Cross-Validated Smooth Multi-Instance Learning

Dayuan Li[1], Lin Zhu[1], and Wenzheng Bao[1]
[1]Institute of Machine Learning and Systems Biology
College of Electronics and Information Engineering
Tongji University
Caoan Road 4800, Shanghai 201804, China
1531811@tongji.edu.cn, dshuang@tongji.edu.cn,
lizhonyx@163.com, baowz55555@126.com

Fei Cheng[2], Yi Ren[2] and De-Shuang Huang[1*]
[2]Beijing E-Hualu Info Technology Co., Ltd
Beijing, China
cf@ehualu.com, reny@ehualu.com

***Abstract*—The problem of object localization in image appear ubiquitously in computer vision applications including image classification, object detection and visual tracking. Recently, it is shown that multiple-instance learning(MIL) which is regarded as the fourth machine learning framework compared with supervised learning, unsupervised learning and reinforce learning has been verified that will get good effect in object localization in images. In this paper, we propose a novel method to solve the classical MIL problem, named Cross-Validated Smooth Multi-Instance learning (CVS-MIL). We treat the positiveness of instance as a continuous variable. The softmax model is used to bring a bridge between instances and bags and jointly optimize the bag label and instance label in a unified framework. The extensive experiments demonstrate that CVS-MIL consistently achieves superior performance on various MIL benchmarks. Moreover, we simply applied CVS-MIL to a challenging vision task, common object discovery. The state-of-the-art results of object discovery on Pascal VOC datasets further confirm the advantages of the proposed method.**

Keywords- Multiple-instance learning; Object localization; Smooth; Cross-Validated

I. INTRODUCTION

Object localization in image has been researched for decade years not only in computer vision but also in pattern recognition[4, 5]. The aim of these problem is to identify all instances of a given object category in an image, and find the location of the objects. It is much more challenging to localize the object compared with image classification. Since the multiple-instance learning (MIL) problem was introduced by Dietterich et al in 1997[6] for the task of drug activity prediction, a large number of researches try to solve object localization problem with MIL techniques emerged in recent years[7, 8]. In contrast to traditional supervised learning, only the labels of bags are given but the labels of instances belonged to one bag are unknown. The learner receives a set of labeled bags, each containing plenty of instances. A bag may be labeled as positive when at least one instance is positive and a bag will labeled as negative if none of the instances is positive in the binary classification task.

Figure 1. The location of objects discovered by CVS-MIL model with different iterations. The first row are original pictures. The second row are pictures divided by SLIC. The third to sixth rows show the detected object locations in iteration 200, 500, 1000, and 2000, respectively. The red boxes show the detected objects that do not enough overlap with ground-truth, and the green boxes show the detected objects that own enough overlap with ground-truth.

There is not a standard for what can be regarded as an instance, but typically we regard an image as a bag and some parts of the image as an instance. The Object that we are interested in considered as a positive instance, and the rest are considered as negative instances.

A discriminative instance model for classifying bags is

*The corresponding author of this paper.

very significant in MIL and experiments with the same classifier but different instance model will result in totally different result. A lot of work have been done to choose an instance model. Latent SVM which is also called multiple instance SVM was used to image categorization[9]. M Sapienza proposed a method to learn discriminative space–time action parts from weakly labelled videos[10]. R Hong annotated image by multiple-instance learning with discriminative feature mapping and selection[11]. All the methods mentioned above regard instance selection and model learning as two separated procedures but we proposed a unified framework to calculate and optimize the label of instance and bag by relaxing the discrete instance label and gradient descent.

In this paper, we utilize Edgebox[12] to propose 100 windows as instances and extract features of each instance with color, shape and texture. The probability of the instance label to be positive is represented by a Probit model. The relation between bag label and instance labels, which is also called MIL constraints, is the softmax function. The procedure of training our classifier follows the Multi-fold MIL training method to prevents training from prematurely locking onto erroneous object locations[13]. We present a detailed demonstration in Section 4. Our method is simple but effective. In the experiments, our approach leads to better localization on the training images using the PASCAL VOC 2007 dataset.

II. RELATED WORKS

Many multi-instance learning algorithms have been developed during the past decade. For example, DD, EM-DD, MIBoosting, miSVM and miGraph, etc. DD[14] finds regions in the instance space with instances from many different positive bags and few instances from negative bags to solve MIL. EM-DD[15] developed DD with expectation maximization (EM). MIBoosting[16] is a boosting approach for MIL assumed that all instances contribute equally and independently to a bag's label. miSVM[17] is an improved SVM designed for MIL. It searches max-margin hyperplanes to separate positive and negative instances in bags. miGraph[18] employs graph kernels to solve MIL problems which regards instances as a non-i.i.d.

Many researchers applied weakly supervised object discovery, especially MIL, to object localization in visual recognition, such as object class discovery[3], face recognition[8], object detection[19]. A multi-fold multiple instance learning procedure is designed to prevent training from prematurely locking onto erroneous object locations. Chunhui Gu's team[20] formed visual clusters from the data that are tight in appearance and configuration spaces and trained individual classifiers for each component, and then learn a second classifier that operates at the category level by aggregating responses from multiple components. Our method is different from the existing object discovery methods, which divides an image with super-pixels and extracts instance features with the color, shape and texture of that super-pixel. We also relaxes MIL problem and trains the CVS-MIL model in a multi-fold way to solve the problem.

III. Multi-fold MIL Training

Multi-fold MIL training[13] is proposed by Ramazan et al. which avoids this rapid convergence to poor local optima. We randomly divide the positive training images(bags) into K disjoint folds, and re-localize the images in each fold using a detector trained using windows from positive images in the other folds. Re-localization detectors never use training windows from the images to which they are applied. Once re-localization is performed in all positive training images, we train another detector using all selected windows. This detector is used for hard-negative mining on negative training images, and returned as the final detector. The traditional MIL training does not execute steps 1 and 2, and re-localizes based on the detector learned in 5.

Algorithm 1 — Multi-fold weakly supervised training

For iteration t = 1 to T

1. Divide positive images randomly into K folds.
2. For k = 1 to K
3. Train using positives in all folds but k.
4. Re-localize positives in fold k using this detector.
5. Train detector using positive windows from all folds
6. Perform hard-negative mining using this detector
7. Return final detector and object windows in train data

Where T is the times of iteration which is defined by the user according to experience. The training in step 3 refers multi-fold MIL training described in section 3.

IV. Smooth MIL

To demonstrate our approach in a more understandable way we first give some notations of it. Given is a set of input patterns $\mathbf{x}_1,\cdots,\mathbf{x}_n$ grouped into bags $\mathbf{B}_1,\cdots,\mathbf{B}_m$, with $\mathbf{B}_I = \{\mathbf{x}_i : i \in I\}$ for given index sets $I \subseteq \{1,\cdots,n\}$. Each bag $\mathbf{B}_I$ is consisted with a set of instance $\mathbf{x}_i = \{\mathbf{x}_{i1},\cdots,\mathbf{x}_{im_i}\}$, where m_i denotes the number of instances in the bag $\mathbf{X}_i$ and it is also associated a label Y_I which can be interpreted in the following way: if $Y_I = -1$, then $y_{ij} = -1$ for all $i,j \in \mathbf{I}$ where y_{ij} is the label of instance $\mathbf{x}_{ij}$, *i.e.*, no instance in the bag is a positive instance. On the other hand, if $Y_I = 1$, then at least one instance $\mathbf{x}_{ij} \in \mathbf{B}_I$ is a positive instance.

The labels of bags or instances in typically MIL problem is discrete and the values of them can only be 1 or 0. We smooth the instance label y_i to be a continues variable in the range of [0,1], which represents the probability of $\mathbf{x}_{ij}$ being positive, denoted as p_{ij}. $\mathbf{x}_{ij}$ is the j-th instance in the i-th bag.

p_{ij} is given by a Probit model.

$$p_{ij} = \Pr(y_{ij} = 1 \mid \mathbf{x}_{ij}; \mathbf{w}) = \Phi(\mathbf{w}^T \mathbf{x}_{ij}) \tag{1}$$

Where Pr denotes probability, and Φ is the cumulative distribution function of the standard normal distribution. $\mathbf{w}$ is the weight vector of the linear model which needs to be optimized through in our formulation. The goal of our problem is to predict whether a bag is positive, so only know the positive probability of instances is far from enough. The given label set is all about bags and we do not know the label of instance. We adopt the softmax model to bring a bridge between these two kinds of labels:

$$P_i = \frac{p_{ij}}{\sum_{j=1}^{m_i} p_{ij}} \tag{2}$$

where m_i is the number of instances in $\mathbf{B}_i$. Assuming that one instance in the bag is predicted as positive, then we can find $P_i = 1$ according to Eq. (2). P_i will equal to 0 if and only if all instances in the bag are predicted as zero.

After smoothing the typical MIL problem, the problem becomes more tractable. There is no discrete variable and we can differentiate all parts in Eq. (3). Our MIL objective function is given as follows with considering the instance-level loss, bag-level loss, and model regularization.

$$\min_{\mathbf{w}} \frac{\beta}{n} \sum_{i=1}^{n} \underbrace{-\{Y_i \log P_i + (1 - Y_i) \log(1 - P_i)\}}_{\text{cost item for i-th bag}} + \frac{1}{n} \sum_{i=1}^{n} \frac{1}{m_i} \sum_{j=1}^{mi} \underbrace{\max(0, [\mathrm{m}_0 - \mathrm{sgn}(p_{ij} - p)\mathbf{w}^T \mathbf{x}_{ij}])}_{\text{cost item for ij-th instance}} + \frac{\lambda}{2} \|\mathbf{w}\|^2 \tag{3}$$

where sgn is the sign function; m_0 is the margin parameter in SVM model used to separate the positive instances and negative instances distant from the hyper line in the feature space; p is a threshold defined before the training to determine positive or instance.

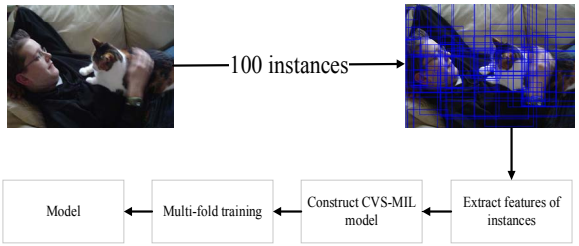

Figure 2. The pipeline of our training procedure. The first picture is the original picture, the second picture is the result of SLIC super-pixel division.

We can optimize Eq. (3) with gradient descent and the partial derivative of the cost item for bag with respect to $\mathbf{w}$ is derived as

$$\frac{\partial \mathcal{L}_{bagi}}{\partial \mathbf{w}} = \frac{\partial \mathcal{L}_{bagi}}{\partial P_i} \cdot \sum_{j=1}^{m_i} \frac{\partial P_i}{\partial p_{ij}} \frac{\partial p_{ij}}{\partial \mathbf{w}}, \tag{4}$$

where $\frac{\partial \mathcal{L}_{bagi}}{\partial P_i}$ and $\frac{\partial P_i}{\partial p_{ij}}$ is given by

$$\frac{\partial \mathcal{L}_{bagi}}{\partial P_i} = -\{\frac{Y_i}{P_i} - \frac{(1 - Y_i)}{1 - P_i}\} = -\frac{Y_i - P_i}{P_i(1 - P_i)}; \tag{5}$$

$$\frac{\partial P_i}{\partial p_{ij}} = \frac{1}{p_{ij} + \sum_{t \in [1, m_i] \setminus j} p_{it}} - \frac{1}{(p_{ij} + \sum_{t \in [1, m_i] \setminus j} p_{it})^2} \tag{6}$$

As for the final expression of partial derivative of $\mathcal{L}_{bagi}$ with respect to $\mathbf{w}$ is

$$\frac{\partial \mathcal{L}_{bagi}}{\partial \mathbf{w}} = -\frac{Y_i - P_i}{P_i(1 - P_i)} \cdot \sum_{j=1}^{m_i} \frac{1}{p_{ij} + \sum_{t \in [1, m_i] \setminus j} p_{it}} - \frac{1}{(p_{ij} + \sum_{t \in [1, m_i] \setminus j} p_{it})^2} \frac{\partial p_{ij}}{\partial \mathbf{w}} \tag{7}$$

And then we can find the partial derivative of $\mathcal{L}_{insij}$ with respect to $\mathbf{w}$ is

$$\frac{\partial \mathcal{L}_{insij}}{\partial \mathbf{w}} = -[\mathrm{sgn}(p_{ij} - p)\mathbf{w}^T \mathbf{x}_{ij} < m] . \mathrm{sgn}(p_{ij} - p)\mathbf{x}_{ij} \tag{8}$$

where $[\mathrm{sgn}(p_{ij} - p)\mathbf{w}^T \mathbf{x}_{ij} < m]$ can only be one or zero which mean its argument is true or false. We update the weight vector using a varied learning rate $s_t = \frac{1}{t\lambda}$, and then $\mathbf{w}_{t+1} = \mathbf{w}_t - s_t \cdot \frac{\beta}{n} \cdot \Delta d$, where Δd is the partial derivative of the objective function Eq. (3) with respect to $\mathbf{w}$.

In summary, the pipeline of our training procedure is shown in Fig.2.

V. Results

In this section, we compare the performance of the proposed method on MIL benchmarks and object discovery in the wild, respectively. Experiments are carried out on a desktop machine with Intel core i7-3770 CPU and 16GB RAM. Given a set of images, we utilize Edgebox[12] to capture plenty of windows as object proposals. This strategy turns the object discovery problem into a well-defined MIL problem in which an image is a bag, an object proposal is an instance, and image label is used as bag label. We extract shape, color and texture features of the instance as input of the CVS-MIL model we proposed before. Furthermore, we regard images containing a shared object as positive bag and randomly select images from the remaining images as negative. The model is trained by the multi-fold weakly supervised training mentioned in section III. At last, after the model is learnt, we report the object proposals with maximal value predicted by our model as the detected object. The final results evaluated via CorLoc measure, which is the percentage of the correct location of objects under the Pascal

Figure 4. Results of object discovery on several class on Pascal07 set. Each row denotes one class. These classes are, from top to bottom, cats, sheep, bicycles, people and cows. The solid green rectangle denotes the matched ground truth; the dashed green rectangle denotes the matched detection.

criteria (intersection over union (IoU) > 0.5 between detected bounding boxes and the ground truth).

A. Pascal

Pascal provides standard image data sets for object class recognition and the Pascal 2006 and 2007 datasets are extremely popular and challenging in computer vison. Two subsets are taken from Pascal 2006 and 2007 train + val dataset, which are divided into various of class and view combinations following the protocol of [21]. There are 2047 images divided into 45 class combinations in Pascal07 while total 2184 images from 33 class in Pascal06. We select all images which contain at least one object instance not marked as truncated or difficult in the ground truth.

Table 1. Object localization results evaluated via CorLoc on Pascal06 and Pascal07.

Dataset	Ours	Multi-fold	ADMM[1]	MIForests[2]	bMCL[3]
Pascal06	**53**	/	43	36	45
Pascal07	**38**	37	27	25	31

We utilize Edgebox to capture plenty of windows as object proposals and extract features by color, texture and SIFT. Our CVS-MIL model is trained with the multi-fold weakly supervised training algorithm with the parameter k equals 10. The parameters are given via $\lambda = 0.0017$, $\beta = 5$,

$p = 0.5$ and $m = 1.5$ on Pascal07 dataset, while $\lambda = 0.0017$, $\beta = 7$, $p = 0.5, T = 2000$ and $m = 1.48$ on Pascal06 dataset. The detectors and CorLoc are shown in Fig. 3 and Table 1, respectively.

B. Drug Activation Prediction

Some molecules with particular shapes can bind well to a target protein and these molecules are very meaningful in pharmacy. Our task is to predict whether a new molecule can bind well to the target protein. Normally, a molecule always exhibits multiple shapes and a good molecule will bind well if at least one of its shapes is right, while a poor molecule will not bind well if none of its shapes can bind. Therefore, our task can be regarded as a MIL problem.

Musk dataset is a popular benchmark for drug prediction. The goal is to learn to predict whether new molecules will be musks or non-musks. Both MUSK1 and MUSK2 are composed of molecules (bags) in multiple conformations (instances). MUSK1 describes a set of 92 molecules of which 47 are judged by human experts to be musks and the remaining 45 molecules are judged to be non-musks and MUSK2 describes a set of 102 molecules of which 39 are judged by human experts to be musks and the remaining 63 molecules are judged to be non-musks.

Table 2. Average prediction accuracy (%) via ten times 10-fold cross validation on MUSK datasets.

Dataset	Ours	miSVM	MISVM
MUSK1	**81.2**	78.0	80.4
MUSK2	**83.1**	70.2	77.5

We set $\lambda = 0.04$, $\beta = 1.8$, $p = 0.5, T = 2000$ and $m = 0.5$ in our model and our results are compared with miSVM and MISVM proposed in [17] in Table 2. All three methods adopt linear kernel function. We can find from the result that the accuracy of our method improved about 10% when comparing with miSVM in MUSK2 dataset. Furthermore on MUSK1 dataset, miSVM, MISVM and our method achieves a similar accuracy

VI. CONCLUSIONS

In this paper, we propose a novel method for MIL and applied it for object localization. In typical MIL problem, the labels of bag and instance are discrete but we smooth this problem into a convex problem and solve it efficiently. The CVS-MIL model is trained in a multi-fold way and the features of instance are extracted with color, shape and texture. Beside of object discovery, other recognition tasks, such as image classification, text categorization and automatic image annotation can also be solved by this method.

Acknowledgments. This work was supported by the grants of the National Science Foundation of China, Nos. 61572447，61472173, 61672203, 61472280, and 61373098, China Postdoctoral Science Foundation Grant, No. 2016M601646.

REFERENCES

[1] X. Wang, Z. Zhang, Y. Ma, X. Bai, W. Liu, and Z. Tu, "Robust subspace discovery via relaxed rank minimization," *Neural computation,* vol. 26, pp. 611-635, 2014.

[2] C. Leistner, A. Saffari, and H. Bischof, "MIForests: Multiple-instance learning with randomized trees," in *European Conference on Computer Vision*, 2010, pp. 29-42.

[3] J.-Y. Zhu, J. Wu, Y. Xu, E. Chang, and Z. Tu, "Unsupervised object class discovery via saliency-guided multiple class learning," *IEEE transactions on pattern analysis and machine intelligence,* vol. 37, pp. 862-875, 2015.

[4] X. Chen, A. Shrivastava, and A. Gupta, "Enriching visual knowledge bases via object discovery and segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 2027-2034.

[5] K. Tang, A. Joulin, L.-J. Li, and L. Fei-Fei, "Co-localization in real-world images," in *2014 IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 1464-1471.

[6] T. G. Dietterich, R. H. Lathrop, and T. Lozano-Pérez, "Solving the multiple instance problem with axis-parallel rectangles," *Artificial intelligence,* vol. 89, pp. 31-71, 1997.

[7] X.-S. Wei, J. Wu, and Z.-H. Zhou, "Scalable multi-instance learning," in *2014 IEEE International Conference on Data Mining*, 2014, pp. 1037-1042.

[8] T. L. Berg, A. C. Berg, J. Edwards, M. Maire, R. White, Y.-W. Teh, *et al.*, "Names and faces in the news," in *Computer Vision and Pattern Recognition, 2004. CVPR 2004. Proceedings of the 2004 IEEE Computer Society Conference on*, 2004, pp. II-848-II-854 Vol. 2.

[9] Y. Chen and J. Z. Wang, "Image categorization by learning and reasoning with regions," *Journal of Machine Learning Research,* vol. 5, pp. 913-939, 2004.

[10] M. Sapienza, F. Cuzzolin, and P. H. Torr, "Learning discriminative space–time action parts from weakly labelled videos," *International Journal of Computer Vision,* vol. 110, pp. 30-47, 2014.

[11] R. Hong, M. Wang, Y. Gao, D. Tao, X. Li, and X. Wu, "Image annotation by multiple-instance learning with discriminative feature mapping and selection," *IEEE transactions on cybernetics,* vol. 44, pp. 669-680, 2014.

[12] C. L. Zitnick and P. Dollár, "Edge boxes: Locating object proposals from edges," in *European Conference on Computer Vision*, 2014, pp. 391-405.

[13] R. G. Cinbis, J. Verbeek, and C. Schmid, "Multi-fold mil training for weakly supervised object localization," in *2014 IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 2409-2416.

[14] O. Maron and T. Lozano-Pérez, "A framework for multiple-instance learning," *Advances in neural information processing systems,* pp. 570-576, 1998.

[15] Q. Zhang and S. A. Goldman, "EM-DD: An improved multiple-instance learning technique," in *Advances in neural information processing systems*, 2001, pp. 1073-1080.

[16] X. Xu and E. Frank, "Logistic regression and boosting for labeled bags of instances," in *Pacific-Asia conference on knowledge discovery and data mining*, 2004, pp. 272-281.

[17] S. Andrews, I. Tsochantaridis, and T. Hofmann, "Support vector machines for multiple-instance learning," in *Advances in neural information processing systems*, 2002, pp. 561-568.

[18] Z.-H. Zhou, Y.-Y. Sun, and Y.-F. Li, "Multi-instance learning by treating instances as non-iid samples," in *Proceedings of the 26th annual international conference on machine learning*, 2009, pp. 1249-1256.

[19] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan, "Object detection with discriminatively trained part-based models," *IEEE transactions on pattern analysis and machine intelligence,* vol. 32, pp. 1627-1645, 2010.

[20] C. Gu, P. Arbeláez, Y. Lin, K. Yu, and J. Malik, "Multi-component models for object detection," in *European Conference on Computer Vision*, 2012, pp. 445-458.

[21] T. Deselaers, B. Alexe, and V. Ferrari, "Weakly supervised localization and learning with generic knowledge," *International journal of computer vision,* vol. 100, pp. 275-293, 2012.