ÜberNet: a 'Universal' CNN for the Joint Treatment of Low-, Mid-, and High-Level Vision Problems

Iasonas Kokkinos,
CentraleSupelec & INRIA.

Computer vision involves a host of tasks, such as boundary detection, semantic segmentation, surface estimation, object detection, image classification, to name a few. While these problems are typically tackled by different Deep Convolutional Neural Networks (DCNNs) a joint treatment of those can result not only in simpler, faster, and better systems, but will also be a catalyst for reaching out to other fields. Such all-in-one architectures will become imperative when considering general AI, involving for instance robots that will be able to recognize objects, navigate towards them, and manipulate them. Furthermore, having a single visual module to address a multitude of tasks will make it possible to explore methods that improve performance on all of them, rather than developping techniques that only apply to limited problems.

In this work we develop a single broad-and-deep architecture to jointly address low-, middle-, and high- level tasks of computer vision. This 'universal' network is intended to act like a 'swiss knife' for vision tasks, and we call it an ÜberNet to indicate its overarching nature. Our current architecture has been systematically evaluated on the following tasks (i) boundary detection (ii) normal estimation (iii) saliency estimation (iv) semantic segmentation and (v) proposal generation and object detection. We are currently in the process of integrating the tasks of (vi) symmetry detection (vii) depth estimation (viii) semantic part segmentation (ix) semantic boundary detection and (x) viewpoint estimation, establishing a full vision decathlon.

Our present system operates in 0.3-0.4 seconds per frame on a GPU and delivers excellent results across all of the five first tasks. In normal estimation and saliency detection in particular we obtain results that directly compete with the most recently published techniques of [3] and [7] respectively. In object detection we improve the performance of a strong faster-rcnn baseline [10] from 78.7 mean Average Precision on the test set of PASCAL VOC 2007 to 80.3 by combining detection with segmentation, while the all-in-one network we finally propose recovers the original score of 78.7.

Our starting point is our recent work on using deep learning for boundary detection [5] where we have shown that a multi-resolution DCNN can yield substantial improvements on the task of boundary detection. Moving on to more tasks, we originally observed that with minor architectural changes the same network can be used to solve quite distinct low/mid-level vision tasks, such as normal estimation or saliency detection with performance that directly competes or outperforms the current state-of-the-art on each of those tasks. We attribute this success to the use of (i) side layers and deep supervision, as proposed in [11] (ii) the use of a multi-scale architecture that is trained end-to-end, as proposed in [5] and (iii) the use of batch-normalization at the side layers, which we started using after [5].

A complementary line of progress has been the fusion of segmentation and detection; for this we use a 'two-headed' network, where one head is fully-convolutional until the end, and addresses the semantic segmentation task, as in [1, 5, 9], while the other relies on region proposals to process a shortlist of interesting positions, as in [10]. Our first experiments in this direction demonstrated that the two tasks can benefit from each other by being solved jointly in two ways; firstly, just training a network that performs two tasks turned out to improve performance - and secondly by interleaving the segmentation and detection tasks we have been able to provide information from one task to the other. Our current results indicate that through this procedure we can improve detection performance from 78.7 to 80.3 on the PASCAL VOC test set, as shown in Table. 1.

Even though it was relatively straightforward to train the low- and high-level task networks in isolation, it turned out to be much more challenging to jointly tackle all of the problems above. At first sight this seems similar to simple training a network with a few more losses for additional tasks, as recently done e.g. in [2, 3, 4]. However, we had to face several technical challenges, the most important of which has been the absence of datasets with annotations for all of the tasks considered. High-level annotation is often missing from the datasets used for the training and testing of low-level tasks, and vice versa. This makes it problematic to jointly fine-tune the network, since the high-level tasks are trying to optimize over a network that is a 'moving target' - while freezing the network weights to some values determined from a high level task always results in worse performance in the low-level tasks. We have adapted backpropagation training to asynchronously update network parameters, so that parameters specific to a task are updated only once sufficient annotated training samples have been observed. The minibatch construction has been accordingly modified so that a proper blend of the different datasets is contained in every minibatch.

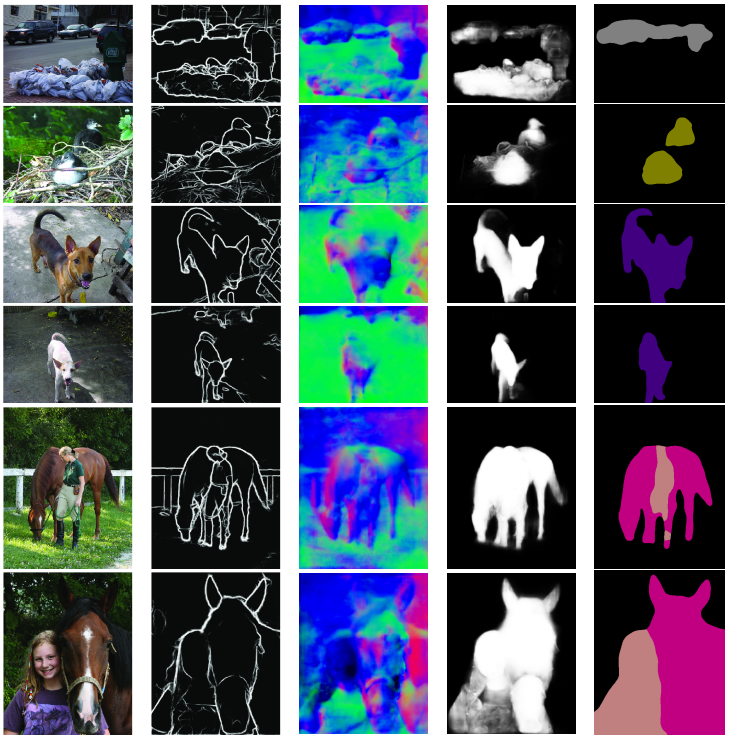

Figure 1: Dense labelling results for four tasks obtained by a single multi-resolution ÜberNet; detection results are not visualized but evaluated below.

	mean Average Precision
Faster-RCNN, MS-COCO + VOC 2007++	78.6
Segmentation & Detection ÜberNet	80.3
All-in-One ÜberNet	78.5

Table 1: AP performance (%) on the PASCAL VOC 2007 test set.

Another major challenge has been the effect of image resolution on performance; we have observed that reducing the image resolution can have an adversarial effect on detection and semantic segmentation performance. However, training with high-resolution images can quickly result in GPU memory issues. We have engineered a two-stage method to train first a network that learns a common low-level, convolutional representation for all tasks, and then freezes the convolutional layers, so as to train in a decoupled manner the subsequent tasks at a higher resolution.

Indicative results are provided below; all of these results, including region proposal generation and object detection are obtained in 0.3 - 0.4 seconds per frame.

We have been exploring (i) how the flow of information across tasks can be exploited, either with feedforward or recurrent connections and (ii)

	Angle Distance		Within t Degrees		
	Mean	Median	11.25	22.5	30
Eigen et. al. [3] - Alexnet	25.9	18.2	33.2	57.5	67.7
Eigen et. al. [3] - VGG	22.2	15.3	38.6	64.0	73.9
ÜberNets (VGG)	21.9	17.2	31.2	63.2	74.1

Table 2: Normal Estimation on NYU-v2 using the ground truth of [6]

	Maximal F-measure
DCL [7]	0.815
DCL + CRF [7]	0.822
ÜberNets	0.826

Table 3: Saliency estimation results on PASCAL Saliency dataset of [8].

automating the choice of which layers of a DCNN are best suited for solving low-, mid- and high- level tasks, obtaining promising results. Due to the preliminary nature of these positive results we cannot elaborate yet on these topics, but they will be covered in a forthcoming technical report.

All of the training and testing code will be made publicly available to accelerate progress in this direction - including methods to simplify the definition of complicated network architectures in Caffe, replacing the construction of prototxt files with thousands of lines by the automatized combination of a few predetermined templates.

[1] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L. Yuille. Semantic image segmentation with deep convolutional nets and fully connected crfs. In *ICLR*, 2015.

[2] Jifeng Dai, Kaiming He, and Jian Sun. Instance-aware semantic segmentation via multi-task network cascades. *CoRR*, abs/1512.04412, 2015. URL http://arxiv.org/abs/1512.04412.

[3] David Eigen and Rob Fergus. Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture. In *ICCV*, 2015.

[4] Georgia Gkioxari, Ross B. Girshick, and Jitendra Malik. Contextual action recognition with r*cnn. In *ICCV*, 2015.

[5] Iasonas Kokkinos. Pushing the boundaries of boundary detection using deep learning. *ICLR*, 2016.

[6] Lubor Ladicky, Bernhard Zeisl, and Marc Pollefeys. Discriminatively trained dense surface normal estimation. In *Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V*, pages 468–484, 2014.

[7] Guanbin Li and Yizhou Yu. Deep contrast learning for salient object detection. In *CVPR*. 2016.

[8] Yin Li, Xiaodi Hou, Christof Koch, James M. Rehg, and Alan L. Yuille. The secrets of salient object segmentation. In *CVPR*, 2014.

[9] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. *CVPR*, 2015.

[10] Shaoqing Ren, Kaiming He, Ross B. Girshick, and Jian Sun. Faster R-CNN: towards real-time object detection with region proposal networks. In *NIPS*, 2015.

[11] Saining Xie and Zhuowen Tu. Holistically-nested edge detection. In *Proc. ICCV*. 2015.