Rich feature hierarchies for object detection and semantic segmentation

Paper Summary

In this paper researchers address the main problems and provide the solutions of these problems in object detection by their new approach of localizing objects with a deep network called R_CNN. In object detection tasks localizing multiple objects in a single image properly and training a high capacity model with only a small quantity of annotated detection data is very tough. From 2010 - 2012 the object detection performance on the PASCAL VOC dataset had stagnated. At the same time it was difficult to gain insight into the representation learned by CNN. A rich hierarchy of image features can provide insight into what the networks learn. The richer features produce better performance. The authors of the paper provide a new approach that provides the rich hierarchy of image features. Also the approach bridging the gap between image classification and object detection. This paper is 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. Progression of methods for visual recognition have been extensively based on low level features such as SIFT and HOG. The better CNN for object detection were systems with complex ensembles. They combined several low level image features with high level context from scene classifier and object detection.

The authors of this paper adopted the CNN from Krizhevsky et al 2012. It produces features via forward propagating a mean subtracted RGB image through five Convolutional layers and two fully connected layers.

Network Architecture :

In this paper authors proposed a multi staged deep neural network called R-CNN. The R-CNN network is made up of three modules. First module is for generating category independent region proposals for determining the candidate detection. The network uses Selective Search method to generate the region proposal and extracted around 2000 bottom-up region proposals. The second module is a large CNN for extracting a fixed length vector from each region. The third module is a set of class specific linear SVM modules. Second module extracts a 4096 dimensional feature vector from each region proposal and converts the image in that region into a fixed 227x227 pixel size. Each region proposal is then warped into a tight bounding box and forward propagated through a large CNN. For each class, scores each extracted feature vector using the SVM trained for that class. Prior to SVM training labels need to be applied.It applies a greedy non maximum suppression for each class independently and provides all score regions in an image.

Working Process :

This system takes an input image, extracts around 2000 bottom-up region proposals, computes features for each proposal using a large CNN and then it classifies each region using class-

specific linear SVMs. R-CNN achieves a mean average precision (mAP) of 53.7% on PASCAL VOC 2010. For comparison reports 35.1% mAP using the same region proposals, but with a spatial pyramid and bag-of-visual-words approach. The popular deformable part models perform at 33.4%. On the 200-class ILSVRC2013 detection dataset, R-CNN's mAP is 31.4%, a large improvement over OverFeat [34], which had the previous best result at 24.3%.

Training data is required for three procedures in R-CNN: (1) CNN fine-tuning, (2) detector SVM training, and (3) bounding-box regressor training. In terms of having insufficient label data the paper approach is the combined supervised pre-tuning and domain specific fine tuning. Supervised pre-tuning involves using a large auxiliary Dataset - ILSVRC 2012 without image label annotation to pre-train the CNN. Domain specific fine tuning continues SGD training of the CNN parameters which allows the CNN to perform object detection. It uses only warped region proposals from VOC. Two properties make object detection efficient. Firstly CNN parameters are shared across all categories and secondly feature vectors are low dimensional compared to other common approaches.

For semantic segmentation three strategies are applied. The "Full" strategy ignores the region shape and computes CNN features directly on the warped window. The "Fg" strategy computes CNN features only on a region's foreground mask. The "Full+Fg" strategy concatenates the full and Fg features.

Limitations:

1. *Label data is scarce.*
2. *Amount of data readily available is insufficient for training a large CNN.*
3. *Training on an auxiliary dataset is that there might be redundancy between it and the test set.*
4. *Training is expensive in space and time.*