Weakly Supervised Deep Detection Networks

Zhang, Liqiang

June 6, 2018

Abstract

Weakly supervised learning of object detection is an important problem in image understanding that still does not have a satisfactory solution. In this paper, we address this problem by exploiting the power of deep convolutional neural networks pre-trained on large-scale image-level classification tasks. This paper propose a weakly supervised deep detection architecture that modifies one such network to operate at the level of image regions, performing simultaneously region selection and classification. The model, which is a simple and elegant end-to-end architecture, outperforms standard data augmentation and fine-tuning techniques for the task of image-level classification as well.

1. Introduction

In recent years, Convolutional Neural Networks (CNN) have emerged as the new state-of-the-art learning framework for image recognition. Key to their success is the ability to learn from large quantities of labelled data the complex appearance of real-world objects. One of the most striking aspects of CNNs is their ability to learn generic visual features that are generalise to many tasks [1]. In particular, CNNs pre-trained on datasets such as ImageNet ILSVRC have been shown to obtain excellent results in recognition in other domains, in object detection, in semantic segmentation, in human pose estimation, and in many other tasks.

In this paper, the author contribute a novel end-to-end method for weakly supervised object detection using pre-trained CNNs which they call a weakly supervised deep detection network (WSDDN) in Fig. 1. their method starts from an existing network, such as AlexNet pre-trained on ImageNet data, and extends it to reason explicitly and efficiently about image regions R [2]. In order to do so, given an image x, the first step is to efficiently extract region-level descriptors $\phi(x; R)$ by inserting a spatial pyramid pooling layer on top of the convolutional layers of the CNN. The recognition and detection scores computed for all the image regions are finally aggregated in order to predict the class of the image as a whole, which is then used to inject image-level supervision in learning.

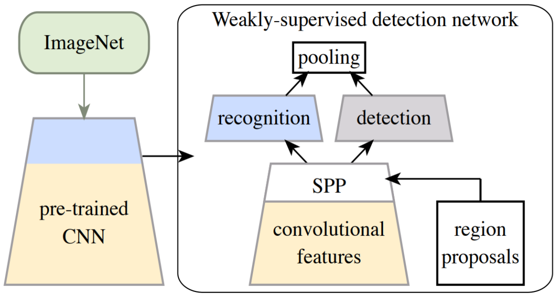

Figure 1. The method starts from a CNN pre-trained for image classification on a large dataset, e.g. ImageNet. It then modifies to reason efficiently about regions, branching off a recognition and a detection data streams. The resulting architecture can be fine-tuned on a target dataset to achieve state-of-the-art weakly supervised object detection using only image-level annotations.

2. Method

The first data stream performs classification of the individual regions, by mapping each of them to a C-dimensional vector of class scores, assuming that the system is trained to detect C different classes. This is achieved by evaluating a linear map ϕ_{fc8c} and results in a matrix of data $x_c \in \mathbb{R}^{C \times |R|}$, containing the class prediction scores for each region. The latter is then passed through a softmax operator, defined as follows:

$$[\sigma_{\text{class}}(x^c)]_{ij} = \frac{e^{x^c_{ij}}}{\sum_{k=1}^{C} e^{x^c_{ij}}}. \tag{1}$$

The second data stream performs instead detection, by scoring regions relative to one another. This is done on a class-specific basis by using a second linear map ϕ_{fc8d}, also resulting in a matrix of scores $x^d \in \mathbb{R}^{C \times |R|}$. It is then passed through another softmax operator, but this time defined as follows:

$$[\sigma_{\mathrm{det}}(x^d)]_{ij} = \frac{e^{x^d_{ij}}}{\sum_{k=1}^{|R|} e^{x^d_{ik}}}. \tag{2}$$

While the two streams are remarkably similar, the introduction of the σ_{class} and σ_{det} non-linearities in the classification and detection streams is a key difference which allows to interpret them as performing classification and detection, respectively [3]. In the first case, in fact, the softmax operator compares, for each region independently, class scores, whereas in the second case the softmax operator compares, for each class independently, the scores of different regions. Hence, the first branch predicts which class to associate to a region, whereas the second branch selects which regions are more likely to contain an informative image fragment.

3. Experiments

	S	**M**	**L**	**Ens.**
SSW	31.1	30.9	24.3 3	3.3
EB	31.5	30.9	25.5	34.2
EB + Box Sc.	33.4	32.7	30.4	36.7
EB + Box Sc. + Sp. Reg.	**34.5**	**34.9**	**34.8**	**39.3**

Table 1. **VOC 2007** test detection average precision (%). The ensemble network is denoted as **Ens.**

Table 1 shows that WSDDN with individual models **S** and **M** are already on par with the state-ofthe-art method and the ensemble outperforms the best previous score in the VOC 2007 dataset. Differently from supervised detection methods, detection performance of WSDDN does not improve with use of wider or deeper networks. In contrast, model **L** performs significantly worse than models **S** and **M** [4].

References

[1] H. Bilen, M. Pedersoli, and T. Tuytelaars. Weakly supervised object detection with posterior regularization. In *BMVC*, pages 1997–2005, 2014. 1

[2] H. Bilen, M. Pedersoli, and T. Tuytelaars. Weakly supervised object detection with convex clustering. In *CVPR*, pages 1081–1089, 2015. 1

[3] R. G. Cinbis, J. Verbeek, and C. Schmid. Weakly supervised object localization with multi-fold multiple instance learning. *arXiv preprint arXiv:1503.00949*, 2017. 2

[4] B. Hariharan, P. Arbelez, R. Girshick, and J. Malik. Simultaneous detection and segmentation. In *ECCV*, pages 297–312, 2014. 2