







Rich feature hierarchies for accurate object detection and semantic segmentation Tech report (v5)

Ross Girshick Jeff Donahue Trevor Darrell Jitendra Malik UC Berkeley

{rbg, jdonahue, trevor, malik}@eecs.berkeley.edu

Abstract

Object detection performance, as measured on the canonical PASCAL VOC dataset, has plateaued in the last few years. The best-performing methods are complex ensemble systems that typically combine multiple low-level image features with high-level context. In this paper, we propose a simple and scalable detection algorithm that improves mean average precision (mAP) by more than 30% relative to the previous best result on VOC 2012—achieving a mAP of 53.3%. Our approach combines two key insights:

(1) one can apply high-capacity convolutional neural net-

R-CNN: Regions with CNN features

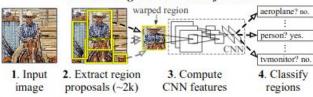


Figure 1: Object detection system overview. Our system (1) takes an input image, (2) extracts around 2000 bottom-up region proposals, (3) computes features for each proposal using a large convolutional neural network (CNN), and then (4) classifies each region using class-specific linear SVMs. R-CNN achieves a mean



Generate Region Proposals





Generate Region Proposals

 Selected regions which "may" have the object





Generate Region Proposals

 Selected regions which "may" have the object





