International Journal of Scientific Research in Computer Science, Engineering and Information Technology
ISSN : 2456-3307 (www.ijsrcseit.com)
doi : https://doi.org/10.32628/IJSRCSEIT

# Self Paced Deep Learning for Weakly Supervised Object Detection

**Mangineni Prasanna*[1], Dr. G. Nirmala[2]**
*[1]M. Tech Student, Sir C R Reddy College of Engineering, Eluru, Andhra Pradesh, India
[2]Associate Professor, Department of CSE, Sir C R Reddy College of Engineering, Eluru, Andhra Pradesh, India

**Article Info**
Volume 8, Issue 1
Page Number : 296-300

**Publication Issue :**
January-February-2022

**Article History**
Accepted : 05 Jan 2022
Published: 30 Jan 2022

## ABSTRACT

In a weakly-supervised scenario object detectors need to be trained using image-level annotation alone. Sincebounding-box-level ground truth is not available, most of the solutions proposed so far are based on an iterative, Multiple Instance Learning framework in which the current classifier is used to select the highest-confidence boxes in each image, which are treated as pseudo-ground truth in the next training iteration. However, the errors of an immature classifier can make the process drift, usually introducing many of false positives in the training dataset. To alleviate this problem, we propose in this paper a training protocol based on the self-paced learning paradigm. The main idea is to iteratively select a subset of images and boxes that are the most reliable, and use them for training.

Keywords - Weakly Supervised Learning, Object Detection, Self-Paced Learning, Curriculum Learning, Deep Learning, Training Protocol.

## I. INTRODUCTION

A well-known problem in object detection is the fact that collecting ground truth data (i.e., object-level annotations) for training is usually much more time consuming and expensive than collecting image-level labels for object classification. This problem is exacerbated in the context of the current deep networks, which need to be trained or "fine-tuned" using large amounts of data. Weakly-supervised techniques for object detection (WSD) can alleviate the problem by leveraging existing datasets which provide image-level annotations only. In the common Multiple Instance Learning (MIL) formalization of the WSD problem, an image I, associated with a label of a given class y, is described as a "bag" of Bounding Boxes (BBs), where at least one BB is a positive sample for y and the others are samples of the other classes (e.g., the background class).

The main problem is how can the classifier, while being trained, automatically guess what the positives in I are. A typical MIL-based solution alternates between 2 phases: (1) optimizing the classifier's parameters, assuming that the positive BBs in each image are known, and (2) using the current classifier to predict the most likely positives in each image [2]. However, a well known problem of MIL-like

solutions is that if the initial classifier is not strong enough, this process can easily drift. For instance, predicted false positives (e.g., BBs on the background) can ake the classifier learn something different than the target class.


Fig. 1. A schematic illustration of how the training dataset increasing recognition skills of the trained network

## II. RELATED WORK

Many recent studies have shown that selecting a subset of "good" samples for training a classifier can lead to better results than using all the samples [1]. A pioneering work in this direction is the curriculum learning approach proposed in [2]. The authors show that suitably sorting the training samples, from the easiest to the most difficult, and iteratively training a classifier starting with a subset of easy samples progressively augmented with more and more difficult samples), can be useful to find better local minima. In [3], easy and difficult images (taken from datasets known to be more or less "difficult")are provided for training a Convolutional Neural Network (CNN) in order to learn generic CNN features using webly annotated data. In [4], different and progressively more complex CNNs are trained for a segmentation task, using more and more difficult data samples together with the output of the previously learned networks. It is worth noting that in thes and in all the other curriculum-learningbased approaches, the order of the samples is decided using additional supervisory information usually provided by a human teacher. Unfortunately, these "image-easiness" metadata are not available for the common large-scale datasets. Curriculum learning was extended to self-paced learning in [5]. The main difference between the two paradigms is that in self-paced learning the order of the samples is automatically computed and it is a priori unknown. The selection of the best "easy" sample set for training is, generally speaking, untractable (it is a subset selection problem). The solution proposed in [6] is based on a continuous relaxation of the problem's constraints which leads to a biconvex optimization of a Structural SVM.

## III. FAST-RCNN AND NOTATION

In this section we review the main aspects of the Fast-RCNN [7] approach which are important to understand our proposal and we introduce notations, used in the rest of the paper.

The network takes as input an image I (raw pixels) and a set of BBs on I: B(I) = fb1; :::; bng . B(I) is computed using an external tool, which usually selects image subwindows taking into account their "abjectness": for instance using Selective Search [8] (also used in all our experiments). If f is the function computed by the network, its outcome is a set of detections:

$$f(I, B(I)) = \{d_{ic}\}_{i=1,\ldots,n, c=1,\ldots,C}, \qquad (1)$$

For more details we refer the reader to [9]. What is important to highlight here is that Fast-RCNN is a strongly supervised method. Conversely, in our weakly-supervised scenario, we do not have BB-level annotations. Hence, in the rest of the article we assume that our training set is T =f(I1; Y1); :::; (Ij; Yj); :::; (IN ; YN )g, where Yj = fy1; :::ynj g is the set of labels associated with image Ij and the number

Since object-level ground truth is not given, we use the network (in the current self-paced training

iteration) to compute the most likely positions of the objects in $I_j$. In the next section we show how these locations are computed and how T is updated following a self-paced learning strategy.

## IV. SELF-PACED LEARNING PROTOCOL

We call W the set of weights of all the layers of the net-work and we initialize our network with $W_0$, which can be obtained using any standard object classification network, trained using only image-level information. At the end of this section we provide more details on how $W_0$ is obtained.

**Algorithm 1** Self-Paced Weakly Supervised Training

**Input:** $T$, $W_0$, $r_1$, $M$
**Output:** Trained network $f_{W_M}$

For $t := 1$ to $M$:
  $P := \emptyset, T_t := \emptyset$
  For each $(I, Y) \in T$:
    Compute $(s_y^I, z_y^I)$ using Eq. 2
    If $y \in Y$, then: $P := P \cup \{(I, s_y^I, z_y^I, y)\}$
  For each $c \in \{1, ..., C\}$ compute $e(c)$ using Eq. 3
  $C_t := r_t C$
  Let $S = \{c_1, c_2, ...\}$ be the subset of the $C_t$ easiest classes according to $e(c)$
  Remove from $P$ those tuples $(I, s, z, y)$ s.t. $y \notin S$
  $N_t := min(r_t N, |P|)$
  Let $P'$ be the $N_t$ topmost tuples in $P$ according to the $s$-score
  For each $(I, s, z, y) \in P' : T_t := T_t \cup \{(I, \{(y, z)\})\}$
  $V_0 = W_{t-1}$
  For $t' := 1$ to $N_t$:
    Randomly select $(I_1, \{(y_1, z_1)\}), ..., (I_m, \{(y_m, z_m)\}) \in T_t$
    Compute a mini-batch $MB$ of BBs using $(I_1, \{(y_1, z_1)\}), ..., (I_m, \{(y_m, z_m)\})$
    Compute $V_{t'}$ using $MB$ and back-propagation on $f_{V_{t'-1}}$
  $W_t := V_{N_t}$
  $r_{t+1} = r_t + \frac{1-r_1}{M}$

The proposed self-paced learning protocol of the net-work is composed of a sequence of self-paced iterations. At a self-paced iteration t we use the current network $f_{Wt1}$ in order to select a subset of easy classes and easy samples of these classes. The result is a new training set $T_t$ which is used to train a new model $W_t$. $W_t$ is obtained using the "standard" training procedure of the Fast-RCNN (Sec. 3), based on mini-batch SGD, but it is applied to $T_t$ only and iterated for only $N_t$ mini-batch SGD iterations, $N_t$ being the cardinality of $T_t$. Note that a mini-batch SGD iteration is different from a self-paced iteration and in each SGD iteration a mini-batch of BBs is formed using the pseudo-ground truth obtained using $f_{Wt1}$ . The proposed protocol is summarized in Alg. 1 and we provide the details below.

**Computing the latent boxes.** Given an image I, its label set Y and the current network $f_{Wt1}$, first we compute:

$$(s_y^I, z_y^I) = \underset{(s_{ic}, p_{ic}) \in f_{W_{t-1}}(I, B(I))}{\arg\max} s_{ic} \qquad (2)$$

In Eq. 2, $(s^I_y; z_y{}^I)$ is the detection in f(I; B(I)) with the highest score $(s^I_y)$ with respect to all the detections obtained starting from B(I) and the subscript y indicates the corresponding class. $z_y{}^I$ is a latent box which specifies the most likely position of an object of the "winning" class y in image I according to $f_{Wt1}$ . Note that the background class is not included in $f_{Wt1}$ (see Eq. 1), thus y 2 f1; :::; Cg.

## V. ANALYSIS OF ASPECTS OF PROTOCOLS

In this section we analyse the influence of different elements of our proposed training protocol by separately removing or modifying important parts of Algorithm 1.

### 1.1 Simplified versions of the training protocol

**Basic-MIL.** In the experiments of this subsection we useboth Pascal VOC 07 and ILSVRC 2013. We start with comparing our method (Self-Paced, SP) with a MIL-based solution (MIL), where: (a) all the images in T are used and (b) in each image the latent boxes are computed by iteratively maximizing the class-specific score of the current iteration's model. Thus, we remove from Alg. 1 all those steps which concern image (and class) selection. Moreover, we also remove the inter-classifier competition, and we in-dependently select the top score box for each label in Y . More in detail, given (I; Y ) 2 T , for each y 2 Y we separately compute:

$$(s_y, z_y) = \underset{\substack{(s_{ic}, p_{ic}) \in f_{W_{t-1}}(I, B(I)), \\ c=y}}{\arg\max} s_{ic}$$

TABLE 8: MAP (%) on Pascal VOC 2007 test computed with different networks $f_{W_t}$ and with respect to different versions of our training protocol and M + 1 iterations.

| Method | $W_0$ | $W_1$ | $W_2$ | $W_3$ | $W_4$ | $W_5$ |
|---|---|---|---|---|---|---|
| MIL | 31.9 | 33.6 | 32.1 | 32.2 | 30.8 | 30.9 |
| Curriculum | 31.9 | 31.3 | 33.8 | 31.6 | 31.3 | 30.5 |
| SP-all-cls | 31.9 | 36.6 | 36.9 | 36.6 | 36.9 | 36.9 |
| SP-rnd-cls | 31.9 | 32.3 | 31.6 | 32.4 | 32.7 | 33.8 |
| No-reg-train | 31.9 | 31.2 | 32.6 | 33.1 | 33.5 | 34.4 |
| No-reg-train-test | 28.3 | 28.3 | 30.1 | 30.9 | 30.7 | 31.4 |
| SP | 31.9 | 35.3 | 37.6 | 37.8 | 38.1 | 38.1 |

TABLE 9: map (%) on ILSVRC 2013 val2 computed with different networks $f_{W_t}$ and with respect to different versions of our training protocol and M + 1 iterations.

| Method | $W_0$ | $W_1$ | $W_2$ | $W_3$ | $W_4$ | $W_5$ |
|---|---|---|---|---|---|---|
| MIL | 9.54 | 9.66 | 9.01 | 8.97 | 8.59 | 8.7 |
| Curriculum | 9.54 | 9.08 | 9.15 | 8.77 | 8.89 | 8.97 |
| SP-all-cls | 9.54 | 10.68 | 10.74 | 11.77 | 11.97 | 12.06 |
| SP | 9.54 | 10.88 | 11.87 | 12.01 | 12.13 | 11.87 |

### 1.2 Multi-label versions of the training protocol

This subsection is dedicated to evaluating the importance of the inter-classifier competition. As explained in Sec. 4 the inter-classifier competition is used in SP to reduce the amount of noisy training boxes by selecting only one box $z_y^I$ per image I, according to the current most confident classifier -(y) on I.

### 1.3 Precision of the selected subsets of training data

In this subsection we evaluate the number of "correct" samples selected for training the network. To this aim we adopt the evaluation protocol suggested in [9], where the authors use ILSVRC 2013 val1 and a Precision metric. The latter is similar to CorLoc, the difference being that in CorLoc one latent box ($z_y$) is computed for each label y 2 Y associated with a training image, while Precision is based on extracting one single latent box ($z_y^I$ ) per image. Using Precision @0:5 IoU we can measure the quantity of latent boxes actually used during training which sufficiently overlap with a real ground truth box with the correct class.

## VI. CONCLUSIONS

We proposed a self-paced learning based protocol for deep networks in a WSD scenario, aiming at reducing the amount of noise while training the DN. Our training protocol extends the self-paced learning paradigm by introducing: (1)Inter-classifier competition as a powerful mechanism to reduce noise, (2) class-selection, in which the easiest classes are trained first, and (3) the use of the Fast-RCNN regression layer for the implicit modification of the bag of boxes.

## VII. REFERENCES

[1]. L. Bazzani, A. Bergamo, D. Anguelov, and L. Torresani. Self-taught object localization with deep networks. In IEEE Winter Conference on Applications of Computer Vision (WACV), 2016.

[2]. Y. Bengio, J. Louradour, R. Collobert, and J. Weston. Curriculum learning. In ICML, pages 41–48, 2009.

[3]. H. Bilen, M. Pedersoli, and T. Tuytelaars. Weakly supervised object detection with posterior regularization. In BMVC, 2014.

[4]. H. Bilen, M. Pedersoli, and T. Tuytelaars. Weakly supervised object detection with convex clustering. In CVPR, 2015.

[5]. H. Bilen and A. Vedaldi. Weakly supervised deep detection networks. In CVPR, 2016.

[6]. K. Chatfield, K. Simonyan, A. Vedaldi, and A. Zisserman. Return of the devil in the details: Delving deep into convolutional nets. In BMVC, 2014.

[7]. X. Chen and A. Gupta. Webly supervised learning of convolu-tional networks. arxiv:1505.01554, 2015.

[8]. R. G. Cinbis, J. J. Verbeek, and C. Schmid. Multi-fold MIL training for weakly supervised object localization. In CVPR, pages 2409– 2416, 2014.

[9]. R. G. Cinbis, J. J. Verbeek, and C. Schmid. Weakly supervised object localization with multi-fold multiple instance learning. IEEE Trans. Pattern Anal. Mach. Intell., 39(1):189–203, 2017.

**Cite this article as :**

Mangineni Prasanna, Dr. G. Nirmala, "Self Paced Deep Learning for Weakly Supervised Object Detection", International Journal of Scientific Research in Computer Science, Engineering and Information Technology (IJSRCSEIT), ISSN : 2456-3307, Volume 8 Issue 1, pp. 296-300, January-February 2022.
Journal URL : https://ijsrcseit.com/CSEIT228210