Object Detection Basics

Resource for learning: R-CNN Explained

R-CNN: Extract region proposals from image using selective search. Instead of using all the cropped image patches through the entire network, it clusters the regions in the image using pre-trained object using **selective search**.

A step process:

1) propose image use regions
2) Warp, transform regions to create larger regions

Assume. the image to be a graph with vertices as pixels & edges as neighborhood of pixels. Weight has some weights based on pixel difference, color difference, etc.

The yellow boxes are regions that might be identified here. Note, more small within-the region on cabinet & two & regions the dachshund.

3) Group adjacent similar regions using the graph created. Bounding box based on region proposals of small & large sizes. This combining is done using vector of similarity:
→ color similarity
→ Texture similarity
→ size similarity
→ shape similarity (nearby are merged with)

Result after the step. But we make the larger image. The horizon. We need have. This is what the step stands for.

Region merging:

The proposals:
Only a subset of the total proposal is "chosen".

II. Extract feature vectors for each region.

Now, we had seen in "understanding images for DL" section we can use pre-trained / have one model. Remove the classification layer & only & use the conv layers. Feed the image into classification then log in the conv, pool, and ReLU.

But, & any "random" vectors is used in every shape & 227 x 227 and our region proposals can be of any shape.

→ 227 → 227 → 227 → 227

→ we can read from region to be warped. & any one from region to warping that may need a random size.

→ 10% (random size) warp → warp

→ 20% (random size) warp → warp

III) Warp region to a fixed size. (warped region) (No warping) (warping)

So, we will have datasets with image where we have the ground truth bounding box and have only 2 & selective search region (200). & 200, we choose the one which has an IOU > 0.5 → the ground truth box.

The ground truth with 2 IS the boxes selective search returned.

After the IOU, here are the boxes that remaining & others which does have enough overlap become background region.

So a combination of positive and background proposals are sampled to train the model only in the batch. (number that will have in training)

The model is trained on detection & background task.

Now, what we have is a model that has region classification size fixed. & convert class to our dataset.

We remove the last classification layer that has the class predictions. We use the only connected network (FC layers) as the feature representation of the region proposals. (the weights are)

As last step:

Break class we use binary SVM one-vs-rest. We assume representation to classify the class & is class. → Yes
So SVM for person class → No other.

SVM can class → Yes
→ No

Deep Net is not updated & frozen. at this stage.

How is the classification done in SVM: for getting the labels

True for class A		Ground Truth Box
	for class A	proposal boxes with IOU < 0.3 with instances of class A

STEPS SO FAR:

1 Train CNN on large classification dataset

2 Fine tune CNN with resized proposals on classes of detection dataset and background class

a Proposals with label as Class X : Proposals with IOU >= 0.5 with GT Box of Class X

b Proposals with label as Class Background : All Remaining Proposals

Train a binary classifier(SVM) for each class on the to layer representation of proposals

a Positive Labelled Proposals for Class X, GT Boxes for that class.

b Negative Labelled Proposals for Class X : Proposals with < 0.3 IOU with ALL GT Boxes of Class X

So the main idea is I classify & SVM classifier is that
→ I can, any IOU > 0.5 was treated as +ve & the
→ IOU < 0.3 as -ve

In SVM, any IOU < 0.3 is considered as -ve & the ground truth box as +ve

Why SVM, SVM?

Improved the results when compared to removing without SVM.

Also to be noted that SVM NMS

We would have matched all these as person. The initial one before using SVM. Because it uses IOU > 0.5 and all these will have IOU > 0. But this leads to wrong identity.

But SVM has only ground truth as +ve. Also, it is also trained using hard -ve mining. SVM is special in classifying IOU < 0.3 as -ve. Hence it leaves better.

Fine tuning & bounding box regression helps improve mAP.

Class specific Bounding Box regression :

To avoid localization errors in our predicted box, I basically is a linear box from the object. I will use class specific bounding box regression. (a linear regression model)

Measure the suggested box coordinates & calculate the P_x, P_y, P_w, P_h & box output the suggested box to the ground truth box. This will help in reducing errors.

P_x, P_y, P_w, P_h coordinates of proposal

Non maximum suppression: (NMS)

To avoid multiple boxes for the same object. It suppresses or filter out redundant boxes.

NMS is performed separately on each class.

So: How it is done ??

→ Sort the scores

→ Generate the predictions based on class. Then we measure the quality of the box predicted using confidence score & the predicted class label.

We pick the highest confidence filter & by overlap with remaining boxes. The one with high overlap is removed. And the rest removed. This is further iterated.

Metric:

mean Average Precision - mAP.

→ Dataset 1:

Extract region proposals using selective search → Train AlexNet on Imagenet classification dataset → Remove softmax layer of the model → Fine tune the model using detection dataset → Train the SVM classifier for each class (+ve on -ve) → Bounding box regression → Extract features using fine-tuning representation → Apply NMS → Filter prediction → Accuracy achieved = 44.7 → After fine-tuning → 54.2

→ After BBR Regression → 58.8

→ On changing Alex net to VGG → 66.0