Weakly Supervised Object Detection with Posterior Regularization

Hakan Bilen
hakan.bilen@esat.kuleuven.be
Marco Pedersoli
marco.pedersoli@esat.kuleuven.be
Tinne Tuytelaars
tinne.tuytelaars@esat.kuleuven.be

KU Leuven, ESAT-PSI, iMinds
Leuven, Belgium

Motivation: In weakly supervised object detection where only the presence or absence of an object category as a binary label is available for training, the common practice is to model the object location with latent variables and jointly learn them with the object appearance model [1, 5]. An ideal weakly supervised learning method for object detection is expected to guide the latent variables to a solution that disentangles object instances from noisy and cluttered background. The learning algorithm should lead the appearance model and the latent variables to best explain the correlation between the training images and their binary labels. However, without complete supervision, maximizing the likelihood of observed data or minimizing the data-dependent cost function during training may result in latent variables that do not capture the expected regularities.

Contributions: In this paper, (i) we show that in a weakly-supervised setting, regulating the latent distribution and properly driving the latent variables are crucial for good performance and lead to state-of-the-art results in both classification and detection, (ii) we show how to introduce in the weakly supervised detection specific prior knowledge that helps to drive the latent variables by means of posterior regularization, and (iii) we better model the weakly-supervised object detection problem via the soft-max where multiple objects in the same image are considered and at the same time the optimization is smoother.

We focus on domain specific prior knowledge for object detection. In particular we exploit the fact that (i) each horizontal mirror of an object is still a valid object (*object symmetry*) and (ii) the same spatial region (in our case a bounding box) cannot represent more than one object class (*mutual exclusion*). We incorporate this prior knowledge via posterior regularization as proposed in [4].

Results: We evaluate our method and compare its performance to previous work [2, 6, 7] in the Pascal VOC 2007 dataset [3]. We first illustrate hard-max and soft-max outputs in Fig. 1, the posterior regularization on symmetry and mutual exclusion in Fig. 2 and Fig. 3 resp. We also report quantitative results in detection and classification tasks in Table 1 and 2 resp. We show the contribution of each added component and compare the final result to the state-of-the-art methods in both detection and classification.

[1] H. Bilen, V.P. Namboodiri, and L. Van Gool. Object and action classification with latent window parameters. *IJCV*, pages 1–15, 2013.

[2] R.G. Cinbis, J. Verbeek, and C. Schmid. Multi-fold mil training for weakly supervised object localization. In *CVPR*, 2014.

[3] M. Everingham, A. Zisserman, C. K. I. Williams, and L. Van Gool. The PASCAL Visual Object Classes Challenge 2007 (VOC2007) Results.

[4] K. Ganchev, J. Graça, J. Gillenwater, and B. Taskar. Posterior regularization for structured latent variable models. *The Journal of Machine Learning Research*, 11:2001–2049, 2010.

[5] M.H. Nguyen, L. Torresani, F. De la Torre, and C. Rother. Weakly supervised discriminative localization and classification: a joint learning process. In *ICCV*, 2009.

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

[7] H.O. Song, R. Girshick, S. Jegelka, J. Mairal, Z. Harchaoui, and T. Darrell. One-bit object detection: On learning to localize objects with minimal supervision. *arXiv preprint arXiv:1403.1024*, 2014.

(a) images (b) hard-max output (c) soft-max output

Figure 1: Visual comparison of max-margin and soft-max margin learning on representative "cow" and "chair" images. While max outputs a single window, soft-max marginalizes over all windows and better represents multiple instances.

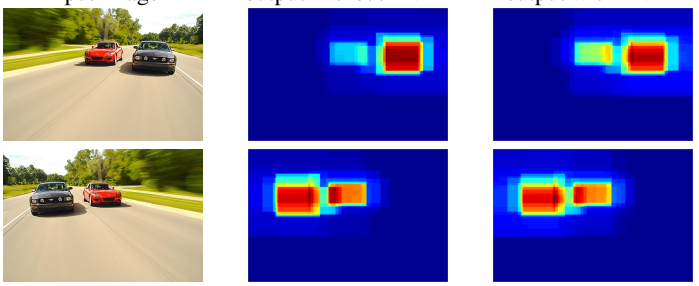

Figure 2: Output maps of "car" detectors on test and flipped images without and with posterior regularization for symmetry. Learning with the symmetrical constraints increase the scores of less confident detections.

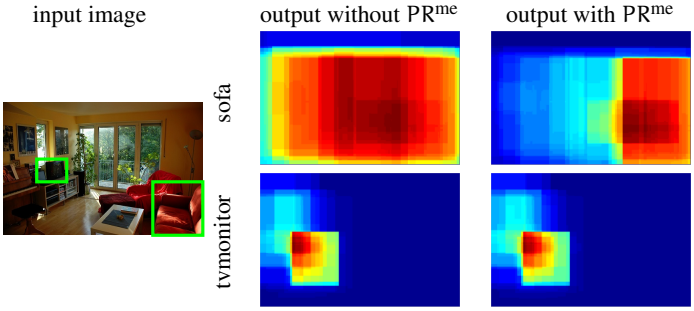

Figure 3: Output maps of "sofa" and "tvmonitor" detectors for input images. Adding the mutual exclusion constraint helps to separate two distributions by penalizing the bounding boxes with high probability for both detectors.

Ours					Others	
hard-max	soft-max	+flip	+PR^{sym}	+PR^{me}	[2]	[7]
22.7	24.0	24.8	26.0	**26.4**	22.4	22.7

Table 1: Weakly supervised detection results on the Pascal VOC 2007 in mean average precision (mAP). +flip indicates of adding horizontally mirrored training images to the training. PR^{sym} and PR^{me} denote the posterior regularization for symmetry and mutual exclusion. The components starting from +flip are consecutively added on the soft-max. Our method outperforms the state-of-the-art weakly supervised detectors [2, 7].

Ours			Others	
SVM	hard-max	Full	[2]	[6]
74.1	77.1	**80.9**	65.6	77.7

Table 2: Classification results on the Pascal VOC 2007 in mAP. SVM denotes training linear SVMs without any localization. hard-max and Full denote the latent SVM formulation and our full model. Our method outperforms the state-of-the-art classifiers [2, 6].