Rich Feature Hierarchies for Object Detection and Semantic Segmentation

Hongzhi Liu

August 3, 2018

Abstract

Object detection performance, as measured on the canonical PASCAL VOC dataset, has plateaued in the last few years before Fast R-CNN appear. 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. During preparation for URPC2018, I read a thesis written by Ross Girshick who is from UC Berkeley. This paper proposes a simple and scalable detection algorithm which combines two key insights, one can apply high-capacity convolutional neural networks to bottom-up region proposals in order to localize and segment objects and the other one is that 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. Overview of R-CNN

The last decade of progress on various visual recognition tasks has been based considerably on the use of SIFT [7] and HOG [2] which are blockwise orientation histograms, a representation ones could associate roughly with complex cells in V1, the first cortical area in the primate visual pathway. But people also know that recognition occurs several stages downstream, which suggests that there might be hierarchical, multi-stage processes for computing features that are even more informative for visual recognition.

In this paper, Girshick and his team use a simple technique to compute a fixed-size CNN input from each region proposal, regardless of the region's shape. Fig. 1 presents an overview of the method and highlights some of their results. Since system combines region proposals with CNNs, they dub the method R-CNN: Regions with CNN features. The second major contribution of the paper is to show that supervised pretraining on a large auxiliary dataset, followed by domain-specific fine-tuning on a small dataset, is an effective paradigm for learning high-capacity CNNs when data is scarce.

In Girshick's experiments, fine-tuning for detection improves mAP performance by 8 percentage points. After fine-tuning, their system achieves a mAP of 54% on VOC 2010 compared to 33% for the highly-tuned, HOG-based deformable part model (DPM) [4]. With minor modifications, they also achieve state-of-the-art results on the PASCAL VOC segmentation task, with an average segmentation accuracy of 47.9% on the VOC 2011 test set.

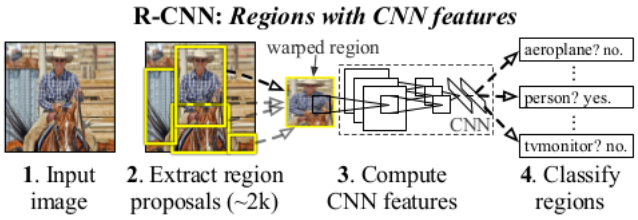

Figure 1. **Object detection system overview.** Girshick's system (1) takes an input image, (2) extracts around 2000 bottom-up region proposals, (3) computes features for each proposal using a large convolutional neural network (CNN), and then (4) classifies each region using class-specific linear SVMs.

2. Object detection with R-CNN

Girshick's object detection system consists of three modules. The first generates category-independent region proposals. These proposals define the set of candidate detections available to his team's detector. The second module is a large convolutional neural network that extracts a fixed-length feature vector from each region. The third module is a set of class specific linear SVMs.

2.1. Module design

R-CNN is agnostic to the particular region proposal method, we use selective search to enable a controlled comparison with prior detection work *e.g.* [8, 9]. The team extract a 4096-dimensional feature vector from each region proposal using the Caffe implementation of the CNN described by Krizhevsky *et al.* [6].

In order to compute features for a region proposal, people must first convert the image data in that region into a form that is compatible with the CNN. Prior to warping, the

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. 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.

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	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 [8]	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	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	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

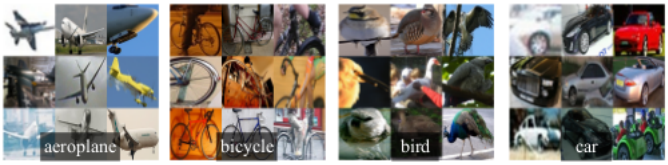

Figure 2. Warped training samples from VOC 2007 train.

team dilate the tight bounding box so that at the warped size there are exactly p pixels of warped image context around the original box. Fig. 2 shows a random sampling of warped training regions. The supplementary material discusses alternatives to warping.

2.2. Test-time detection

At test time, Girshick and his team run selective search on the test image to extract around 2000 region proposals and warp each proposal and forward propagate it through the CNN in order to read off features from the desired layer. Then they score each extracted feature vector using the SVM trained for that class for each one. Given all scored regions in an image, they apply a greedy non-maximum suppression that rejects a region if it has an intersection-over-union (IoU) overlap with a higher scoring selected region larger than a learned threshold.

Two properties make detection efficient. First, all CNN parameters are shared across all categories. Second, the feature vectors computed by the CNN are low-dimensional when compared to other common approaches, such as spatial pyramids with bag-of-visual-word encodings. The features used in the UVA detection system, for example, are two orders of magnitude larger than theirs [8]. The result of such sharing is that the time spent computing region proposals and features (13s/image on a GPU or 53s/image on a CPU) is amortized over all classes.

It is also interesting to contrast R-CNN with the recent work from Dean *et al.* on scalable detection using DPMs and hashing. They report a mAP of around 16% on VOC 2007 at a run-time of 5 minutes per image when introducing 10k distractor classes. With the approach, 10k detectors can run in about a minute on a CPU, and because no approximations are made mAP would remain at 59%.

3. Experiments

Following the PASCAL VOC best practices [3], the team validated all design decisions and hyperparameters on the VOC 2007 dataset. For final results on the VOC 2010-12 datasets, they fine-tuned the CNN on VOC 2012 train and optimized the detection SVMs on VOC 2012 trainval. They submitted test results to the evaluation server only once for each of the two major algorithm variants.

Tab. 1 shows complete results on VOC 2010. They compare the method against four strong baselines, including SegDPM [5], which combines DPM detectors with the output of a semantic segmentation system and uses additional inter-detector context and image-classifier rescoring [1]. Compared to their multi-feature, non-linear kernel SVM approach, they achieve a large improvement in mAP, from 35.1% to 53.7% mAP, while also being much faster. Their method achieves similar performance (53.3% mAP) on VOC 2011/12 test.

References

[1] J. Carreira, C. Rui, J. Batista, and C. Sminchisescu. Semantic segmentation with second-order pooling. In *ECCV*, 2012. 2

[2] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *CVPR*, 2005. 1

[3] M. Everingham, L. V. Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL visual object classes (VOC) challenge. *IJCV*, 2010. 2

[4] P. F. Felzenszwalb, R. B. Girshick, D. Mcallester, and D. Ramanan. Object detection with discriminatively trained part-based models. *IEEE TPAMI*, 2010. 1

[5] S. Fidler, R. Mottaghi, A. Yuille, and R. Urtasun. Bottom-up segmentation for top-down detection. In *CVPR*, 2013. 2

[6] A. Krizhevsky, I. Sutskever, and G. E. Hinton. ImageNet classification with deep convolutional neural networks. In *NIPS*, 2012. 1

[7] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 2004. 1

[8] J. R. Uijlings, K. E. Sande, T. Gevers, and A. W. Smeulders. Selective search for object recognition. *IJCV*, 2013. 1, 2

[9] X. Wang, M. Yang, S. Zhu, and Y. Lin. Regionlets for generic object detection. In *ICCV*, 2014. 1