In addition to daily listening, speaking, reading and writing training, I read some articles about image processing and learned the following:

The essence of the neural network learning process is to learn the data distribution. Once the distribution of training data and test data is different, the generalization ability of the network is also greatly reduced; on the other hand, once the distribution of training data in each batch is different (the gradient of the batch drops) , then the network will have to learn to adapt to different distributions at each iteration, which will greatly reduce the training speed of the network. One of the solutions is to do a normalized preprocessing of the data. Adding batch normalization after each convolutional layer greatly improves the convergence speed and reduces the dependence on other regularization methods.

Sliding window sampling is performed on the convolution map, and each center predicts 9 different sizes and proportions of the proposed box. Since all convolutions do not require reshape, spatial information is well preserved. Each feature point of the final feature map corresponds to each cell of the original image. Moreover, replacing predicted direct coordinates with predicted relative offsets simplifies the problem and facilitates network learning. All in all, remove the fully connected layers (to get more spatial information) use the anchor boxes to predict the bounding boxes.

In the use of anchors, the number and breadth dimension of the anchor boxes in Faster-RCNN are often hand-picked priors. If they can be selected from the beginning, they are better and more representative. A priori boxes dimension, then the network should be easier to learn the exact prediction position. Using the K-means clustering method, the ground truth box was found by clustering the ground truth box in the data set. The number of clusters k is the number of anchor boxes, and the width and height of the boxes of k cluster centers are the dimensions of the anchor box.