CNN convolutional neural Net works（大容量卷积神经网络），Previously used for image classification.

Combined with Region Proposals, RCNN can be used for image target detection.

RCNN has three modules: our object detection system consists of three modules. The first to generate **Category-Independent Region Proposals**. These suggestions define the candidate detection set available to our detector. The second module is a large convolutional neural network, which extracts a fixed-length feature vector (4096 dimensions) from each region. The third module is a set of class-specific linear SVMs.

The RCNN process can be divided into four steps

1. Generate 1000 to 2000 region proposal in a picture (using the selective search)

2. For each region proposal, use the CNN to extract features

3. The SVM classifier that sends features to each category determines whether it belongs to the category

4. Finely correct the candidate box position using the BB

1. Generate:

Use SS to obtain some original areas through image segmentation, and then merge these areas to obtain a hierarchical regional structure, which contains objects that may be required.

2. Use deep network to extract features for each region proposal

Zoom the 2000 region proposal to 227x277pixel (the reason for scaling to a uniform size is that the size of the pictures it processes must be consistent), and then enter the candidate area to the trained AlexNet CNN. The network obtains the characteristics of the 4096 dimension to obtain the 2000x4096 dimension matrix.

3. The features are sent to the SVM classifier of each category to determine the category.

Multiply the 2000x4096 dimension feature with the weight matrix composed of 20 SVMs (4096x20, the first column and second column, each column represents an item to be tested) to obtain the probability matrix. Each row represents the probability of a recommendation box attributable to each target category. Each column in the above 2000x20-dimensional matrix, that is, each class, to remove the overlapping recommendation boxes, to get some of the highest scores in the column of technical modification. The first column of the probability matrix represents the probability that all candidate boxes are the first item, and so on.

RCNN's training strategy: adopt the supervised pre-training + domain specific fine-tuning strategy: first pre-train CNN (supervised pre-training) on the large dataset ILSVRC, and then fine-tuning on the small dataset PASCAL VOC. The reason for this is that when the RCNN design was completed, there were not enough datasets that can be used to train it, so the relatively classic big dataset is used for supervision pre-training, and then the target detection data set is fine-tuned. (this just for CNN)

The SVM should be trained too.

Specific fine-tuning method: Change the 1000 class of the last layer originally trained on image Net to N+1 layer. When using VOC, N is 20, and +1 refers to the background. AlexNet has five convolutional layers and three full connection layers. The last layer has a total of 1,000 classes, which comes from the previous pre-training on the big data set.

Training a network requires enough data. ImageNet has a lot of data, and each picture will mark its category; but there are fewer VOCs and fewer annotations, but the boxes and categories of the target are marked, which is different from the special picture classification.

There is a available way for CNN's training: for this figure, there are two marked ground truth boxes on the figure. When the SS algorithm generates 2,000 candidate boxes to train people, you can use the marked box and candidate box above to make IoU. The IoU result is greater than 0.5 is a positive sample, and the rest is a negative sample. In this way, we can get a lot of positive and negative samples. In the process of iteration, 32 positive samples and 96 negative samples are forced to be sampled. When this pre-training is completed, it can be used to extract the characteristics of the data set on VOC.

SVM training: When training the vector machine model, the paper takes the ground truth bounding box as a positive sample, and the area with IoU less than 0.3 as a negative sample. In order to control the ratio of positive and negative samples, the standard hard negative method is used to select negative samples.

Why should CNN and SVM conduct a training on classification respectively?

Classifying directly with CNN will lead to lower mAP, which may be because each positive sample (jitter area) is not the exact position coordinate of each target, which will lead to errors. And when training CNN, negative samples are randomly selected. Samples of CNN prediction errors can be used to retrain an SVM.

Question:

1. What is Selective Search: Selective Search is a region proposal generation algorithm used to generate **category-independent region proposals**, by J.R.R. Uijlings and others proposed it in 2012. Its core goal is to efficiently provide Region Proposals that may contain objects for target detection tasks, so as to avoid the problem of excessive calculations of the traditional Sliding Window method. However, it also has a problem, that is, it relies too much on image segmentation, so the speed is slower. In Fast-RCNN, the SS algorithm has been eliminated.

2. What is AlexNet? (AlexNet is a kind of convolutional neural network. It is used to extract the characteristics of each candidate region in RCNN to form a matrix. It has five convolutional layers, three fully connected layers, and some pooled layers in the middle. The original AlexNet is based on imag eNet is trained, so it will output a 1000 image classification categories of imageNet in the last layer, but we only need it to extract image features, so we don't need a third full-connection layer.)

3. What is Affine Warping?

4. More specific analysis of IoU

(IoU (Intersection over Union) is a formula that=IMG_256Its first function is for training. Its second function is to carry out non-maximum suppression in the third step to eliminate overlapping region proposals. Let me give you an example. This is a picture. Here are its two candidate boxes. The first one we call it A, and the second one is B. A and B each have a probability score to evaluate which type of object they are most likely to be. Let's assume that A has a high score (0.98 and 0.88). When the IoU of both is greater than the threshold given (this threshold is set by ourselves), the two goals are considered to be the same goal, delete the low-rated one, which is to select the most accurate and perfect boundary box for each category. Do it once for each column. Used to make overlapping region proposals.

)

5. What is mAP (mean average precision)?

6. What is SVM? (clear)

Bounding Box (BB): Most of the reasons for target recognition errors are inaccurate positioning, and BB can solve this problem to a certain extent. To predict an accurate location based on an inaccurate location, such as this figure,

Px and Py are the center point of the red box x coordinate y coordinate, Pw is its width, and Ph is its height.

The green box is ground truth.

