector. For single object detection output vector dimension is 4+No. Of classes. Where the first 4 vectors for object location such as (x, y, height, width). Similarly, for N object detection It would be N*(4 + No. of classes).