Is object localization for free? – Weakly-supervised learning with convolutional neural networks

Maxime Oquab[1], Léon Bottou[2], Ivan Laptev[1] Josef Sivic[1]
[1]WILLOW project, INRIA Paris, France. [2]Microsoft Research NYC, USA.

Successful methods for visual object recognition typically rely on training datasets containing lots of richly annotated images. Detailed image annotation, e.g. by object bounding boxes, however, is both expensive and often subjective. We describe a weakly supervised convolutional neural network (CNN) for object classification that relies only on image-level labels, yet can learn from cluttered scenes containing multiple objects. We quantify its object classification and object location prediction performance on the Pascal VOC 2012 (20 object classes) and the much larger Microsoft COCO (80 object classes) datasets. We find that the network (i) outputs accurate image-level labels, (ii) predicts approximate locations (but not extents) of objects (see figures 1 and 3) , and (iii) performs comparably to its fully-supervised counterparts using object bounding box annotation for training. We build on the fully supervised network architecture of [3] that consists of five convolutional and four fully connected layers and assumes as input a fixed-size image patch containing a single relatively tightly cropped object. To adapt this architecture to weakly supervised learning we introduce the following three modifications. First, we treat the fully connected layers as convolutions, which allows us to deal with nearly arbitrary-sized images as input. Second, we explicitly search for the highest scoring object position in the image by adding a single global max-pooling layer at the output (see figure 2). Third, we use a cost function that can explicitly model multiple objects present in the image.

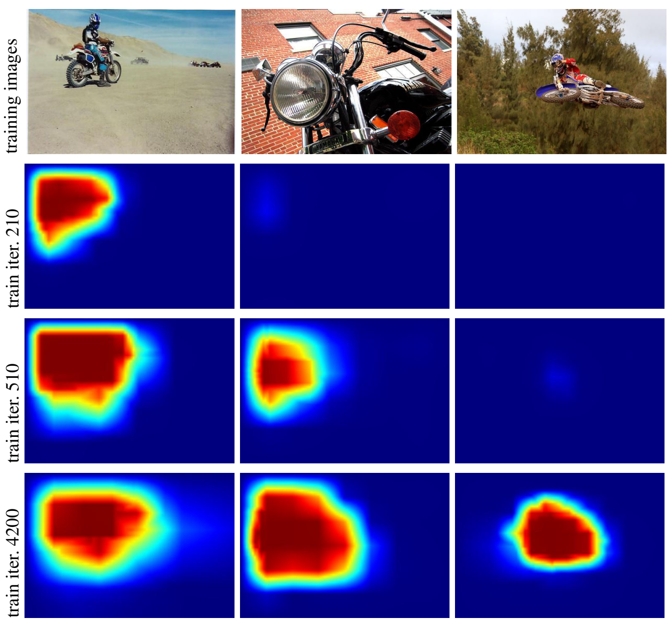

Figure 1: Evolution of localization score maps for the `motorbike` class over iterations of our weakly-supervised CNN training. Note that the network learns to localize objects despite having no object location annotation at training, just object presence/absence labels. Note also that locations of objects with more usual appearance (such as the motorbike shown in left column) are discovered earlier during training.

Figure 3: **Example location predictions** for images from the Microsoft COCO validation set obtained by our weakly-supervised method. Note that our method does not use object locations at training time, yet can predict locations of objects in test images (yellow crosses). The method outputs the most confident location per object per class. **Please see additional results on the project webpage [1].**

We apply the proposed method to the Pascal VOC 2012 object classification task and the recently released Microsoft COCO dataset. Our approach obtains one of the highest overall object classification mAP (86.3%) among single network methods on the Pascal VOC 2012 test set. Furthermore, the proposed weakly supervised architecture outputs score maps for different objects (see figure 1), which can be used to predict the x, y position (but not extent) of the dominant objects in the image (figure 3) with a comparable accuracy to methods trained from images annotated with object bounding boxes [2]. The results open-up the possibility of large-scale reasoning about object relations without the need for detailed object level annotations.

[1] http://www.di.ens.fr/willow/research/weakcnn/, 2014.

[2] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. 2014.

[3] M. Oquab, L. Bottou, I. Laptev, and J. Sivic. Learning and transferring mid-level image representations using convolutional neural networks. 2014.

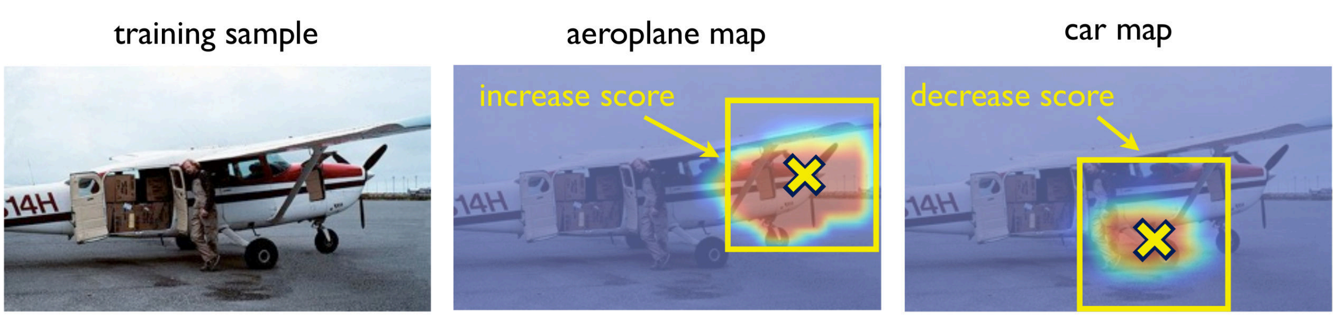

Figure 2: **Illustration of the weakly-supervised learning procedure.** At training time, given an input image with an aeroplane label (left), our method increases the score of the highest scoring positive image window (middle), and decreases scores of the highest scoring negative windows, such as the one for the car class (right).

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

[2] : Léon Bottou is now with Facebook AI Research, New York.