WEAKLY SUPERVISED OBJECT DETECTION WITH CORRELATION AND PART SUPPRESSION

Fang Wan[1*], *Pengxu Wei*[1*], *Zhenjun Han*[1+], *Kun Fu*[2], *Qixiang Ye*[1+]

[1]School of Electronics, Electrical and Communication Engineering
University of Chinese Academy of Sciences, Beijing, China.
[2] Institute of Electronics, Chinese Academy of Sciences.
{hanzhj,qxye}@ucas.ac.cn

ABSTRACT

In weakly supervised object detection, conventional methods treat object location in each image as a latent variable and use non-convex optimization to solve the latent variable. However, as the optimization objective is image-level instead of sample-level, the learning procedure tends to choose object parts as false positive samples. Furthermore, when multiple classes of objects appear in the same images, the models could invite class-correlations and lose discriminative capability. In this paper, we propose a simple but effective suppression strategy that mines hard negative samples in the learning procedure to ease the above problems. We propose using a spatial-voting strategy to help finding negative samples to suppress the impact of object parts. We also use regions from class-correlated images as negative samples to suppress the impact of class-correlations. Experiments show that our approach significantly improves the baseline by 6% and achieves state-of-the-art performance.

Index Terms— Weakly Supervised Object Detection, Correlation and Part Suppression, SLSVM

1. INTRODUCTION

Object detection is one of the most fundamental computer vision task. Nevertheless, most existing object detection methods require a considerable amount of manual annotations of samples (classes and locations) at learning stage [1, 2, 3]. Considering that the image-level annotations are widely available on the Internet, weakly supervised object detection (WSOD), which uses image-level annotations about the presence or absence of a class of objects, has attracted increasing attentions.

Multi-instance learning (MIL) [4, 5, 6], latent SVM (LSVM) [7, 8, 9, 10, 11] and clustering [9, 12, 13] are three kinds of widely used WSOD methods. MIL treats each training image as a "bag" of samples and iteratively selects high-scored samples from each bag to learn models. Recent proposed multi-fold MIL [5, 6] uses division of training set and cross validation to prevent training from prematurely locking onto erroneous object locations. Various clustering approaches target at finding a single compact cluster for the object class and multiple ones for the related backgrounds. Wang *et al*. [12, 13] calculate clusters of image windows with probabilistic latent Semantic Analysis (pLSA) on the regions of positive samples, as well as employing these clusters and a voting-like approach to determine positive subcategories. Bilen and Song [9, 11] use clustering to initialize their latent variables (*i.e*. object regions, part configurations and sub-categories) and learn object detectors based on this initialization.

Latent SVM has been the most popular WSOD method, which learns weakly supervised models with a non-convex optimization function. To prevent the optimization from getting stuck to local minimum, Bilen *et al*. [8] propose regularizing object symmetry and class mutual exclusion informa-

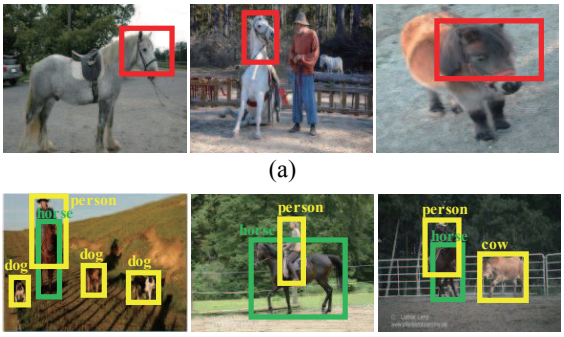


Fig. 1. Illustration of object part and class-correlation problems. (a) Object parts (red boxes) could follow a SLSVM model, and thus be falsely detected as positives. (b) The presence of object from other classes (yellow boxes) introduces correlation to the "horse" object class (green boxes). (Better view in color version)

*Contributed equally to this work.
+Corresponding authors.

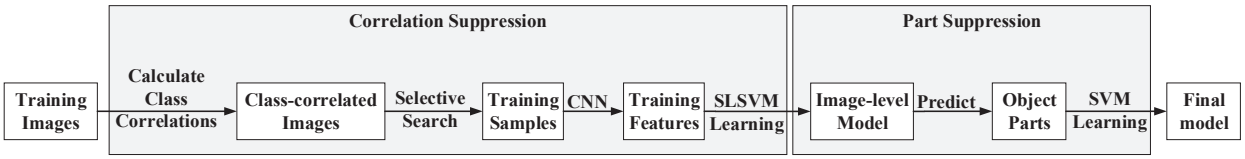


Fig. 2. Flowchart of our proposed approach.

tion in the objective function. Song $et\ al.$ [10] improve the optimization of latent SVM by applying Nesterov's smoothing technique. Bilen $et\ al.$ [9] improve the solution of latent SVM by regularizing a term for convex clustering.

The conventional LSVM based approaches [8, 9, 10, 11] aim to optimize the image-level classification instead of the sample-level classification. Once the image-level classification objective function reaches optimization, whether or not the sample-level classification is optimized, the learning procedure stops. Considering that all positive images contain the object parts but none of negative images does, LSVM is easy to falsely select object parts as "positive" samples.

In addition, a single training image might contain multiple classes of objects, thus the image has multiple image-level labels. When learning a detector for the target class of objects using images of multiple labels, the non-target objects could be falsely introduced into target class by the latent SVM, and bring correlations to the learned detector.

In this paper, we propose a simple but effective suppression strategy to ease the above problems. First, we suppress the object class correlation by statistically sampling the negative images with the correlation information and train a correlation-suppressed model. We then predict the object location by a spatial voting scheme and obtain the negative object parts for retraining a part-suppressed model.

The remainder of this paper is organized as follows. In Section 2, we review the Soft-max LSVM (SLSVM) approach, followed by our proposed learning approach with part and correlation suppression. Section 3 presents the experimental results and Section 4 concludes the paper.

2. APPROACH

The flowchart of our approach is shown in Fig. 2. It incorporates the proposed correlation and part suppression strategies into the SLSVM-based weakly supervised detection framework. In the learning stage, we separately train a detector for each object class. We use the Selective Search algorithm [14] to extract regions as [3, 15], which produces about 2000 samples for each image. Each region is represented with a 4096 dimensional feature vector from FC7 layer of the CNN pretrained on the ILSVRC image classification dataset [16]. In correlation suppression, we use all positive images and correlated negative samples for SLSVM training. In part suppression, we use object parts as negatives samples and train an SVM model.

2.1. Review of SLSVM

Let $x \in X, y \in Y, h \in H$ denote image, its binary label and the object location (bounding box) which is a latent variable to be solved, respectively. The joint feature vector [8] of image x, label y and location h is defined as follows:

$$\Phi(x, y, h) = \begin{cases} \phi(x, h) & if\ y = 1 \\ \vec{0} & if\ y = -1, \end{cases} \tag{1}$$

where ϕ denotes the feature representation. For image x, its label y and the region h of the target object can be predicted via a detection model w by maximizing the flowing function

$$\{y^*,\ h^*\} = \underset{y \in Y,\ h \in H}{\arg\max}\ w \cdot \Phi(x, y, h)\,. \tag{2}$$

Concretely, the objective function for learning the model w on the training image set $S = \left\{\left(x^i, y^i\right), i = 1, \ldots, N\right\}$ is defined as:

$$L(w, S) = \frac{1}{2} \parallel w\parallel^2 + \lambda \sum_{i=1}^{N} l_m\left(w, x^i, y^i\right), \tag{3}$$

where the regularization term of the image classification loss l_m is defined as

$$\begin{aligned} l_m(w, x^i, y^i) = &\log \sum_{y,h} \exp(w \cdot \Phi(x^i, y, h) + \Delta(y^i, y)) \\ &- \log \sum_{h} \exp(w \cdot \Phi(x^i, y^i, h)), \end{aligned} \tag{4}$$

where $\Delta(y^i, y) = 0$ if $y = y^i$, 1 else.

In learning procedure, a concave-convex optimization for max-margin latent SVM is used to solve the non-convex problem. More details can be found in [7, 8].

2.2. Correlation Suppression

The correlation problem is caused by multiple labels of images. Conventional WSOD methods that simply use all (or randomly sampled subset) of the negative images in the learning stage do not consider the class-correlation problem. In this paper, we experimentally find that class-correlations do effect the WSOD performance and the object distributions in the positive images are suitable to describe such correlations. Considering that the correlation between two object classes is not symmetric, $e.g.$, half of the "horse" images contain the object "person" while only one in ten of the "person" images contain the object "horse", we have to calculate the correlation vector for each class of objects individually. The pipeline of correlation suppression is described in Algorithm 1. For

Algorithm 1 Correlation Suppression

Input: Training image set X of C object classes
Output: Corrlation-suppressed model $\{w_c, c{=}1, 2, \ldots, C\}$

for each object class c **do**
  $X_c^+, X_c^- \leftarrow X$ // Class-specifically divided.
  $L_c^+ \leftarrow X_c^+$ // L_c^+ is the positive image labels set
  $\rho_c \leftarrow L_c^+$ // ρ_c is the class correlation for class c
  $X_{cT}^+ \leftarrow X_c^+$ // X_{cT} is the training image set
  $X_{cT}^- \leftarrow \emptyset$
  for each object class k except c **do**
    $X_{k\rho} \leftarrow \rho, X_c^-$ // Negative sampling for class k
    $X_{cT}^- \leftarrow X_{cT}^- \cup X_{k\rho}$
  end for
  $w_c \leftarrow Training\ SLSVM\ using\ X_{cT}^+, X_{cT}^-$
end for

the target object class c, we divide the training image set X into positive set X_c^+ and negative set X_c^-. To calculate how many correlated (negative for class c) objects appear in the positive images, we get the label set L_c^+ by counting all the image labels except label c in positive set X_c^+. We then calculate the correlation vector ρ_c, which indicates the relation between positive objects and correlated objects, by

$$\rho_c = hist\left(L_c^+\right) / \left|L_c^+\right|, \tag{5}$$

where $hist\left(\cdot\right)$ denotes the calculation of histograms, and the function $|\cdot|$ returns the total number of elements in a set. We then sample the negative samples for each correlated object class k except c (because images with label c are positive) by randomly sample $\rho_c \left|X_c^-\right|$ images of class k in X_c^-.

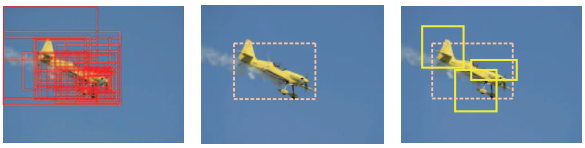
After obtaining the positive and negative image set X_c^+ and X_c^-, the correlation-suppressed SLSVM model for object c can be trained. We apply such a procedure to all classes of objects to obtain the correlation-suppressed SLSVM models.

2.3. Part Suppression

We also propose suppressing the impact of object parts by sampling object parts as negative training samples. In WSOD settings, however, there are no precise locations for positive training samples and the top-scored samples obtained by the SLSVM model may be incorrect (as shown in Fig. 1(a)), which makes it difficult to directly extract the object parts. We propose using a spatial "voting" scheme to predict the positive object location and then localize object parts as negatives (shown in Fig. 3). We retrain a part-suppressed detection model using the previous obtained samples.

We use the SLSVM model trained in section 2.2 as an initial model, and calculate the detection score for each sample. We then choose a set samples H_S by using the SLSVM scores, as:

$$H_S = \left\{score(h) > \theta |\ h \in H\right\}, \tag{6}$$



(a) (b) (c)

Fig. 3. (a) The high-scored boxes; (b) The predicted positive sample (yellow dotted box); (c) Selection of object parts as negative samples (yellow solid boxes). (Better view in color version)

where θ is the score threshold (as shown in Fig. 3(a)). Given N_{H_S} samples $\left\{h_i = (p_1^i, q_1^i, p_2^i, q_2^i), i = 1, ..., N_{H_S}\right\}$, where (p_1^i, q_1^i) and (p_2^i, q_2^i) denote the left-top and right-bottom coordinates of sample h_i, the predicted object location h^* (shown in Fig. 3(b)) is

$$h^* = (\overline{p_1}, \overline{q_1}, \overline{p_2}, \overline{q_2}), \tag{7}$$

where $\overline{p_1} = 1/N_{H_S} \sum_{i \in [1, N_{H_S}]} p_1^i$ is a coordinate value by spatial voting and $\overline{q_1}$, $\overline{p_2}$ and $\overline{q_2}$ are similar to it. Then, the negative object parts can be obtained by the predicted object location h^* as

$$\left\{h\,|0 < overlap\,(h, h^*) \leq \delta, h \in H\right\}, \tag{8}$$

where $overlap(\cdot)$ denotes the intersection-over-union overlap between two boxes (samples) and δ denotes the overlap threshold.

To learn a part-suppressed detector, we choose the fully supervised SVM model. In the SVM learning procedure, predicted object regions (with Eq. 7) from all positive images are chosen as positive training samples. The predicted object part samples are merged with the samples from negative images as negative training samples.

3. EXPERIMENT

We evaluate our approach on the challenging PASCAL VOC 2007 dataset [17], which totally contains 20 object classes and 9963 images. The dataset is are divided into three subsets: $train$, val and $test$ with 2501, 2510 and 4952 images, respectively. Following the previous works of WSOD, we use the $trainval$ set for training and $test$ set for testing. Each object class has an image-level label where 1 means that there is at least one target objects in the image, -1 means none and 0 means the "difficult" image for recognition. We exclude the images with label 0 for each object class in the training process as [8].

In the PASCAL VOC dataset, there are about 350 positive images and 4650 negative images for each object class. In experiments, we propose using all positive images for each object class and class-correlated negative images obtained by the correlation suppression strategy. Using class-correlated negatives not only benefits the learned detection models but also reduces the memory cost during the learning stage. In

Table 1. The detection results of the previous works and our proposed method on PASCAL VOC 2007

Method	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	hors	mbik	pers	plnt	shp	sofa	train	tv	mAP
Song *et al*. [10]	27.6	41.9	19.7	9.1	10.4	35.8	39.1	33.6	0.6	20.9	10.0	**27.7**	29.4	39.2	9.1	19.3	20.5	17.1	35.6	7.1	22.7
Song *et al*. [11]	36.3	47.6	23.3	12.3	11.1	36.0	46.6	25.4	0.7	23.5	12.5	23.5	27.9	40.9	14.8	19.2	24.2	17.1	37.7	11.6	24.6
Bilen *et al*. [8]	42.2	43.9	23.1	9.2	12.5	**44.9**	45.1	24.9	8.3	24.0	**13.9**	18.6	31.6	43.6	7.6	**20.9**	26.6	20.6	35.9	29.6	26.4
Bilen *et al*. [9]	**46.2**	46.9	24.1	**16.4**	12.2	42.2	47.1	35.2	7.8	28.3	12.7	21.5	30.1	42.4	7.8	20.0	**26.8**	**20.8**	35.8	**29.6**	27.7
SLSVM [8]	39.8	42.0	22.7	9.1	**12.9**	42.1	42.1	23.4	**9.3**	21.3	8.0	17.9	27.6	41.9	10.8	18.8	19.9	18.4	33.5	18.2	24.0
Our slsvm	36.1	36.4	18.1	10.5	10.0	35.1	43.3	31.8	2.7	21.7	10.6	19.1	30.9	34.3	14.9	14.2	17.5	8.3	30.1	13.9	22.0
Our slsvm+C	35.2	48.2	20.2	14.0	2.5	41.2	43.9	26.8	5.3	**29.3**	9.5	21.9	35.3	38.1	11.9	17.3	25.4	14.5	37.6	20.5	24.4
Our slsvm+C+P	44.9	**52.2**	**24.5**	14.4	11.2	40.6	**52.2**	**35.2**	3.4	28.9	3.4	25.9	**39.4**	**44.4**	**24.5**	17.2	19.1	18.2	**40.7**	24.9	**28.3**

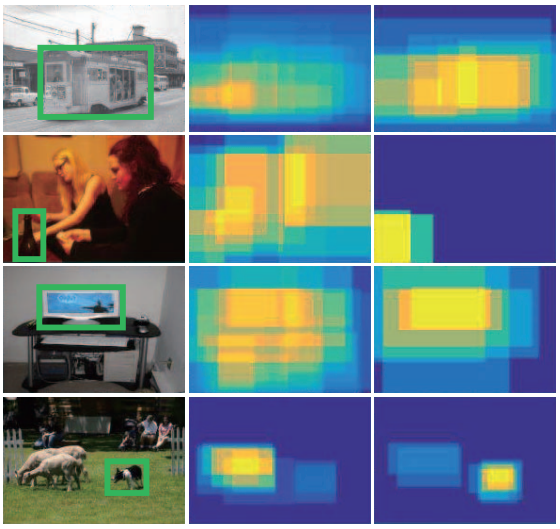


Fig. 4. Illustration of the effectiveness of correlation suppression. From left to right: object locations, score heat-maps of SLSVM, score heat-maps after correlation suppression. (Better view in color version)

the part suppression procedure, we experimentally choose an overlap threshold δ defined in Eq. 8. Experiments show that with $\delta = 0.3$ we can obtain the best performance.

Table 1 shows the mean average precisions (mAPs) of the proposed approach and other state-of-the-art approaches. In Table 1, "SLSVM" denotes the SLSVM baseline in [8] and "Our slsvm", "Our slsvm+C" and "Our slscm+C+P" denote the SLSVM baseline, SLSVM with correlation suppression, and SLSVM with both correlation and part suppression, respectively. It shows our implemented baseline for SLSVM model "Our slsvm" gets 22% mAP, which is 2% lower than "SLSVM" [8]. It can be seen in Table 1 that "Our slsvm+C" improves the baseline by 2.4% while "Our slsvm+C+P" further improves 3.9%. Our approach finally yields a detection mAP of 28.3%, which significantly improves the baseline result by 6.3% and slightly improves the state-of-the-art approach [9] result by 0.6%.

Fig. 4 and Fig. 5 show the detection results. It can be seen in Fig. 4 that by using the correlation suppression strategy, the class-correlated objects ("car", "person", "table" and "sheep") are suppressed and the target objects ("bus", "bottle", "TV" and "dog") are more precisely detected. Fig. 5 shows that the proposed part suppression strategy reduces the response of object parts in the learning stage, which consequently improves the detection models and precision.

Fig. 5. Illustration of the effectiveness of part suppression. Red boxes are detection results by the SLSVM. Green boxes are results of our approach with part suppression. (Better view in color version)

4. CONCLUSION

In this paper, we revisit the challenging weakly supervised object detection problem. Most of the conventional methods using a non-convex optimization approach to learn models that achieve best image-level classification could be impacted by object parts and class correlations. Adopting the soft-max latent SVM as the learning method, we propose a part and correlation suppression strategy which uses negative samples to alleviate the impact of object part and class correlation. Experiments show that our proposed strategy is effective, indicating that object part and class-correlation are two important factors in weakly supervised object detection. To facilitate future development, we have made the source code publicly available at $www.ucassdl.cn$.

5. ACKNOWLEDGMENT

This work is supported in part by the National Science Foundation of China under Grants 61271433 and 61202323, and Beijing Municipal Science & Technology Commission.

6. REFERENCES

[1] Navneet Dalal and Bill Triggs, "Histograms of oriented gradients for human detection," in *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*. IEEE, 2005, vol. 1, pp. 886–893.

[2] Pedro F Felzenszwalb, Ross B Girshick, David McAllester, and Deva Ramanan, "Object detection with discriminatively trained part-based models," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 32, no. 9, pp. 1627–1645, 2010.

[3] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jagannath Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*. IEEE, 2014, pp. 580–587.

[4] Stuart Andrews, Ioannis Tsochantaridis, and Thomas Hofmann, "Support vector machines for multiple-instance learning," in *Advances in neural information processing systems*, 2002, pp. 561–568.

[5] Ramazan Gokberk Cinbis, Jakob Verbeek, and Cordelia Schmid, "Multi-fold mil training for weakly supervised object localization," in *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*. IEEE, 2014, pp. 2409–2416.

[6] Ramazan Gokberk Cinbis, Jakob Verbeek, and Cordelia Schmid, "Weakly supervised object localization with multi-fold multiple instance learning," *arXiv preprint arXiv:1503.00949*, 2015.

[7] Chun-Nam John Yu and Thorsten Joachims, "Learning structural svms with latent variables," in *Proceedings of the 26th Annual International Conference on Machine Learning*. ACM, 2009, pp. 1169–1176.

[8] Hakan Bilen, Marco Pedersoli, and Tinne Tuytelaars, "Weakly supervised object detection with posterior regularization," in *British Machine Vision Conference*, 2014.

[9] Hakan Bilen, Marco Pedersoli, and Tinne Tuytelaars, "Weakly supervised object detection with convex clustering," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 1081–1089.

[10] Hyun Oh Song, Ross Girshick, Stefanie Jegelka, Julien Mairal, Zaid Harchaoui, and Trevor Darrell, "On learning to localize objects with minimal supervision," *arXiv preprint arXiv:1403.1024*, 2014.

[11] Hyun Oh Song, Yong Jae Lee, Stefanie Jegelka, and Trevor Darrell, "Weakly-supervised discovery of visual pattern configurations," in *Advances in Neural Information Processing Systems*, 2014, pp. 1637–1645.

[12] Chong Wang, Weiqiang Ren, Kaiqi Huang, and Tieniu Tan, "Weakly supervised object localization with latent category learning," in *Computer Vision–ECCV 2014*, pp. 431–445. Springer, 2014.

[13] Chong Wang, Kaiqi Huang, Weiqiang Ren, Junge Zhang, and Steve Maybank, "Large-scale weakly supervised object localization via latent category learning," *Image Processing, IEEE Transactions on*, vol. 24, no. 4, pp. 1371–1385, 2015.

[14] Jasper RR Uijlings, Koen EA van de Sande, Theo Gevers, and Arnold WM Smeulders, "Selective search for object recognition," *International journal of computer vision*, vol. 104, no. 2, pp. 154–171, 2013.

[15] Thomas Deselaers, Bogdan Alexe, and Vittorio Ferrari, "Weakly supervised localization and learning with generic knowledge," *International journal of computer vision*, vol. 100, no. 3, pp. 275–293, 2012.

[16] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[17] Mark Everingham, Luc Van Gool, Christopher K-I Williams, John Winn, and Andrew Zisserman, "The pascal visual object classes (voc) challenge," *International journal of computer vision*, vol. 88, no. 2, pp. 303–338, 2010.