**Paper Title:** Rich feature hierarchies for accurate object detection and semantic segmentation

**Authors:** Ross Girshick, Jeff Donahue, Trevor Darrell, Jitendra Malik, UC Berkeley

Published on 6 January, 2016

In this paper, they propose "R-CNN: Regions with Convolutional Neural Network" features to detect object and semantic segmentation. They use ILSVRC2013 dataset to train their model which contain 200 classes. The number of images in train, validation and test set is 395918, 20121 and 40152 respectively. The validation and test dataset contain labeled images. The train dataset contains both labeled and unlabeled images, and each class has an extra set of negative images. The training data is required for three procedures in R-CNN: (1) CNN fine-tuning, (2) detector SVM training, and (3) bounding-box regressor training.

They achieve mean average precision (mAP) of 53.3% which is more than 30% relative to the previous best result on VOC 2012. This paper has two key findings, (1) we can apply high-capacity CNNs to bottom-up region proposals to localize and segment objects and (2) if labeled data is not found then using supervised pre-training model followed by domain specific and fine tuning can boost the performance significantly.

For object detection using R-CNN has three modules, (1) generates category-independent region proposals, (2) large CNN which can extract a fixed-length feature vector from each region, and (3) a set of class-specific linear SVMs. They followed the same region proposal approach that was used for detection on PASCAL. They found an average of 2409 region proposals per image with a 91.6% recall on validation by running a selective search [1] in "fast mode". They extract a 4096-dimensional feature vector from each region proposal for feature extraction using Caffe [2] implementation. They computed the features using forward propagating a mean-subtracted 227 × 227 RGB image through five convolutional layers and two fully connected layers. By visualizing each layer, they found that much of the CNN’s representational power comes from its convolutional layers, rather than from the much larger densely connected layers.

This approach was very effective in terms of time and accuracy. The previous approach on VOC 2007 takes 5 minutes per image when introducing 10k distractor classes with mAP of around 16%. But in this approach, 10k detectors can run in about a minute on a CPU with a mAP of around 59%. This object detection system takes an input image and extracts around 2000 bottom-up region proposal, then computes features for every proposal using the large CNN and classifies each region using the class-specific linear SVMs.

They used O2P (second-order pooling) open-source framework for the semantic segmentation. O2P uses constrained parametric min-cuts (CPMC) to generate 150 region proposals per image and then predicts the quality of each region, for each class, using support vector regression (SVR). They achieve an average accuracy of 47.9% on VOC 2011.

The main limitation of this R-CNN model is speed. OverFeat has a significant speed advantage over R-CNN. OverFeat is about 9 times faster than R-CNN.

**References**

[1] J. Uijlings, K. van de Sande, T. Gevers, and A. Smeulders. Selective search for object recognition. IJCV, 2013.

[2] Y. Jia. Caffe: An open source convolutional architecture for fast feature embedding. http://caffe.berkeleyvision.org/, 2013.