Rich feature hierarchies for accurate object detection and semantic segmentation

Ross Girshick, Jeff Donahue, Trevor Darrell, Jitendra Malik

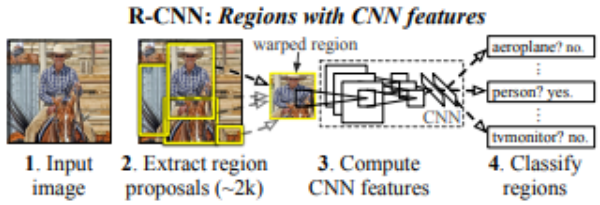
https://arxiv.org/abs/1311.2524
https://arxiv.org/pdf/1311.2524.pdf

TL;DR

This paper proposes a simple and scalable object detection algorithm that significantly improves mean average precision. The approach combines the idea that high-capacity convolutional neural networks can be applied to bottom-up region proposals in order to localize and segment objects and supervised pre-training followed by domain-specific fine-tuning yields a performance boost when labeled data is scarce. This newly proposed method is called R-CNN: Regions with CNN features. In tests, it outperformed OverFeat(another proposed object detection architecture) by a large margin.

Introduction

To apply CNNs to object detection, two problems must be solved: localizing objects with a deep network and training a high-capacity model with only a small quantity of annotated detection data. Object detection requires localizing many objects within an image. Many previous approaches like regression or a sliding window detector have not performed well. R-CNN operates within the ”recognition using regions” framework. At test time, R-CNN first generates around 200- region proposals for the input image then extracts a fixed-length feature vector and lastly classifies each region with SVMs.



Object detection with R-CNN Semantic Segmentation

R-CNN is made of three modules. The first generates category-independent region proposals. The second is a large CNN that extracted a feature vector from each region. The final module is a set of specific linear SVMs. R-CNN is agnostic to the particular region proposal method but uses selective search(selective search is a recursive greedy algorithm used to segment an image. It starts with small regions, and greedily combines similar regions to make larger ones.) to enable a more controller comparison with prior work in object detection. R-CNN extracts a 4096-dimensional feature vector from each region proposal using the Caffe implementation of the CNN. The CNN is composed of five convolutional layers and two fully connected layers. Before a region is fed into the CNN, it is warped into a tight bounding box. Each feature vector is classified using an SVM trained for a class. Given all scored regions in an image, a greedy

non-maximum suppression is applied to reject a region if it has an intersection-over-union overlap with a higher scoring selected region larger than a learned threshold.

Training

The CNN is pre-trained. To adapt it to the new task, SGD is continued using warped region proposals.

Results

VOC 2010 test	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	mbike	person	plant	sheep	sofa	train	tv	mAP
DPM v5 [20]†	49.2	53.8	13.1	15.3	35.5	53.4	49.7	27.0	17.2	28.8	14.7	17.8	46.4	51.2	47.7	10.8	34.2	20.7	43.8	38.3	33.4
UVA [39]	56.2	42.4	15.3	12.6	21.8	49.3	36.8	46.1	12.9	32.1	30.0	36.5	43.5	52.9	32.9	15.3	41.1	31.8	47.0	44.8	35.1
Regionlets [41]	65.0	48.9	25.9	24.6	24.5	56.1	54.5	51.2	17.0	28.9	30.2	35.8	40.2	55.7	43.5	14.3	43.9	32.6	54.0	45.9	39.7
SegDPM [18]†	61.4	53.4	25.6	25.2	35.5	51.7	50.6	50.8	19.3	33.8	26.8	40.4	48.3	54.4	47.1	14.8	38.7	35.0	52.8	43.1	40.4
R-CNN	67.1	64.1	46.7	32.0	30.5	56.4	57.2	65.9	27.0	47.3	40.9	66.6	57.8	65.9	53.6	26.7	56.5	38.1	52.8	50.2	50.2
R-CNN BB	**71.8**	**65.8**	**53.0**	**36.8**	**35.9**	**59.7**	**60.0**	**69.9**	**27.9**	**50.6**	**41.4**	**70.0**	**62.0**	**69.0**	**58.1**	**29.5**	**59.4**	**39.3**	**61.2**	**52.4**	**53.7**

Table 1: Detection average precision (%) on VOC 2010 test. R-CNN is most directly comparable to UVA and Regionlets since all methods use selective search region proposals. Bounding-box regression (BB) is described in Section C. At publication time, SegDPM was the top-performer on the PASCAL VOC leaderboard. †DPM and SegDPM use context rescoring not used by the other methods.

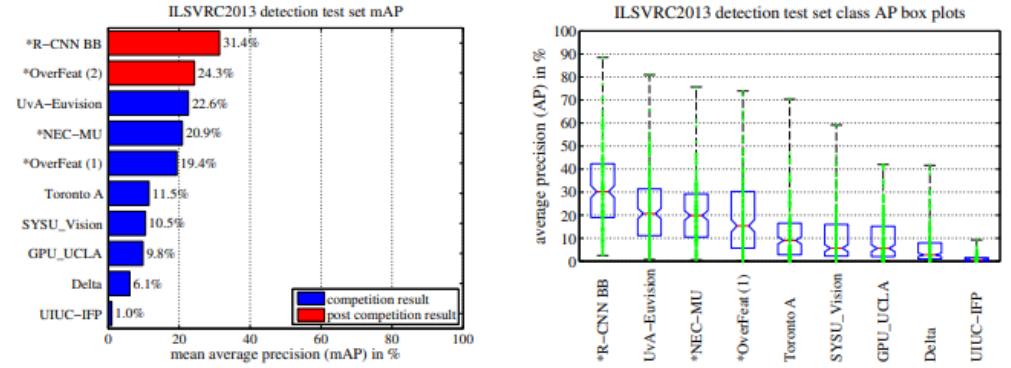


Figure 3: (Left) Mean average precision on the ILSVRC2013 detection test set. Methods preceeded by * use outside training data (images and labels from the ILSVRC classification dataset in all cases). **(Right) Box plots for the 200 average precision values per method.** A box plot for the post-competition OverFeat result is not shown because per-class APs are not yet available (per-class APs for R-CNN are in Table 8 and also included in the tech report source uploaded to arXiv.org; see `R-CNN-ILSVRC2013-APs.txt`). The red line marks the median AP, the box bottom and top are the 25th and 75th percentiles. The whiskers extend to the min and max AP of each method. Each AP is plotted as a green dot over the whiskers (best viewed digitally with zoom).

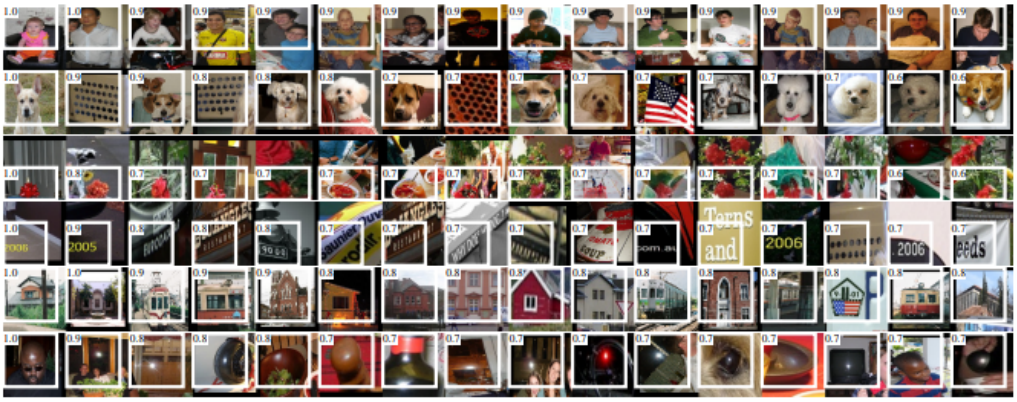


Figure 4: Top regions for six pool$_5$ units. Receptive fields and activation values are drawn in white. Some units are aligned to concepts, such as people (row 1) or text (4). Other units capture texture and material properties, such as dot arrays (2) and specular reflections (6).