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

Abstract

This paper focuses on the problem of object detection when the annotation at training time is restricted to presence or absence of object instances at image level. We present a method based on features extracted from a Convolutional Neural Network and latent SVM that can represent and exploit the presence of multiple object instances in an image. Moreover, the detection of the object instances in the image is improved by incorporating in the learning procedure additional constraints that represent domain-specific knowledge such as symmetry and mutual exclusion. We show that the proposed method outperforms the state-of-the-art in weakly-supervised object detection and object classification on the Pascal VOC 2007 dataset.

1 Introduction

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 a model for the object appearance [2, 15, 19]. An ideal weakly supervised learning algorithm 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, in practice, 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. Thus, it is crucial to regulate the latent distribution such that they encode useful structures of the data.

A possible way to properly learn the latent variables is to provide a meaningful initialization so that the optimization does not get stuck in a local minimum. In weakly supervised object detection the initialization of the latent variables involves providing good initial location for each object instance. Initialization strategies for the latent variables in weakly supervised object detection have been presented in several works [3, 15, 24]. However, this is an intertwined problem because if we knew where the object is, the task would no longer be a weakly supervised problem. Other methods [14, 17] focus on the optimization procedure,

to make it smooth and as convex as possible (continuation methods). Although we do not deny the importance of the two previously mentioned factors, we argue that, for an effective learning of the latent variables and improved localization performance, controlling the latent variable distribution is crucial. In particular we focus on relaxing too strict constraints of the model and we introduce additional prior knowledge that has not been previously considered.

In [2, 15] it is implicitly assumed that each image contains only one object. This is a very restrictive assumption that does not allow the method to exploit the complete training data because additional instances of the same class in the same image are discarded. Moreover, and even worse, the additional instances are considered as background and therefore the model is forced to learn them as "negative samples". An elegant way to overcome this problem is by substituting the maximization over the latent variables with a soft-max function. The soft-max is a smooth, differentiable approximation of the max function that amplifies high scoring configurations and reduces the influences of the low scoring ones. By definition, it does not restrict the number of "chosen" elements. In this way we have the two-fold advantage of (i) being able to consider multiple objects in a single image and (ii) using a smooth function that makes the optimization easier.

Furthermore, prior knowledge about the problem is a valuable source of information that can be exploited to better drive the latent variables. However, generally this additional general knowledge about the problem is not considered because either it is difficult to represent or it makes the inference computationally very expensive. Instead we include this prior knowledge via posterior regularization as proposed in [11]. In our case we want to 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*).

To exploit the first point the commonly used solution is to add horizontally flipped versions of the training images as extra training data. While the additional data guides the model to score high for some latent configuration on both original and flipped images, it does not enforce to select the same (but mirrored) location for both images. We can say that the problem is *too relaxed*. On the other side one can localize by jointly evaluating an image part with its mirror together. In this case the algorithm can evaluate only perfectly symmetric representations. This limits the optimization space and makes the optimization more difficult, *i.e* the problem is *too constrained*. Instead, we introduce a posterior regularization that enforces the scores on each bounding box and its mirrored one to have the similar values. In section 4 we show the advantage of our formulation compared with the others.

Spatial co-occurrence of multiple objects has been exploited in [7] for supervised object detection. Desai *et al*. [7] learn the relative spatial relationships between pairs of object instances and categories. However, this learning of relationships relies on the supervision of the object locations and involves a computationally expensive optimization. For the weakly supervised case, jointly estimating the object localizations and interactions between different instances is a significantly more challenging problem. Therefore instead of parameterizing the spatial relations and learning those parameters, we enforce an additional posterior regularization strategy [11]. Namely, we transform the individual binary detection problems for all object categories into a single multi-label problem and impose a regularization term that discourages configurations where multiple categories have high probability for the same bounding box. In section 4 we empirically show the contribution of this proposed regularization.

The main contributions of the paper are: (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.

The rest of the paper is structured as follows: in section 2 we discuss related work. In section 3 we formally introduce our method for learning and inference. Finally experiments are detailed in section 4 and conclusions are drawn in section 5.

2 Related Work

Discriminative supervised classifiers have been proven to be very effective and accurate tools for learning the correlation between input and precisely annotated outputs. In the literature there has been a substantial amount of work that proposes weakly supervised algorithms for classification. The proposed weakly supervised method aims to jointly label the missing annotations and learn a classifier based on these labellings. This results in solving a non-convex problem which is typically optimized via an alternating Expectation Maximization (EM) kind of algorithm. Due to the non-convexity, these methods are prone to converge into a local minimum. To overcome the problems, previous work can be divided into two groups that focus on clever initialization strategies and on converting the optimization into a convex or smooth one, which is in general easy to optimize.

The non-convex optimization is notoriously known for its sensitivity to initialization. Many previous works [3, 15, 24] focus on developing initialization strategies for the weakly supervised detection problem. Deselaers *et al.* [8] initialize the latent variables with the highest "objectness" score [1]. Kumar *et al.* [15] propose a self paced method that starts learning on a small set of samples and gradually adds harder samples to training. Bilen *et al.* [3] employ exemplar background detectors to initialize latent variables. Song *et al.* [24] initialize the latent variables by discovering discriminative and relevant windows via a sub-modular search. Cinbis *et al.* [5] show that an iterative multi-fold multiple instance learning (MIL) prevents overfitting on the initial estimates.

The second line of research focuses on developing smoother approximations of the non-convex problem and thus making the problem less sensitive to initialization. Joulin *et al.* [14] propose a convex relaxation method and solve the transformed problem with semi-definite programming. Miller *et al.* [17] employ an entropy measure to minimize the uncertainty over the distribution of latent parameters. Song *et al.* [24] apply the smoothing technique of Nesterov [18] to the latent SVM's object function.

Designing initialization strategies and smoothing the non-convex optimization problems have proven to be useful in many learning applications. Likewise, we show that smoothing the learning provides a better learning for object detection and improves the results. This improvement is also related to the less constrained modelling of the problem, where multiple objects in the same image can be considered. Moreover, in this paper, inspired by [11], we focus on efficiently incorporating indirect supervision via prior knowledge. Ganchev *et al.* [11] introduce a probabilistic framework that can express and enforce data-dependent constraints on latent variable distribution with posterior regularization. The framework is shown to improve the performance on natural language processing problems such as word alignment and multi-view learning. Here we show that posterior regularization is also useful for weakly-supervised detection.

3 Learning and Inference

In this section we formally introduce the problem of learning to localize object classes in images as a weakly-supervised discriminative problem based on latent structural SVM [26] (section 3.1), how to relax the problem (section 3.2) and finally how to include posterior regularization (section 3.3).

3.1 Latent Structural SVM optimization

Let $x \in \mathcal{X}$, $y \in \{-1,1\}$ and $h \in \mathcal{H}$ denote the input, output and latent variables of our problem respectively. In our context, x, y, h correspond to an image, its label and a window (or bounding box). Following the setting in [26], we consider a linear prediction rule in a joint input/output/latent space $\mathcal{O}$. We define an input/output/latent mapping $\Phi(x,y,h) \in \mathcal{X} \times \mathcal{Y} \times \mathcal{H} \rightarrow \mathcal{O} \subset \mathcal{R}^d$ such that

$$\Phi(x,y,h) = \begin{cases} \vec{0} & \text{if } y \leqslant 0 \\ \phi(x,h) & \text{if } y > 0 \end{cases}$$

where $\vec{0}$ is a d dimensional 0 vector and $\phi(x,h)$ is a representation of a window h cropped from image x, *i.e.* DeCAF features [6] with $d = 4096$ in our experiments.

The prediction y^* of a previously unseen input x is for a given parameter vector $w \in \mathcal{R}^d$ defined as:

$$y^* = \arg\max_{y \in \mathcal{Y}} (\max_{h \in \mathcal{H}} w \cdot \Phi(x,y,h)). \tag{1}$$

Given a set of training samples $\mathcal{S} = \{(x^i, y^i), i = 1, \ldots, N\}$, the learning task is to find the parameter vector w^* that best predicts the outputs for the inputs. For a given w, the loss on the i-th sample is measured by a loss function $l(w, x^i, y^i)$. The empirical loss for the training samples with an additional ℓ_2 regularizer (denoted as $\|.\|$) is given as

$$\mathcal{L}(w;\mathcal{S},\lambda) = \frac{1}{2\lambda}\|w\|^2 + \sum_{i=1}^{n} l(w, x^i, y^i) \tag{2}$$

where λ is the regularization constant. The optimal parameter vector w^* can be found by minimizing the regularized loss

$$w^* = \arg\min_{w} \mathcal{L}(w;\mathcal{S},\lambda) \tag{3}$$

There are various ways to design the measure of mismatch $l(w, x^i, y^i)$ between the input and output. One popular measure is the max-margin formulation of [26]. Its margin rescaling form with latent variables amounts to:

$$l_m(w, x^i, y^i) = \max_{y,h} \left(w \cdot \Phi(x^i, y, h) + \Delta(y^i, y) \right) - \max_{h} w \cdot \Phi(x^i, y^i, h) \tag{4}$$

where $\Delta(y^i, y)$ is zero-one error, *i.e.* $\Delta(y^i, y) = 0$ if $y^i = y$, 1 else. The basic idea is to learn the weights such that score of (x^i, y^i, h^*) with $h^* = \arg\max_h w \cdot \Phi(x^i, y^i, h)$ is bigger than (x^i, y, h) for all alternatives $y \in \mathcal{Y} \setminus \{y^i\}$ and $h \in \mathcal{H}$ with a larger margin for those y with larger loss.

3.2 Relaxing the problem with soft-max

While the max-margin latent learning is justified for various applications, in the context of weakly supervised detection, it is somehow restricted to exploit only a single positive bounding box per image, since each positive image is represented only by the max scoring latent parameter. In order to alleviate this restriction, we replace the "hard" max in (4) with the "soft" max as in [12, 23]:

$$l_s(w, x^i, y^i) = \log \sum_{y,h} \exp\left(w \cdot \Phi(x^i, y, h) + \Delta(y^i, y)\right) - \log \sum_{h} \exp\left(w \cdot \Phi(x^i, y^i, h)\right) \quad (5)$$

Unlike Eq. 4 that finds the max scoring latent variable, the soft-max margin approach Eq. 5 marginalizes over the latent variable space $\mathcal{H}$. Thus it can represent an image with multiple latent variables and utilize more positive data in training. At the same time, using soft-max helps to make Eq. 2 smoother and this enables the use of fast gradient-based Quasi-Newton optimization techniques such as the LBFGS algorithm [16]. One can still use the alternating optimization techniques such as Concave Convex Procedure (CCCP) [27] or Block Coordinate descent [10] with these Quasi-Newton methods. However, we empirically find that directly optimizing the soft-max formulation with the LBFGS converges faster and gives slightly better performance for our experiments.

3.3 Posterior Regularization

Both the max and soft-max margin formulations that are based on a dot product between the parameter vector w and the joint feature map Φ. So they are in practice very simplistic models for object detection. On the positive side, we have the aforementioned optimization tools to efficiently minimize the optimization problem in Eq. 3. However, without direct supervision, an inherent problem with such learning is that the optimization may not guide the latent variables towards the intended regularities of the application at hand. We focus on incorporating regularization strategies on latent variables arising from prior knowledge. To do so, we define two regularization terms that avoid certain latent configurations by restricting the model space.

The symmetry term is based on the assumption that if an image part contains an object instance, its horizontal mirroring should also be a valid representation for the object class. Horizontally mirroring training images and adding them to learning is a common practice to augment the training data and to boost the results. One would expect a learning algorithm to have the same posterior distribution of latent variables for an image and its flipped version in object detection. However, this is usually not the case unless one uses perfectly invariant features to flipping. Here, we propose a way to impose such behavior. We first define the posterior for a latent parameter given an input x, its label y and parameter vector w by using the soft-max formulation of Eq. 5:

$$p(h|x, y, w) = \frac{\exp\{w \cdot \Phi(x, y, h)\}}{\sum_{h \in \mathcal{H}} \exp\{w \cdot \Phi(x, y, h)\}}. \quad (6)$$

Then we introduce a Kullback-Leibler (KL) divergence term that encourages similarity between the probability of latent variables between a positive (*i.e.* $y^i = 1$) image (x^i) and its flipped version ($\bar{x}^i$):

$$D_{kl}(x^i, \bar{x}^i, y^i, w) = \sum_{h \in \mathcal{H}} p(h|x^i, y^i, w)\left(\log(p(h|x^i, y^i, w)) - \log(p(\bar{h}|\bar{x}^i, y^i, w))\right) \quad (7)$$

where h and $\bar{h}$ correspond to a window and its flipped version. The posterior regularization term PR^{sym} for symmetry is written as a symmetrized KL divergence for (x^i, y^i):

$$PR^{sym}(x^i, y^i, w) = \frac{1}{2}\left(D_{kl}(x^i, \bar{x}^i, y^i, w) + D_{kl}(\bar{x}^i, x^i, y^i, w)\right). \tag{8}$$

The mutual exclusion term is used when multiple object categories are present in an image. We use the prior knowledge that instances of different object categories do not lie in the same part of the image (same window), therefore we penalize the cases when a window is highly probable for more than one class. This decision involves a joint optimization for all classes (*i.e.* a multi-label setting). Let K denote the number of object categories and $Y^i \in \{-1,1\}^K$ denote a K dimensional binary array that contains the labels of image x^i such that $Y^i_k \in \{-1,1\}$ is the label for the object category k. Modifying the loss in Eq. 5 for the multi-label setting gives:

$$l^{ml}_s(w, x^i, Y^i) = \sum_{k=1}^{K} l_s(w_k, x^i, Y^i_k), \tag{9}$$

where w_k is the parameter vector of the object category k and w is the concatenation of the w_k for all categories. Now we can define the mutual exclusion term for positive label pairs as:

$$PR^{me}(x^i, Y^i, w) = \sum_{\substack{1 \leqslant j < k \leqslant K \\ Y^i_j, Y^i_k = 1}} \sum_{h \in \mathcal{H}} \sqrt{p(h|x^i, Y^i_j, w_j)\, p(h|x^i, Y^i_k, w_k)} \tag{10}$$

This term produces large costs when a window h of an image x has high probability for multiple labels. We rewrite our cost function in Eq. 2 by considering Eq. 8, 9 and 10:

$$\mathcal{L}(w; \mathcal{S}, \lambda) = \frac{1}{2\lambda}\|w\|^2 + \sum_{i=1}^{N}\left(l^{ml}_s(w, x^i, Y^i) + \sum_{\substack{k=1 \\ Y^i_k = 1}}^{K} PR^{sym}(x^i, Y^i_k, w_k) + PR^{me}(x^i, Y^i, w)\right). \tag{11}$$

It should be noted that our method is not limited to a multi-label case. Likewise it can be modified to exploit consistency on learning hierarchical categories.

4 Experiments

In this section, we first give details about the dataset and our implementation and then we empirically evaluate our method. We analyse the contribution of each added component and then compare our results to the state-of-the-art for weakly supervised detection and classification.

4.1 Implementation Details

We evaluate our method in the Pascal VOC 2007 dataset [9] which allows us to compare our results to previous work [5, 20, 24]. We follow the standard VOC procedure [9] and report average precision (AP) on the Pascal VOC 2007 *test* split. Note that some previous works [8, 21] have shown results only on VOC 2007 *trainval* split. Differently from Cinbis

	LSVM	SLSVM	+flip	+PR^{sym}	+PR^{me}
aeroplane	37.8	39.8	39.0	40.4	**42.2**
bicycle	40.1	42.0	42.3	**43.9**	**43.9**
bird	**24.9**	22.7	23.8	22.7	23.1
boat	8.1	9.1	9.1	**9.7**	9.2
bottle	12.1	12.9	**13.0**	12.6	12.5
bus	40.7	42.1	42.5	44.7	**44.9**
car	43.0	42.1	42.1	45.0	**45.1**
cat	**28.2**	23.4	26.4	25.9	24.9
chair	0.8	**9.3**	8.0	7.0	8.3
cow	8.1	21.3	23.7	**24.7**	24.0
diningtable	11.2	8.0	5.9	8.9	**13.9**
dog	**21.3**	17.9	17.4	19.7	18.6
horse	28.7	27.6	28.3	30.0	**31.6**
motorbike	**44.4**	41.9	42.8	43.9	43.6
person	11.2	10.8	8.5	**12.5**	7.6
pottedplant	13.1	18.8	20.2	20.6	**20.9**
sheep	15.7	19.9	26.2	**26.6**	**26.6**
sofa	18.6	18.4	18.8	19.6	**20.6**
train	34.8	33.5	33.7	35.8	**35.9**
tvmonitor	10.8	18.2	25.1	24.8	**29.6**
mean	22.7	24.0	24.8	26.0	**26.4**

Table 1: Weakly supervised detection results on the Pascal VOC 2007. LSVM and SLSVM denote max-margin in Eq. 4 and soft-max margin formulations in Eq. 5 resp. +flip indicates of adding horizontally mirrored training images to the learning. PR^{sym} and PR^{me} denote the posterior regularization for symmetry and mutual exclusion. The components starting from +flip are consecutively added on the SLSVM.

et al. [5] that use the extra "truncated" flag to exclude the images with only truncated object instances, we train with all images but the ones marked "difficult".

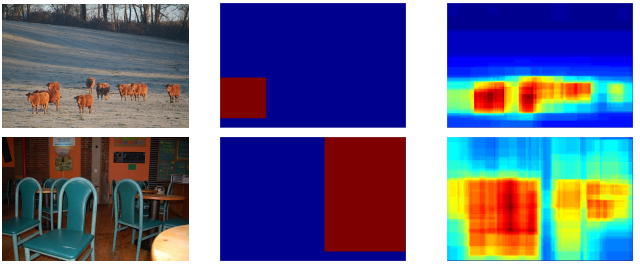
We use recently proposed convolutional neural network (CNN) features from DeCAF [6] that are pre-trained on the ImageNet ILSVRC 2012. We use the selective search windows ($\approx$ 1500 windows per image) from [25] to generate the candidate object locations. Similarly to [13], we crop images using these windows and run the pre-trained CNN on these cropped images to extract features. The initialization of the model is done setting the initial model parameter w to $\vec{0}$ which corresponds to giving the same probability to be an object to every bounding box.

4.2 Results

As the first baseline, we use our own implementation of [26] in margin rescaling configuration with hard negative mining (as in [10]). We denote the max-margin latent SVM with LSVM in Table 1. As shown in Table 1, LSVM obtains results comparable with the current state-of-the-art on weakly-supervised detection based on CNN [24] or Fisher Vectors [5].

As explained in Eq. 5, we smooth the LSVM learning by replacing the hard max with a soft one and denote this setting with SLSVM in Table 1. The SLSVM formulation obtains 24.0% mAP and improves the LSVM baseline with 1.3 points. Significant improvement can be seen in the categories with multiple instances such as "chair", "cow", "sheep" and "tv-monitor". We also illustrate max and soft-max outputs for representative "cow" and "chair" images in Fig. 1 and observe that soft-max provides a better representation for images with multiple instances.

In the next step, we add the flipped versions of VOC 2007 *trainval* images. As expected,



(a) images (b) 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. Best viewed in colour.

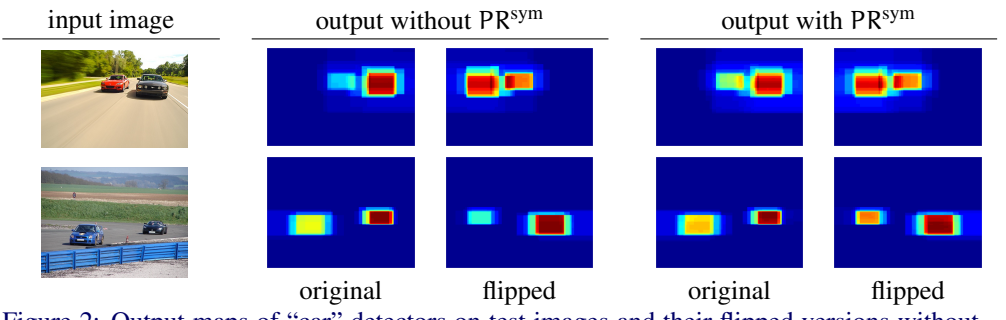


Figure 2: Output maps of "car" detectors on test images and their flipped versions without and with posterior regularization for symmetry. Learning with the symmetrical constraints increase the scores of less confident detections. Best viewed in colour.

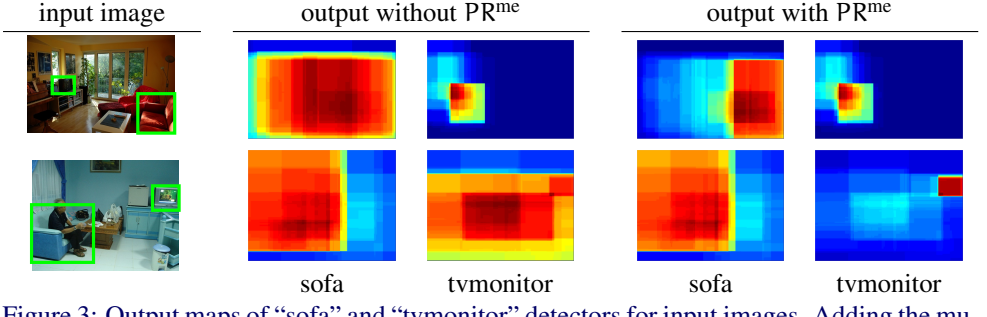


Figure 3: Output maps of "sofa" and "tvmonitor" detectors for input images. Adding the mutual exclusion regularization helps to separate two distributions by penalizing the bounding boxes with high probability for both detectors. Best viewed in colour.

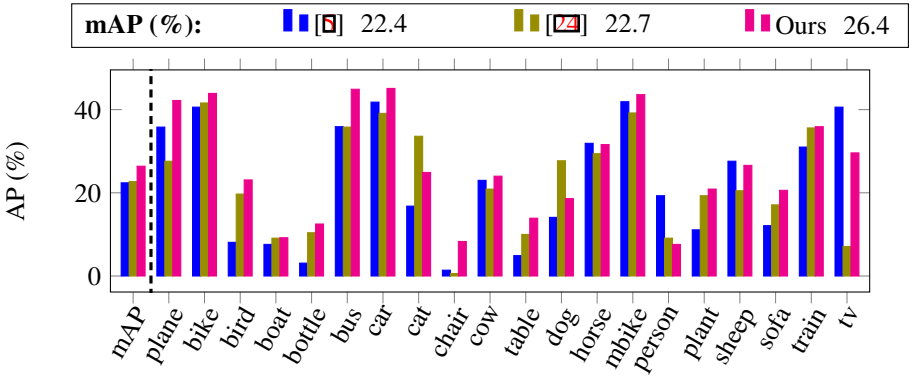


Figure 4: A comparison of our full model to the state-of-the-art weakly supervised detection methods on the Pascal VOC 2007 test in terms of AP. Our method significantly outperforms two methods.

adding the flipped images to the training set helps the soft-max margin learning, obtaining 24.8% mAP and 0.8 point improvement on average (indicated as +flip in Table 1). Applying the regularization on symmetry further improves the mean result by 1.2% (indicated as PR^{sym} in Table 1). We illustrate the basic idea of symmetry regularization in Fig. 2 and show that symmetry improves localization when an image or its flipped version have low confidence on some object instances. We see that low scoring objects become more visible in the output images of Fig. 2.

We also compare the proposed PR^{sym} to an additional baseline that concatenates a cropped image and its mirrored version to represent a window during training and testing. The feature dimensionality of this baseline and thus its inference running time are double of our method. This baseline obtains a mAP of 24.9% (not shown in Table 1), *i.e.* a negligible improvement over adding flipped images to the training set. This indicates that the symmetry regularization leads to a better learning and faster inference time than the baseline.

We finally show that adding the second regularization term PR^{me} leads to a further improvement of 0.4 point and to a mAP of 26.4%. It obtains significant further improvements for "diningtable", "sofa" and "tvmonitor" categories. These categories often co-occur, and therefore serve as context information for one another. This complicates accurate localization. By adding the mutual exclusion term, the detectors are forced to differentiate from one another and learn what is specific to each individual category. We illustrate the effect of the mutual exclusion constraint on a representative "sofa"-"tvmonitor" pair in Fig. 3. Adding the mutual exclusion forces the sofa and tvmonitor detectors to have different probability maps.

Comparison to state-of-the-art weakly supervised detection: In this part, we compare our full configuration (*i.e.* the right-most column in Table 1) to two recent works [5, 24] that report the best results on this dataset. Those methods also use the window proposals from [25]. While [5] represents these windows with the high-dimensional Fisher Vectors [23], [24] relies on the CNN features that are also used in this paper. We show in Fig. 4 that our full configuration significantly outperforms both baselines.

Comparison to state-of-the-art classification: Since we optimize our weakly supervised detection based on the image-level labels, the same trained models for detection can also be used for image classification. Differently from detection, the score for a given image

Ours			Others	
SVM	LSVM	Full	[6]	[20]
74.1	77.1	**80.9**	65.6	77.7

Table 2: Classification results on the Pascal VOC 2007 in mean AP. SVM denotes training linear SVMs without any localization. LSVM and Full denote the formulation in Eq. 4 and our full model in Eq. 11. Our method outperforms the state-of-the-art classifiers [6, 20].

is the log sum of all windows scores since we use a soft-max margin learning. We show the classification results on the Pascal VOC 2007 in Table 2.

We compare our results to our baselines and previous work [6, 20]. For all of our baselines, we use the DeCAF features [6]. We do not use the better performing CNN features from [13] that are fine-tuned on the ground-truth bounding boxes of VOC 2007. As a first baseline, we train linear SVM on entire images without any localization and denote it as "SVM". This obtains 74.1% mAP. As our second baseline, we learn localization with the Latent Structural SVM [26] (indicated as "LSVM"). This shows that localization improves classification with 3 points by eliminating confusing background. Finally we run our full configuration and yield 80.9% mAP and a 6.5% improvement over the SVM setting.

We also compare our results to two recent works [6, 20] in Table 2 and outperform both the methods. Interestingly, the work by Oquab *et al.* [20] exploits a different kind of localization by generating mid-level image representations and requires the *ground truth bounding box annotations* to choose better image patches during training. A recent work [4] that rigorously analyses architecture details of different CNN techniques achieves 82.4% mAP in the Pascal VOC 2007. This approach is complementary to our method and we plan to integrate it to our work as future plan.

5 Conclusion

In this paper we propose a set of ideas that addresses and improves the limitations of weakly supervised learning for object detection. We first show that relaxing the commonly used latent max-margin classifier with a soft-max enables a smoother optimization and better representation of images during learning. Then we show that further improvements can be obtained by introducing domain-specific knowledge that softly enforces symmetry and mutual exclusion on the posterior latent distribution at learning. A joint learning with the proposed components contributes to obtain state-of-the-art results in both classification and weakly supervised detection. It is interesting to see that our weakly supervised results for object detection come close to what used to be the state-of-the-art under a fully supervised setting just a couple of years ago [10].

Acknowledgements: The authors acknowledge the support of the FP7 ERC Starting Grant COGNIMUND and the EU FP7 project AXES.

References

[1] B. Alexe, T. Deselaers, and V. Ferrari. What is an object? In *CVPR*, pages 73–80, 2010.

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

[3] H. Bilen, M. Pedersoli, V.P. Namboodiri, T. Tuytelaars, and L. Van Gool. Object classification with adaptable regions. *CVPR*, 2014.

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

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

[6] Jeff D., Y. Jia, O. Vinyals, J. Hoffman, N. Zhang, E. Tzeng, and T. Darrell. Decaf: A deep convolutional activation feature for generic visual recognition. *arXiv preprint arXiv:1310.1531*, 2013.

[7] Chaitanya Desai, Deva Ramanan, and Charless C Fowlkes. Discriminative models for multi-class object layout. *International journal of computer vision*, 95(1):1–12, 2011.

[8] T. Deselaers, B. Alexe, and V. Ferrari. Localizing objects while learning their appearance. In *ECCV*, pages 452–466. Springer, 2010.

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

[10] P.F. Felzenszwalb, R.B. Girshick, D.A. McAllester, and D. Ramanan. Object detection with discriminatively trained part-based models. *TPAMI*, 32(9):1627–1645, 2010.

[11] 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.

[12] K. Gimpel and N. A. Smith. Softmax-margin training for structured log-linear models. Technical Report CMU-LTI-10-008, Carnegie Mellon University, 2010.

[13] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. *arXiv preprint arXiv:1311.2524*, 2013.

[14] A. Joulin and F. Bach. A convex relaxation for weakly supervised classifiers. *arXiv preprint arXiv:1206.6413*, 2012.

[15] M.P. Kumar, B. Packer, and D. Koller. Self-paced learning for latent variable models. In *NIPS*, pages 1189–1197, 2010.

[16] D.C. Liu and J. Nocedal. On the limited memory bfgs method for large scale optimization. *Mathematical programming*, 45(1-3):503–528, 1989.

[17] K. Miller, M P. Kumar, B. Packer, D. Goodman, and D. Koller. Max-margin min-entropy models. In *AISTATS*, 2012.

[18] Y. Nesterov. Smooth minimization of non-smooth functions. *Mathematical programming*, 103(1):127–152, 2005.

[19] 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.

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

[21] M. Pandey and S. Lazebnik. Scene recognition and weakly supervised object localization with deformable part-based models. In *ICCV*, pages 1307–1314, 2011.

[22] F. Perronnin, J. Sánchez, and T. Mensink. Improving the fisher kernel for large-scale image classification. In *ECCV*, pages 143–156, 2010.

[23] P. Pletscher, C.S. Ong, and J.M. Buhmann. Entropy and margin maximization for structured output learning. In *Machine Learning and Knowledge Discovery in Databases*, pages 83–98. Springer, 2010.

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

[25] JRR Uijlings, KEA van de Sande, T Gevers, and AWM Smeulders. Selective search for object recognition. *International journal of computer vision*, 104(2):154–171, 2013.

[26] C. John Yu and T. Joachims. Learning structural svms with latent variables. In *ICML*, pages 1169–1176, 2009.

[27] A.L. Yuille and A. Rangarajan. The concave-convex procedure. *Neural Computation*, 15(4):915–936, 2003.