You Only Look Once Unified, Real-Time Object Detection ax1605

Divides the image into a regular grid of cells, 7 x 7 in the paper and a fixed number of bounding boxes are predicted for each cell, 2 in the paper

For each bounding box, 5 numbers are predicted to indicate the location and confidence of the box and for each grid cell, C additional numbers are predicted to represent the class probabilities where C=20= number of classes;

the entire image is process at once unlike the sliding window kernel system used in RCNN and SSD so that the network produces a single output tensor of size 7 x 7 x 30 in this case;

only the grid cell within which the center of a ground truth bounding box lies is responsible for predicting that box

The confidence a score of the box is supposed to represent both the object Ness score and its intersection over union with the ground truth box;