Common models about object detection

How to help computer understand the world through image is an important issue in several fields, such as robotics, medical and autopilot. With the development of machine learning, lots of machine learning model about object detection has emerged.

Technically, there are three levels of understanding image. The first level called classification, which means we can use some strings or ID to describe the image. This is the easiest level of understanding image and lots of excellent machine learning models have already got high accuracy result in this level. The second level called detection. The difference between classification and detection is that the classification only give a description of whole picture and the detection focus on the specific object. The detection will not only give the class of object but also give the location of object in the picture. We will use bounding boxes to represent the location of specific object. The third level is segmentation which includes semantic segmentation and instance segmentation. The semantic segmentation means segment he process of linking each pixel in an image to a class label. And the instance segmentation is and extension of detection, it will describe the edge of the object which provides more accuracy than bounding boxes. And in this report we will focus on detection level.

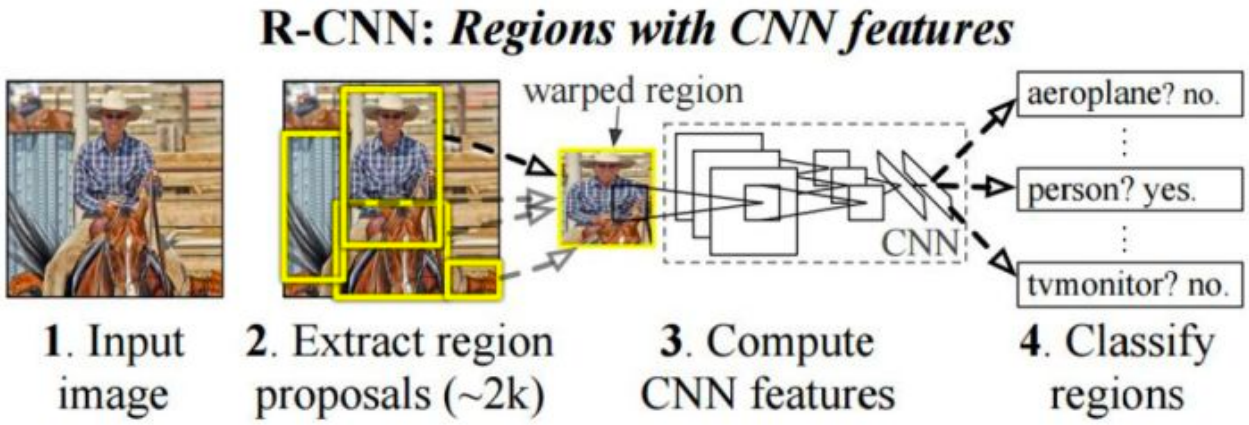
2-stage network

Generally, we can classify the detection models into two ways. The main model of 2-stage model is R-CNN model. With development of model, researchers have developed Fast R-CNN, Faster R-CNN and etc based on R-CNN model. The development of R-CNN model family goes like this:

R-CNN -> SPP Net -> Fast R-CNN -> Faster R-CNN -> Mask R-CNN

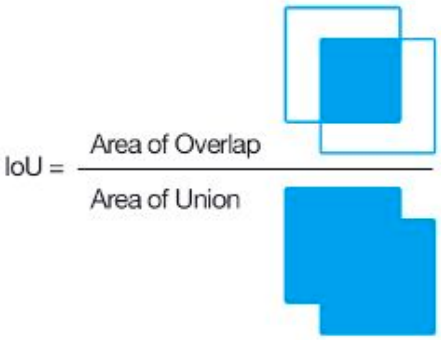
R-CNN

The R-CNN model creatively use CNN network to localize and segment objects. Also, when the number of supervised training samples is scarce, the pre-trained model on the additional data can achieve good results after fine-tuning. In traditional computer vision filed, people like to set some feature manually like SIFT and HOG, and machine learning model prefer study feature from the images.



Network of R-CNN model

R-CNN model separates the detection into two processes. The first one is giving some regions which could include objects and we call this process region proposal. In the original article, it is based on selective search algorithm. The second process is using classification network named AlexNet in these proposal regions.



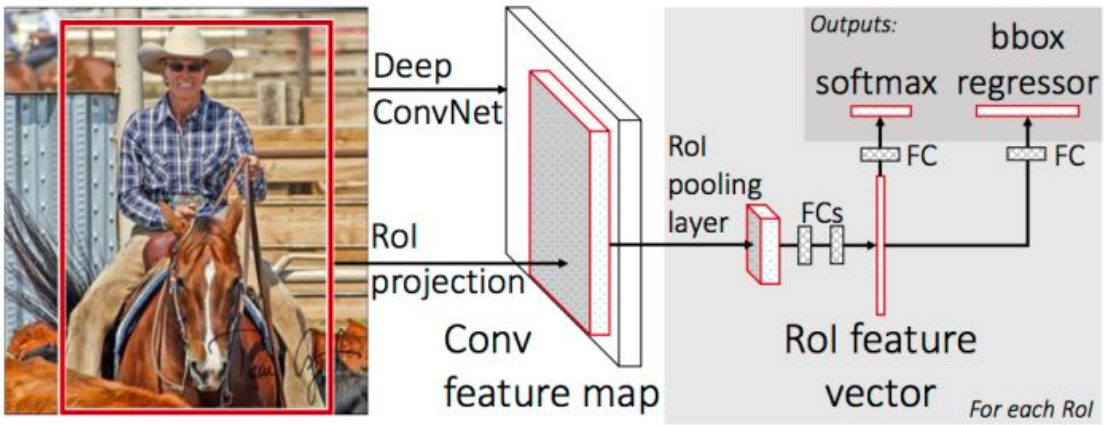
Calculation of IoU

Before entering CNN, we need to mark the proposed Region Proposal according to Ground Truth. The indicator used here is IoU (Intersection over Union). IoU calculates the ratio of the area of the intersection of the two regions to their sum, and describes the degree of coincidence between the two regions.

Another point is the Bounding-Box Regression, which is the adjustment of Region Proposal to Ground Truth. The log/exp transformation is added to achieve the loss level at a reasonable level. It can be regarded as a Normalization operation.

Fast R-CNN

After R-CNN, people found that the R-CNN spends lot of time on process each proposal. To accelerate the computing, some researchers propose that after the basic network is run on the whole picture, it is then transmitted to the R-CNN sub-network, sharing most of the calculations.

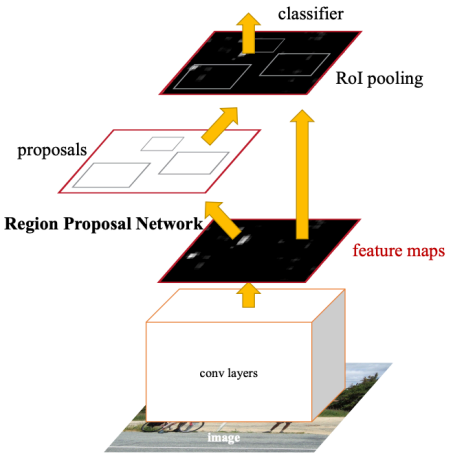


Fast R-CNN Network Structure

The image is obtained by the feature extractor, and the Selective Search algorithm is run on the original image. The RoI (Region of Interset, which can be mixed with Region Proposal) is mapped to the feature map, and RoI Pooling is performed for each RoI. The operation obtains the feature vector of the same length, and the obtained feature vector is arranged for positive and negative samples (maintaining a certain proportion of positive and negative samples), and the batch is transferred to the parallel R-CNN sub-network, and simultaneously classified and returned, and Unify the losses between the two.

Faster R-CNN

Then people propose a new network called Faster R-CNN which basically can be presented as: Faster R-CNN = RPN + Fast R-CNN. RPN means Regional Proposal Networks which can be instead of SS algorithm. RPN network treat proposal as a binary classification problem. The first step is to generate an anchor box of different size and aspect ratio on a sliding window (as shown in the right part of the figure above), determine the threshold of IoU, and calibrate the positive and negative of these anchor boxes according to Ground Truth. Thus, the sample data passed into the RPN network is organized into an anchor box and whether each anchor box has an object (two classification labels). The RPN network maps each sample to a probability value and four coordinate values. The probability value reflects the probability that the anchor box has an object, and the four coordinate values are used to regress the position of the defined object. Finally, the losses of the two classification and coordinate regression are unified and trained as the target of the RPN network.



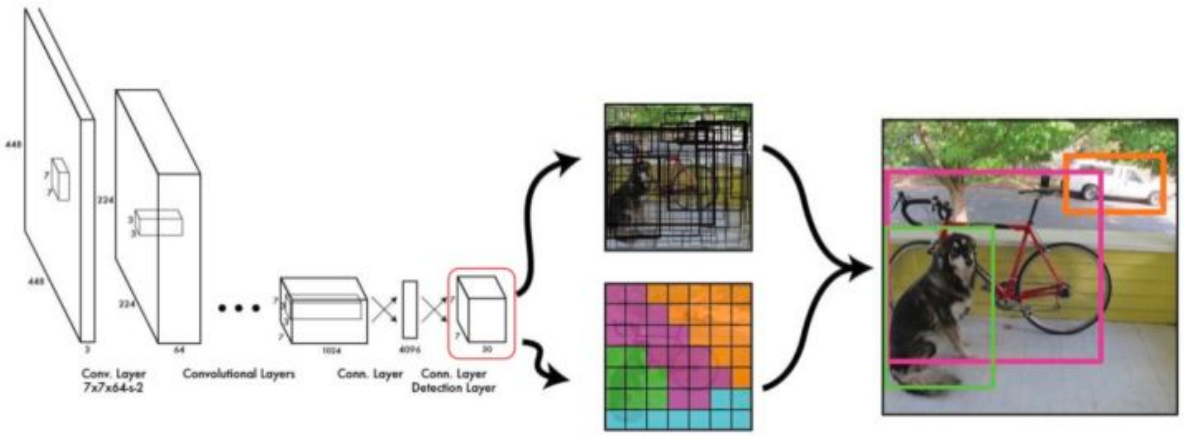
Faster R-CNN Structure

The Region Proposal obtained by the RPN is filtered into a sub-network based on the probability value and passed through the R-CNN sub-network for multi-classification and coordinate regression. The loss of the two is also combined by multi-task loss.

1-Stage network

YOLO

The 1-stage model means it can get the result from picture directly. The most famous 1-stage model is YOLO(you only look once) model. This model transfer the detection problem into an end to end regression problem. The advantage of this kind of model is fast and it can be used in some real-time scene.



Network of YOLO

The structures of YOLO model includes threes parts:

1. Preparation of date: The image will be scaled and divided into equally grids. Each grid is assigned to the sample to be predicted by the IoU of Ground Truth.
2. Convolutional network: baed on GoogLeNet, each grid predicts a conditional probability value for each category, and generates B boxes on a grid basis, each box predicts five regression values, four representation locations The fifth characterizes the accuracy and position (indicated by IoU) of the box containing objects. When testing, the score is calculated as follows:

$$\Pr(\text{Class}_i|\text{Object}) * \Pr(\text{Object}) * \text{IOU}_{\text{pred}}^{\text{truth}} = \Pr(\text{Class}_i) * \text{IOU}_{\text{pred}}^{\text{truth}}$$

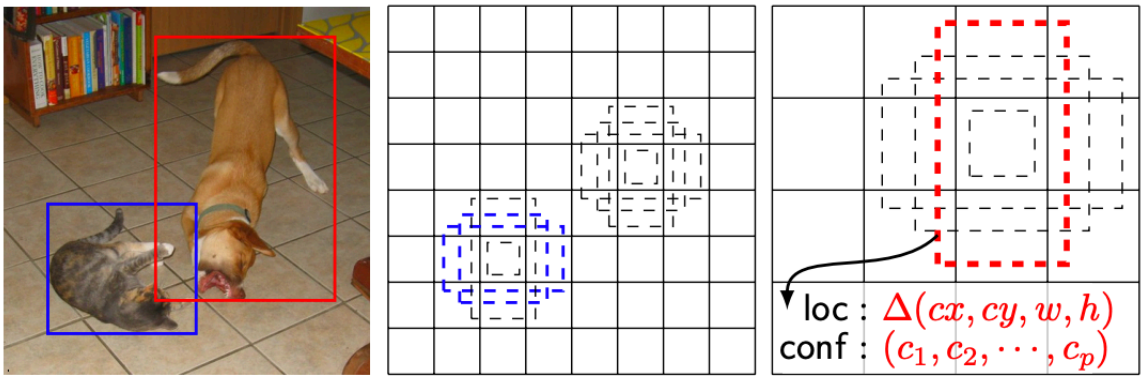
The first item on the left side of the equation is predicted by the grid, and the last two items are predicted by each box. The scores of each box containing different categories of objects are obtained by conditional probability. Therefore, the total number of predicted values output by the convolution network is S×S×(B×5+C), where S is the number of grids, B is the number of boxes generated for each grid, and C is the number of categories. Obviously, the confidence will be zero when there is no object in the bounding boxes.

3. Post-processing: Use NMS (Non-Maximum Suppression) to get the final prediction box.Compared with 2-staged network, the speed advantage of YOLO is obvious, and the real-time features are impressive. However, YOLO itself has some problems. For example, the mesh is rough, and the number of boxes generated by each mesh limits the detection of small-scale objects and similar objects.

In other word, the feature map extracted after the image is convoluted is divided into SxS blocks, and then each block is classified by an excellent classification model, and each grid is processed to remove overlapping frames by using NMS (non-maximum suppression) algorithm. Get our results.

SSD

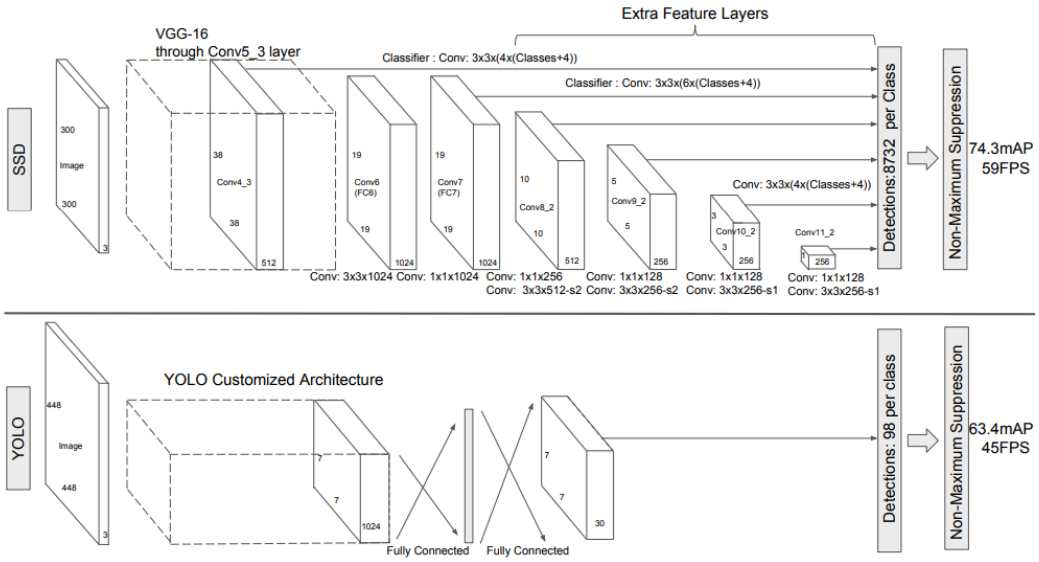
To solve these problem, people invent a new network called SSD(Single Shot Multibox Detector). SSD provides more anchor boxes, each grid point generates boxes of different sizes and aspect ratios, and the category prediction probability is based on box prediction (YOLO is on the grid), and the number of output values obtained is (C+4) ×k×m×n, where C is the number of categories, k is the number of boxes, and m×n is the size of the feature map.



(a) Image with GT boxes (b) 8×8 feature map (c) 4×4 feature map

Based on picture above, we can notice that SSD synthesis the YOLO network and Faster R-CNN Anchor and this greatly improves the detection of small objects.

Another major advancement is the combination of features extracted by Feature Maps of different sizes and then predicted. This is the first attempt by the FPN network to make a Feature Pyramid. This feature pyramid combines information from different layers to combine feature information of different sizes and sizes.



The Structure comparison between SSD and YOLO

Reference:

[1] Girshick, R., Donahue, J., Darrell, T. and Malik, J. (2019). *Rich feature hierarchies for accurate object detection and semantic segmentation*. [online] arXiv.org. Available at: https://arxiv.org/abs/1311.2524 [Accessed 16 Oct. 2019].

[2] Girshick, R. (2019). *Fast R-CNN*. [online] arXiv.org. Available at: https://arxiv.org/abs/1504.08083 [Accessed 16 Oct. 2019].

[3] Ren, S., He, K., Girshick, R. and Sun, J. (2019). *Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks*. [online] arXiv.org. Available at: https://arxiv.org/abs/1506.01497 [Accessed 16 Oct. 2019].

[4] Redmon, J., Divvala, S., Girshick, R. and Farhadi, A. (2019). *You Only Look Once: Unified, Real-Time Object Detection*. [online] arXiv.org. Available at: https://arxiv.org/abs/1506.02640 [Accessed 16 Oct. 2019].

[5] Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C. and Berg, A. (2019). *SSD: Single Shot MultiBox Detector*.