R-CNN minus R

Karel Lenc
http://www.robots.ox.ac.uk/~karel
Andrea Vedaldi
http://www.robots.ox.ac.uk/~vedaldi

Department of Engineering Science,
University of Oxford,
Oxford, UK.

Deep convolutional neural networks (CNNs) have had a major impact in most areas of image understanding, including object category detection. In object detection, methods such as R-CNN have obtained excellent results by integrating CNNs with region proposal generation algorithms such as selective search. In this paper, we investigate the role of proposal generation in CNN-based detectors in order to determine whether it is a necessary modelling component, carrying essential geometric information not contained in the CNN, or whether it is merely a way of accelerating detection. We do so by designing and evaluating a detector that uses a trivial region generation scheme, constant for each image. Combined with SPP, this results in an excellent and fast detector that does not require to process an image with algorithms other than the CNN itself. We also streamline and simplify the training of CNN-based detectors by integrating several learning steps in a single algorithm, as well as by proposing a number of improvements that accelerate detection.

Object detection is one of the core problems in image understanding. Until recently, the best performing detectors in standard benchmarks such as PASCAL VOC were based on a combination of handcrafted image representations such as SIFT, HOG, and the Fisher Vector and a form of structured output regression, from sliding window to deformable parts models. Recently, however, these pipelines have been outperformed significantly by the ones based on deep learning that acquire representations automatically from data using Convolutional Neural Networks (CNNs). Currently, the best CNN-based detectors are based on the R-CNN construction of [3]. Conceptually, R-CNN is remarkably simple: it samples image regions using a proposal mechanism such as Selective Search (SS; [6]) and classifies them as foreground and background using a CNN.

The first question that we address here is whether CNN contain sufficient geometric information to localise objects, or whether the latter must be supplemented by an external mechanism, such as region proposal generation. There are in fact two hypothesis. The first one is that the only role of proposal generation is to cut down computation by allowing to evaluate the CNN, which is expensive, on a small number of image regions. The second hypothesis is that, instead, proposal generation provides geometric information essential for accurate object localisation which is not represented in the CNN. This is not unlikely, given that CNNs are often trained to be highly invariant to even large geometric deformations and hence may not be sensitive to an object's location.

The second question is whether the R-CNN pipeline can be simplified. While conceptually straightforward, in fact, R-CNN comprises many practical steps that need to be carefully implemented and tuned to obtain a good performance. R-CNN builds on a CNN pre-trained on an image classification tasks such as ImageNet ILSVRC [1], such as the AlexNet network [5]. This CNN is ported to detection by: i) learning an SVM classifier for each object class on top of the last fully-connected layer of the network, ii) fine-tuning the CNN on the task of discriminating objects and background, and iii) learning a bounding box regressor for each object class. We simplify these steps, which require running a mix of different software on cached data, by training a single CNN addressing all required tasks.

The third question is whether R-CNN can be accelerated. A substantial speedup was already obtained in *spatial pyramid pooling* (SPP) by [4] by realising that convolutional features can be shared among different regions rather than being recomputed. However, this does not accelerate training, and in testing the region proposal generation mechanism becomes the new bottleneck. Our first improvement is that we are able to skip the SVM training step, which involves hard negative mining. Furthermore, at the test time the whole detector, including detection of multiple object classes and bounding box regression, *reduces to evaluating a single CNN* (implemented in MatConvNet [7]) which already brings a significant speedup shown in Table 1.

The Table 1 shows that for the SPP detector the main bottleneck is bounding box generation. We show, that picking a constant set of boxes (by clustering the ground truth boxes locations) and retraining the network and bounding box regressor, the drop in performance on PASCAL VOC 2007 data [2] is only about 6% mAP points and results in an overall detection speedup of more than 16×, from about 2.5s per image down to 160ms. The Figure 1 shows the performance of the resulting detectors versus the number of bounding box proposals.

Impl. [ms]	SelS	Prep.	Move	Conv	SPP	FC	BBR	Σ − SelS
SPP MS	$1.98 \cdot 10^3$	23.3	67.5	186.6	211.1	91.0	39.8	**619.2** ±118.0
OURS MS		23.7	17.7	179.4	38.9	87.9	9.8	**357.4** ±34.3
SPP SS		9.0	47.7	31.1	207.1	90.4	39.9	**425.1** ±117.0
OURS SS		9.0	3.0	30.3	19.4	88.0	9.8	**159.5** ±31.5

Table 1: Timing (in *ms*) of the original SPP-CNN and our streamlined full-GPU implementation, broken down into selective search (SS) and preprocessing: image loading and scaling (Prep), CPU/GPU data transfer (Move), convolution layers (Conv), spatial pyramid pooling (SPP), fully connected layers and SVM evaluation (FC), and bounding box regression (BBR).

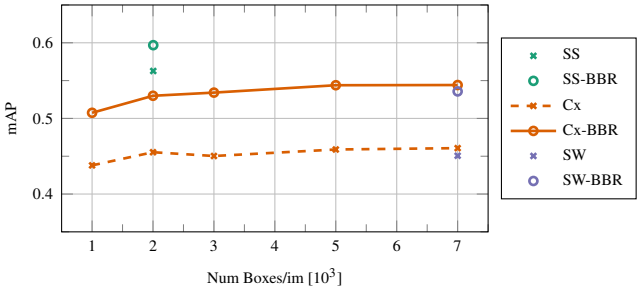

Figure 1: mAP on the PASCAL VOC 2007 test data as a function of the number of candidate boxes per image, proposal generation method, and using or not bounding box regression. In all cases, the CNN is fine-tuned for the particular bounding-box generation algorithm.

Our most significant finding is that current CNNs do contain sufficient geometric information for accurate object detection, although in the convolutional rather than fully connected layers. This finding opens the possibility of building state-of-the-art object detectors that rely exclusively on CNNs, removing region proposal generation schemes such as selective search, and resulting in integrated, simpler, and faster detectors. Our current implementation of a proposal-free detector is already much faster than SPP-CNN, and very close, but not quite as good, in term of mAP. However, we have only begun exploring the design possibilities and we believe that it is a matter of time before the gap closes entirely.

[1] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *Proc. CVPR*, 2009.

[2] M. Everingham, A. Zisserman, C. Williams, and L. Van Gool. The PASCAL visual obiect classes challenge 2007 (VOC2007) results. Technical report, Pascal Challenge, 2007.

[3] R. B. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proc. CVPR*, 2014.

[4] K. He, X. Zhang, S. Ren, and J. Sun. Spatial pyramid pooling in deep convolutional networks for visual recognition. In *Proc. ECCV*, 2014.

[5] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Proc. NIPS*, 2012.

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

[7] A. Vedaldi and K. Lenc. Matconvnet – convolutional neural networks for MATLAB. *CoRR*, abs/1412.4564, 2014.