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

Maxime Oquab[*] Le´on Bottou[†] INRIA Paris, France MSR, New York, USA

Abstract

Successful methods for visual object recognition typi- cally rely on training datasets containing lots of richly an- notated images. Detailed image annotation, e.g. by object bounding boxes, however, is both expensive and often sub- jective. We describe a weakly supervised convolutional neu- ral 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 classi- fication 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) pre- dicts approximate locations (but not extents) of objects, and (iii) performs comparably to its fully-supervised counter- parts using object bounding box annotation for training.

Learning and Transferring Mid-Level Image Representations using Convolutional Neural Networks

Maxime Oquab[1,*] Leon Bottou[2] Ivan Laptev[1,*] Josef Sivic[1,*]

[1]INRIA, Paris, France

Abstract

Convolutional neural networks (CNN) have recently shown outstanding image classification performance in the large- scale visual recognition challenge (ILSVRC2012). The suc- cess of CNNs is attributed to their ability to learn rich mid- level image representations as opposed to hand-designed low-level features used in other image classification meth- ods. Learning CNNs, however, amounts to estimating mil- lions of parameters and requires a very large number of annotated image samples. This property currently prevents application of CNNs to problems with limited training data.

In this work we show how image representations learned with CNNs on large-scale

annotated datasets can be effi- ciently transferred to other visual recognition tasks with limited amount of training data. We design a method to reuse layers trained on the ImageNet dataset to compute mid-level image representation for images in the PASCAL VOC dataset. We show that despite differences in image statistics and tasks in the two datasets, the transferred rep- resentation leads to significantly improved results for object and action classification, outperforming the current state of the art on Pascal VOC 2007 and 2012 datasets. We also show promising results for object and action localization.