

Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation

Zhang, Liqiang

June 10, 2018

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%. The approach of this paper combines two key insights: (1) one can apply high-capacity convolutional neural networks (CNNs) to bottom-up region proposals in order to localize and segment objects and (2) when labeled training data is scarce, supervised pre-training for an auxiliary task, followed by domain-specific fine-tuning, yields a significant performance boost.

1. Introduction

Features matter. The last decade of progress on various visual recognition tasks has been based considerably on the use of SIFT and HOG. But if we look at performance on the canonical visual recognition task, PASCAL VOC object detection, it is generally acknowledged that progress has been slow during 2010-2012, with small gains obtained by building ensemble systems and employing minor variants of successful methods.

Fukushima's neocognitron, a biologically inspired hierarchical and shift-invariant model for pattern recognition, was an early attempt at just such a process. The neocognitron, however, lacked a supervised training algorithm. Building on Rumelhart *et al.* [1] showed that stochastic gradient descent via backpropagation was effective for training convolutional neural networks (CNNs), a class of models that extend the neocognitron.

Fig. 1 presents an overview of the method and highlights some of the results. Since their system combines region proposals with CNNs, the author dub the method R-CNN:

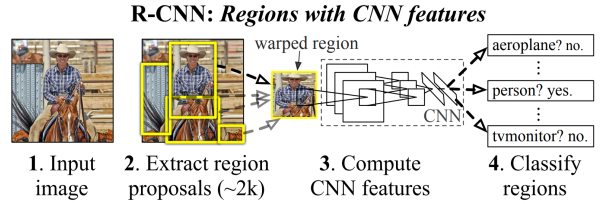


Figure 1. Object detection system overview. Our system (1) takes an input image, (2) extracts around 2000 bottom-up region proposals, (3) computes features for each proposal using a large convolutional neural network (CNN), and then (4) classifies each region using class-specific linear SVMs. R-CNN achieves a mean average precision (mAP) of 53.7% on PASCAL VOC 2010. For comparison, [2] 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%.

Regions with CNN features.

Before developing technical details, we note that because R-CNN operates on regions it is natural to extend it to the task of semantic segmentation. With minor modifications, we also achieve competitive results on the PASCAL VOC segmentation task, with an average segmentation accuracy of 47.9% on the VOC 2011 test set.

2. Method

This loss function is used to match the corresponding image-phrase pairs. Given an input image I_i and a set of corresponding phrases (both positive and negative ones) $\{P_i^1, P_i^2, P_i^3, \dots, P_i^n\}$, the author compute the discriminative loss as:

$$L_{disc} = -Y_i^j \cdot \text{Sigmoid}(\phi_V(I_i) \cdot \phi_L(P_i^j)), \quad (1)$$

where $Y_i^j \in \{-1, 1\}$ is the indicator variable denoting whether P_i^j is a negative/positive match to I_i and $\phi_V(I)$ and $\phi_L(P)$ denote the visual and language code, respectively. The positive phrases are those in the parse tree associated with the input image I_i while the negative phrases are ran-

VOC 2007 test	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	mbike	person	plant	sheep	sofa	train	tv	mAP
R-CNN pool ₅	51.8	60.2	36.4	27.8	23.2	52.8	60.6	49.2	18.3	47.8	44.3	40.8	56.6	58.7	42.4	23.4	46.1	36.7	51.3	55.7	44.2
R-CNN fc ₆	43.6	50.4	32.2	26.0	9.8	58.5	50.4	30.9	7.9	36.1	18.2	31.7	41.4	52.6	8.8	14.0	26.3	34.0	53.1	58.0	46.2
R-CNN fc ₇	57.6	57.9	38.5	31.8	23.7	51.2	58.9	51.4	20.0	50.5	40.9	46.0	51.6	55.9	43.3	23.3	48.1	35.3	51.0	57.4	44.7
R-CNN FT pool ₅	57.1	52.0	31.5	7.6	11.5	55.0	53.1	34.1	1.7	33.1	49.2	42.0	47.3	56.6	15.3	12.8	14.6	40.6	53.1	56.4	47.3
R-CNN FT fc ₆	53.1	57.1	32.4	12.3	15.8	58.2	56.7	39.6	0.9	44.8	39.9	31.0	54.0	62.4	4.5	20.6	34.2	50.0	57.7	63.0	53.1
R-CNN FT fc ₇	64.2	69.7	50.0	41.9	32.0	62.6	71.0	60.7	32.7	58.5	46.5	56.1	60.6	66.8	54.2	31.5	52.8	48.9	57.9	64.7	54.2
R-CNN FT fc ₇ BB	68.1	72.8	56.8	43.0	36.8	66.3	74.2	67.6	34.4	63.5	54.5	61.2	69.1	68.6	58.7	33.4	62.9	51.1	62.5	64.8	58.5

Table 1. Detection average precision (%) on VOC 2007 test. Rows 1-3 show R-CNN performance without fine-tuning. Rows 4-6 show results for the CNN pre-trained on ILSVRC 2012 and then fine-tuned (FT) on VOC 2007 trainval [3]. Row 7 includes a simple bounding box regression (BB) stage that reduces localization errors. The first uses only HOG, while the next two use different feature learning approaches to augment or replace HOG [4].

domly sampled from those in the parse tree associated with any other image.

Table 1 rows 1-3 reveals that features from fc₇ generalize worse than features from fc₆. This means that 29%, or about 16.8 million, of the CNNs parameters can be removed without degrading mAP. More surprising is that removing both fc₇ and fc₆ produces quite good results even though pool₅ features are computed using only 6% of the CNNs parameters.

References

- [1] B. Alexe, T. Deselaers, and V. Ferrari. Measuring the objectness of image windows. *IEEE TPAMI*, 34(11):2189–2202, 2012. 1
- [2] J. Carreira, R. Caseiro, J. Batista, and C. Sminchisescu. Semantic segmentation with second-order pooling. In *ECCV*, 2012. 1
- [3] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *CVPR*, 2005. 2
- [4] J. Donahue, Y. Jia, O. Vinyals, J. Hoffman, N. Zhang, E. Tzeng, and T. Darrell. DeCAF: A deep convolutional activation feature for generic visual recognition. In *ICML*, 2013. 2