Paper Notes

RICH FEATURE HIERARCHIES FOR ACCURATE OBJECT DETECTION AND SEMANTIC SEGMENTATION

Authors of the Paper: Ross Girshick, Jeff Donahue, Trevor Darrell, Jitendra Malik

COMPILED BY: Mohit Jain, Aniket Agarwal

Discussion held on 29th October 2018.

This paper marked a major breakthrough in the field of object detection as it was the first to show that a CNN can lead to dramatically higher object detection performance on PASCAL VOC as compared to systems based on simpler HOG-like features, which had been the major standard on various visual recognition tasks.

Motivation

1. Previous paper observations:
    a. **Krizhevsky et al., 2012:** Used deep convolutional networks for large-scale image recognition for the first time.
    b. **Overfeat, 2013:** Used sliding window technique on the image and passed every snippet of the image through the CNN so as to classify the image.
2. Solved the CNN localization problem by operating within the 'recognition using regions' paradigm.

Object Detection Scheme:

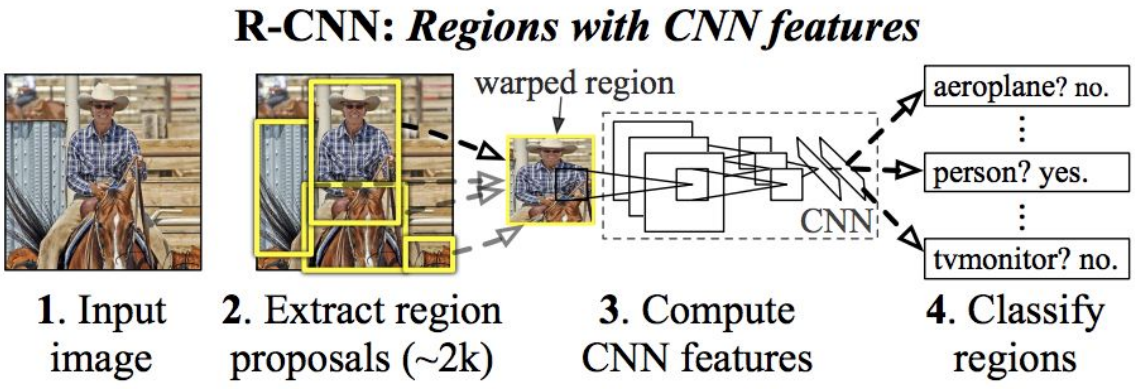

Figure 1: Object detection system overview (taken from paper)

1. The system consisted of three modules:
    a. The first generates category-independent region proposals.
    b. The second module is a large CNN that extracts a fixed-length feature vector from each region.
    c. The third module is a set of class-specific linear SVMs.
2. Usage of **selective search** to extract around 2k region proposals.
3. **Preprocessing:** subtracting the mean RGB value computed on the training set.
4. The CNN takes a 227*227 RGB image as input. The region proposals are transformed into the suitable size by warping the pixels in a tight bounding box around it to the required size. Prior to warping, dilation of tight bounding box takes place so that at the warped size there are exactly p pixels of warped image context around the original box (p=16 in the paper).
5. At test time, the features computed through the CNN are scored on a **class-specific** basis using the SVM trained for that class.
6. After the computation of scores, **non-max suppression** is applied for each class which rejects a region that has an IoU overlap with a higher scoring region larger than a selected threshold.
7. The various steps in the training of the network are as follows:

a. **Supervised Pre-training:** Pre-training the CNN on a large auxiliary dataset (ILSVRC 2012 dataset). The performance was quite like that of AlexNet.
b. **Domain-specific fine-tuning:** The 1000-way classification layer was replaced with a randomly initialized (N+1)-way classification layer(N is the number of object classes, plus 1 for background). All region proposals >= 0.5 IoU overlap with the ground truth were considered positives for the box's classes and the rest as negatives. Also, SGD was used with a learning rate of 0.001.
c. **Object Category Classifiers:** Usage of one linear SVM per class for classification. Also, the IoU threshold here for consideration of an image to be a positive example was set as 0.3, quite different from the one during fine-tuning, which is majorly just an experimentation result.

Miscellaneous Details

1. The various units of the "pool5 layer" were also visualized so as to get a clear idea of what activated a particular unit.
2. Through experimentation, it was also verified that without fine-tuning the model, "fc6" and "fc7" layers generalized quite poorly and the removal of these layers from the network did not affect the overall mAP much. However fine-tuning of the model increased mAP of the model by about 8%, and the boost was much larger for "fc6" and "fc7" than for "pool5".
3. Network architecture by both Zisserman et al. (VGGnet) and Krizhevsky et al. (Alexnet) was tested upon in the model, with both having certain advantages as well as disadvantages.
4. Bounding box regression was used to compute the box using CNN passed features. For this purpose, the training algorithm is a set of *N* training pairs *{(P, G)}* where *P* and *G* refer to the box coordinates, in the R^4 dimension, of the region proposal and the ground-truth respectively. Hence a transformation function is trained using the examples so as to transform *P* to *G.*

Observations

1. Usage of region proposal technique proved to be quite beneficial in improving the accuracy as well as the time taken by the algorithm, compared to Overfeat (2013).

2. Usage of linear classifier like SVM over softmax improved the performance of the model.
3. Fine tuning of the model improved the accuracy of the model by a huge margin, especially the generalization achieved by the “fc6” and “fc7” layers.
4. For converting the dimensions of the region proposals to the suited dimensions for CNN inputs, we can use warping on the content image.

Drawbacks

1. Even though the results and run-time are improved as compared to the Overfeat model, it is still not satisfactory as for every region proposal a full forward pass of the network is to be made which leads to a pretty **high run-time,** hence making the model infeasible for real-time object detection. This problem was recognized and improved in the further versions of this network, namely Fast-RCNN and Faster-RCNN.
2. The requirement of a fixed input size for the CNN also led to a decrease in the accuracy due to warping of the image. This was also further improved on in the Fast-RCNN model.

Conclusion

Usage of region proposal technique can help to increase the accuracy of a network for object detection and also decrease the time taken by the network at test time. However, the model is not end-to-end, suggesting that further improvements can be made.