segDeepM: Exploiting Segmentation and Context in Deep Neural Networks for Object Detection

Yukun Zhu, Raquel Urtasun, Ruslan Salakhutdinov, Sanja Fidler
Department of Computer Science, University of Toronto.

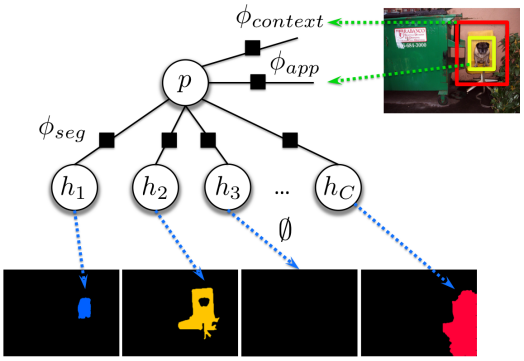

Figure 1: Our SegDeepM exploits appearance, context and segmentation.

In the past two years, Convolutional Neural Networks (CNNs) have revolutionized computer vision. They have been applied to a variety of general vision problems, such as recognition [4, 6], stereo [7], flow [9], etc, consistently outperforming past work. This is mainly due to their high generalization power achieved by learning complex, non-linear dependencies across millions of labelled examples.

For object detection, a successful approach has been to generate a large pool of candidate boxes [8] and classify them using CNNs [4]. Our approach builds on this work. The quality of such a detector thus largely depends on the quality of the object hypotheses. Interestingly, however, using much better proposals obtained via a high-end bottom-up segmentation approach [5] has resulted only in small improvements in accuracy.

In this paper, we show how to exploit a small number of accurate object segment proposals in order to significantly improve object detection performance. We frame the detection problem as inference in a Markov Random Field (Figure 1), in which each detection hypothesis scores object appearance as well as contextual information using CNNs. Following [3], we allow each hypothesis to choose and score a segment out of a small pool of accurate object segmentation proposals. This enables our approach to place more accurate object bounding boxes in parts of the image where object segmentation [2] exists or where strong contextual cues are available. We additionally show that a significant performance boost can be obtained by a sequential approach, where the network iterates between adjusting its spatial scope (the bounding box) and classifying its content. This strategy reduces the dependency on the initial candidate boxes obtained by [8] and enables our approach to recover from the potentially bad initial localization.

Following R-CNN [4], we compute Selective Search boxes [8] yielding approximately 2000 object candidates per image. For each box we extract the last feature layer of the CNN network [6], that is fine-tuned on the PASCAL dataset as proposed in [4]. We also fine-tune another CNN network on boxes enlarged by a fixed percentage on both directions. This network is meant to capture the contextual information around the object. We use these features as potentials in our MRF. We obtain object segment proposals via the CPMC approach [1], although our approach is independent of this choice. Following [2], we take the top 150 proposals given by an object-independent ranker, and train class-specific classifiers for all classes of interest by the second-order pooling method O2P [2]. We remove all segments that have less than 1500 pixels. Our method makes use of these segments along with their class-specific scores. We use several potentials in our MRF capturing compatibility between the candidate detection hypothesis and a segment. These potentials count the number of segment pixels contained inside and outside the box, as well as the number of background pixels inside and outside. The segmentation potentials encourage a detection hypotheses to tightly fit around object segmentation if it's available for an image.

This is an extended abstract. The full paper is available at the Computer Vision Foundation webpage.

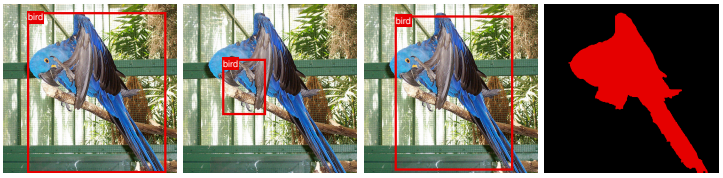

Figure 2: We show the top scoring detections for each ground-truth class. For our method, we also show the segment chosen by our model.

We evaluate our model, called segDeepM, on PASCAL VOC 2010 *val* and show that it outperforms the baseline R-CNN [4] approach by 3.2% in Table 1. The table evaluates each potential function used, which we denote with *seg* (segmentation) and expanded network *exp* (context). We also compare our iterative bbox regression approach, referred to as *ibr*, to the standard bbox regression, referred to as *br*, [4]. We get a total of 5% improvement by incorporating contextual information at the cost of doubling the running time of the method. On PASCAL VOC 2010 *test*, our method achieves 4.1% improvement over R-CNN and 1.4% over the current state-of-the-art. The results and comparison with other methods from the PASCAL Leaderboard are in Table 1. Fig. 2 shows a qualitative example.

	mAP
segDeepM-16 layers	**67.2**
segDeepM-8 layers	57.8
BabyLearning	63.8
R-CNN (breg)-16 ly.	62.9
R-CNN-16 layers	59.8
Feature Edit	56.4
R-CNN (breg)	53.7
R-CNN	50.2

	seg	exp	ibr	br	mAP
RCNN					58.8
segDeepM	√				59.8
segDeepM		√			61.8
segDeepM	√	√			62.2
RCNN				√	62.0
segDeepM	√			√	62.4
segDeepM		√		√	64.4
segDeepM	√	√		√	**64.5**

Table 1: **Left**: State-of-the-art detection results (in %AP) on PASCAL VOC 2010 *test*. The 16 layer models adopt OxforNet, the rest use 8-layer AlexNet. **Right**: Detection results (in % AP) on PASCAL VOC 2010 *val* for RCNN and segDeepM using 16 layer OxfordNet CNN.

[1] Joao Carreira and Cristian Sminchisescu. Constrained parametric min-cuts for automatic object segmentation. In *CVPR*, pages 3241–3248. IEEE, 2010.

[2] João Carreira, Rui Caseiro, Jorge Batista, and Cristian Sminchisescu. Semantic segmentation with second-order pooling. In *ECCV*, pages 430–443. Springer, 2012.

[3] Sanja Fidler, Roozbeh Mottaghi, Alan Yuille, and Raquel Urtasun. Bottom-up segmentation for top-down detection. In *CVPR*, 2013.

[4] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. *arXiv preprint arXiv:1311.2524*, 2013.

[5] Bharath Hariharan, Pablo Arbeláez, Ross Girshick, and Jitendra Malik. Simultaneous detection and segmentation. In *ECCV*, 2014.

[6] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, pages 1097–1105, 2012.

[7] R. Memisevic and C. Conrad. Stereopsis via deep learning. In *NIPS*, 2011.

[8] Koen EA Van de Sande, Jasper RR Uijlings, Theo Gevers, and Arnold WM Smeulders. Segmentation as selective search for object recognition. In *ICCV*, pages 1879–1886. IEEE, 2011.

[9] Philippe Weinzaepfel, Jerome Revaud, Zaid Harchaoui, and Cordelia Schmid. DeepFlow: Large displacement optical flow with deep matching. In *ICCV*, 2013.