As my colleague Ying said before, the main goals of object detection are classification and position. So, starting from our training dataset, we label it with the following parameters: the category of the object and its position information. Among them, the position can be described by the x and y coordinates of center point, and width and height of bounding Box. In this way, we can get a box named Ground Truth Box to contain the entire object as shown in the figure. And the box predicted by our trained model is called the Prediction Box.

First, we consider how to select the regions where the object may appear. For example, there is an image, we want to find out the position of the object, but since we are not sure of the real position and real size of the object. The traditional method is using the sliding Windows to traverse the entire image several times. t's not efficient. Thus, YOLO comes on the scene, as its name You Only Look Once, it only needs to traverse the image once.
YOLO reads the entire image at once and divides it into S*S grids of the same size. If the center point of an object is in the grid, the grid will be responsible for detecting the object. And in each grid, we can use the k-means algorithm to generate some Anchor Boxes that fits our training set.

What is Anchor Box？
An Anchor Box is a five-tuple, which are the x and y coordinates of the center point, the height and width of the Box, and the confidence. Confidence is related to IoU, which we will talk about later.
If there are multiple objects of different sizes in one image at the same time, we can artificially specify several groups of Boxes of different sizes to extract features. For example, the first is used to detect small targets, the second is used to detect medium-sized targets, and the third is for large targets. As shown in the figure, there are 3*3 boxes of different shapes and sizes, so that the probability of "boxing" the target will be increased, and the efficiency of the model will be improved at the same time.

For a detector, we need some rules to evaluate the quality of that, but the output of the model is not structured, and we cannot know the position and size of output data in advance.
As mentioned before, we have a predicted value and a true value. And we can introduce IoU to quantify the degree of coincidence between the predicted and the true value. This concept comes from the set in mathematics, which is the ratio of the intersection of two sets to their union. If the IoU of the two Boxes is equal to 1, it means that the two Boxes are completely coincident. Conversely, when the IoU is equal to 0, it means that the two Boxes are completely non-overlapping. We usually choose an IoU threshold to decide whether a Prediction Box is valid.

Finally, we deal with the situation where the confidence levels of many Prediction Boxes meet the threshold. In order to make the results more clearly, we introduce non-maximum suppression NMS. But the calculation of NMS is more complicated, we can realize the function by calling the library function.

We don't care about the structure of the YOLO model, we only consider the input and output data. The raw data we input is an rgb3 channel image with 448px times 448px, and the output is a 7*7*30.
We can analysis the data from different places: 7*7 is the size when we cut the image into grids, and 30 is a tensor with "depth" information.
The first five data are the 5-tuples of the first Anchor Box, the xy coordinates of the center point, the height and width of the Box, and the confidence.
Then the 5 is the 5-tuples data of the second horizontal Anchor Box.
The last 20 bits of data are the 20 categories of the VOC dataset, and the value is the probability of the corresponding category. And the confidence is multiplied by the probability of each classification to get the final category.

For YOLO's loss function, the formula is as follows. It is a weighted sum of three parts: loss of position, loss of confidence, and loss of classification. The general idea is to sum the squares of the difference between the actual value and the predicted value.