Improvising Weakly Supervised Object Detection (WSOD) using Deep Learning Technique

Jyoti G. Wadmare, Sunita R. Patil

Abstract: ***Object detection is closely related with video and image analysis. Under computer vision technology, object detection model training with image-level labels only is challenging research area.Researchers have not yet discovered accurate model for Weakly Supervised Object Detection (WSOD). WSOD is used for detecting and localizing the objects under the supervision of image level annotations only.The proposed work usesself-paced approach which is applied on region proposal network of Faster R-CNN architecture which gives better solution from previous weakly-supervised object detectors and it can be applied for computer visionapplications in near future.***

Keywords: MIL, Object Detection, Weakly Supervised Learning, WSOD.

I. INTRODUCTION

Object detection is a computer vision technique that is used to identify and locate objects within an image or video. Specifically, object detection draws bounding boxes around detected objects, which allows locating said objects (or how they move through) in an input image or scene. It has many applications such as self-driving cars, event detection, remote sensing, medical image analysis, etc.

The state-of-the-art fully supervised object detector's training is totally based on the accurate bounding box annotations. However, the task of manual annotation is very tedious process. Due to this reason, enhancing detection algorithm is now in demand.

In this paper, we have presented a self paced approach to train the weakly supervised object detection (WSOD) model using image-level annotations. WSOD mitigates the problem of accurate bounding box annotation for each instance. This particular framework requires fewer efforts compared fully supervised object detection. However, image-level supervision for object detection remains a challenging task and continues to struggle with small objects, especially those bunched together with partial occlusions. Real-time detection classification and localization accuracy again remains challenging.Till date the methods available for object detection from images and video lack in correctly identifying type of object and localization accuracy.

The instance learning like Multiple Instance Learning (MIL) in which instead of labeling individual instances, it works with a set of instances called as bags and labeling is provided only for entire set. If all the instances are negative then the particular bag is labeled as negative and the positive bag contains at least one positive instance.

Revised Manuscript Received on January 21, 2020.

Ms. Jyoti G. Wadmare, Assistant Professor, Department of Computer Engineering, KJSIEIT, university of Mumbai, India.
E-mail: jyoti@somaiya.edu

Dr. Sunita R. Patil, Vice Principal, Professors at KJSIEIT, Mumbai, University of Mumbai (UoM), India.
E-mail: &vice_principal@somaiya.edu

If an image has a label of class X which contains bounding boxes and at least one bounding box is positive for the class X and other boxes are belong to other classes.

However the problem with MIL is if the initial classifier is not robust enough, then it can't predict the accurate target class. If the classifier predicted false positive for aninstance i.e.bounding box on the background which makes inaccurate prediction of target class. The crowded background causes ambiguity between the target object and background so that it affects on the localizationaccuracy.

To overcome the problem of cluttered background, decoupled attention-based object representation model is used which has important advantages: The attention map diminishes the impact from crowded backgrounds, and generates region based object representation more particular. The extracted attention maps gives idea about different image annotation [1].

Self-paced learning, used in research and it is similar to curriculum learning [2] and it is nothing but learning the simplest concepts first and gradually learns difficult concepts as per our human learning process. A self-paced learning approach is used for handling ambiguity related to localization of the objects in the training images in a weakly supervised detection method. "simpler" bounding boxes (without clutter)is interpreted as "more positive or accurate" localization. Self-paced learning based on gradual learning which is basically used to reduce the noise while trainingdeepnetworks [3].

II. RELATED WORK

Marc-Andr´eCarbonneauproposed Multiple Instance Learning (MIL) is a kind of weakly supervised learning where training data is arranged into sets called as bags and label is given to bag. Due to this gaining interest in various problem solutions which works on weakly labeled data.

The MIL assumption states that all negative instances are contained in negative bags, and at least one positive instance should contain in the positive bags. Let P be the label of a bag Q, defined as a set of feature vectors Q = {q_1, q2, qN}. Each instance (i.e. feature vector) q_i corresponds to a label p_i. The label of the bag is given by:

$$P = +1 \text{ if } \exists p_i: p_i = +1;$$
$$-1 \text{ if} \forall p_i: p_i = -1. \quad (1)$$

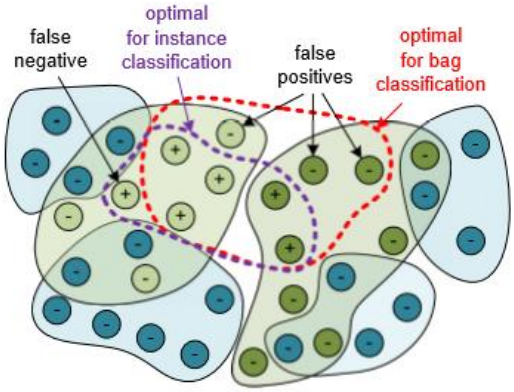

Figure1: Dotted boundary for optimal instance and optimal bag classification

Under the MIL assumption, when the witness is present in a bag it is considered as a positive label and remaining labels of instances are neglected. In such a case, classification accuracy does not depend on false positives and false negatives. In the case of negative bags, misclassification will be possible if it contains a single false positive.The green circles represent positive bags, while the blue circle represents negative bags. Both dotted lines decision boundaries are best for bag and instance classification as shown in Fig. 1. However, only one purple boundary achieves perfect instance classification. So MIL uses bag accuracy is an optimization criteria[4].

RamazanGokberkCinbisproposedMultifold MIL, It contains two detectors. The first detector uses a multi-fold procedure to train the detector, similar to cross-validation, within the MIL iterations [5]. This technique divides the positive training images into K disjoint folds and it uses a trained detector for re-localizing the images from positive images. In the next step, the second detector is trained using all selected windows. This detector is used for mining on negative training which acts as the final detector. In this approach, the number of folds plays a very important role. It has pros and cons too 1) Re-localization accuracy increases when the number of training samples per fold increases. 2) But more folds increase the computational cost too. The drawback of the multifold MIL procedure is eliminated with the help of the window refinement method. It is used for improving the localization accuracy [6].

YunhangShen, RongrongJi proposed Object specific pixel gradient method. it focuses on the input of the neural network rather than the output of the neural network. The input of the neural network is given in terms of object-specific pixel gradient called an OPG map and the output of the neural network is called a score map. Object is strongly localized when the resolution of OPG map is equal to the resolution of an image.

Another technique uses selective search for extracting region of interest as per the best classification score. In contrast, this technique masks the original images as per the pixel contribution to the output object class.In these two classifiers are used for classification and localization task. For fine-tuning it requires more iteration. It also requires additional forward and backward propagation for testing[7].

Wenhui Jiang &Zhicheng Zhao proposed the framework of Attention based object detection. It takes input as an image and a set of regions that are extracted with the help of selective search [8]. This framework has convolution layers that are used for extracting the features from input image. These are given as input to the object representation block.The main task of object representation block is to find attention maps with the help of fully convolution network. Attention pooling layer plays important role for object localization task.Here only the image level labels are used for learning.This attention based object detection method enhances occlusions by mitigating noisy backgroundas per [1].

As surveyed from literature, it is observed that Object detection is a major and essential task yet to be deciphered in computer vision. Remarkable results have been achieved on large-scale detection benchmarks by fully-supervised object detection (FSOD) methods, especially with the convenience of deep convolutional neural networks (CNNs), whose success mainly benefits from the flexibility of deep learning models and an abundance of instance-level annotations in extensive datasets. However, annotating such large scale datasets is expensive and time-consuming. More importantly, the performance of FSOD is profoundly affected by the quality of these annotations. For imperfect bounding box annotations or missing annotations of objects in training images can have a drastic impact on FSOD performance as stated in [9].

It is also observed that with the advancement of deep learning, many CNN based methods have been proposed to solve the FSOD problem, such as Fast RCNN, Faster RCNN [10], SSD, YOLO2, and many of their variants. Faster RCNN is a typical proposal based detection CNN, which balances both detection performance and computational efficiency but the performance of FSOD detectors is largely affected by the number of missing annotations and this can be solved by WSOD.

IV. PROPOSED METHOD

Existing methods for object detection from image or video has low detection rate to identify the type of object and its localization. A CNN based region proposal uses bounding box annotations to train a proposal network, where sliding window style method is used to identify the object locationand it comes under fully supervised approach.

The detection time of region proposal method using selective search and RCNN is still relatively slow so novel region proposal network of Faster RCNN requires 3 to 10 ms time to detect an object in an image.The accuracy of the network is fully dependent on the region proposal network of Faster RCNN.

We propose a self-paced learning algorithm for training a region proposal network for WSOD which is computationally powerful. This approach will be able to boost model training by pseudo-labeling which helps to enhance detection performance without manual annotations. Fig. 2.Shows the architecture of a weakly supervised object detector using Faster RCNN.It is divided into two networks i.e. Fast RCNN and Region Proposal Network (RPN).

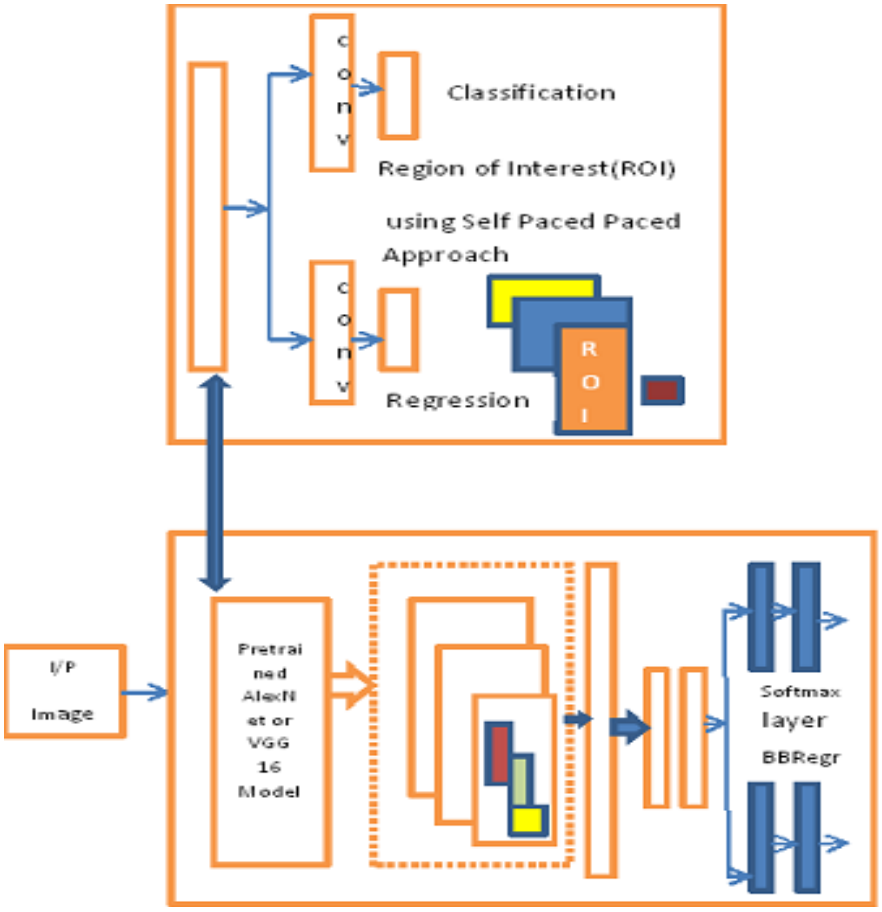

Figure 2: Proposed architecture for the weakly supervised object detection.

1) Training the RPN network effectively is very important because the accuracy of this proposed architecture depends on the RPN network (upper part of Figure 2) so instead of traditional approach of training.we propose self-paced approach to fine tune the RPN network.
2)Train the Fast RCNN (lower part of Figure 2) using another convolution network, it passes feature maps to a fully connected layer ofsoftmax and a linear regression layer. It predicts the accurate localization of the identified object.
3) Classification and regression loss is back propagated to minimize the loss.
4) We apply our method on KITTI contains 7481 training images and Cityspaces dataset contains 2945 training images.

V. ALGORITHM

The institution behind self-paced learningcomes from human learning which is useful for vision problems. In this, learning starts with easy examples and then gradually learns hard examples[11].

Simultaneously estimates easiness of example and learn the parameters. Easiness is the property of datasets, not singe instance.

Self-Paced Learning Algorithm steps.

1. Start with Initial estimate W_0
2. Update $h_i = \max_{h \in H} W^T \Psi (xi,yi , h)$
3. Set the indicative variables
   $V_i \in \{0,1\}$
4. Update w_{t+1} by solving a convex problem
   Min $\|W\|^2 + C \sum_i V_i \xi_I - \sum V_i / K$
   $W^T \Psi (x_i,y_i,h_i) - W^T \Psi (x_i,y,h) \geq \Delta (y_i,y,h) - \xi i$
5. Choose next value of K for next set of images.
6. Repeat step 2 to 4 till it converge
7. Stop

In this h is latent variable, (x_i,y_i,h_i) is feature vector of input image and (x_i,y,h) is predicted output.

Atthe initial step keep K that is self-paced learning rate at maximum value and keep on decreasing it as per iteration. First $V_i$set contains easy samples, train the model for the same and repeat the steps 2 to 4 for another set of sample hard images till it converges.

The self-paced iteration will use for selecting subset of easy classes as well as easy samples of these classes. This approach is applied to region proposal network as weakly supervised network of Faster R-CNN without using selective search, which made the algorithm much faster.

III. RESULTS

Studies of the literature in the form of parameters such as detection paradigms used, mAP/WR, Datasets used and its limitations are given in Table 1.

Table 1: Deep learning techniques for weakly supervised object detection

Algorithm	Detection Paradigms	Mean Average Precision(mAP)/Witness Rate(WR)	Datasets	Backbone Architecture	Limitation
Multiple Instance Learning	Saliency Detector	25.5%(WR)	SILVAL	NA	1)Accuracy of object detector depends on the initial classifier. 2)The crowded background causes ambiguity between targrt object and background
Multifold Multiple Instance Learning	Fisher Vector + CNN	47.3% mAP (training) 27.4 % mAP(Testing)	PASCAL VOC 2007	--	More folds increases the computational cost too
Object Specific Pixel Gradient	OPG+FRCNN	44.5% mAP	PASCAL VOC 2007	AlexNet	For fine tuning of both the classifiers (classification and localization) requires more no of iterations.
Attention based object detection	FCNN	30.6 to 32.9% mAP (training) 32.9 % mAP (Testing)	PASCAL VOC 2007	AlexNet	1)Gradual learning due to selective search algorithm so it requires high computation time. 2)Enhances occlusions by eliminating noisy background.
		31.0% mAP	PASCAL VOC 2010		
		31.3% mAP	PADCAL VOC 2012		

VI. CONCLUSION

Object detector training with image-level labels only is an essential task in a range of applications like Self driving cars,Security: video surveillance, Retail,etc. The traditional weakly supervised object detection approacheswrongly predicting and localizing the target class. To eliminate this problem, Self-paced learning based approach for region proposal network of Faster RCNN in a WSOD scenario is proposed. This self-paced learning paradigm train easiest samples first and gradually train hard samples, so it is a powerful mechanism to reduce noise while training so as to get an accurate result. Finally, it usesSoftmax and linear regression layer of Faster-RCNN predicting the bounding boxes for the identified objects.

REFERENCES

1. Wenhui Jiang & Zhicheng Zhao & Fei Su,"Weakly \supervised detection with decoupled attention-based deep representation",Multimed Tools Appl (2018) 77:3261–3277 DOI 10.1007/s11042-017-5087.
2. Y. Bengio, J. Louradour, R. Collobert, and J. Weston, "Curriculum learning," in Proc. Int. Conf. Mach. Learn., 2009, pp. 41–48
3. Enver Sangineto , Moin Nabi,"Self Paced Deep Learning for Weakly Supervised Object Detection",IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 41, no. 3, march 2019
4. Marc-Andr´e,Carbonneau,Veronika Cheplygina,"Multiple Instance Learning: A Survey of Problem Characteristics and Applications"arXiv:1612.03365v1 [cs.CV] 11 Dec 2016
5. Ramazan Gokberk Cinbis, Jakob Verbeek, and Cordelia Schmid,"Weakly Supervised Object Localization with Multi-Fold Multiple Instance Learning"IEEE Transactions on Pattern Analysis and machine intelligence, vol. 39, no. 1, January 2017
6. R. G. Cinbis, J. Verbeek, and C. Schmid, "Multi-fold mil training for weakly supervised object localization," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2014, pp. 2409–2416.
7. Yunhang Shen, Rongrong Ji,Changhu Wang,"Weakly Supervised Object Detection via Object-Specific Pixel Gradient",IEEE Transactions on Neural Networks and Learning Systems, vol. 29, no. 12, December 2018
8. J. R. R. Uijlings, K. E. A. van de Sande, T. Gevers, and A. W. M. Smeulders, "Selective search for object recognition," Int. J. Comput. Vis., vol. 104, no. 2, pp. 154–171, Apr. 2013
9. MengmengXuYanchengBai,"Missing Labels in Object Detection" Institute of Software, Chinese Academy of Sciences (CAS)
10. Bin Liu, Wencang Zhao and Qiaoqiao,"Study of Object Detection Based On Faster R-CNN" Chinese Automation Congress (CAC)-2017
11. M. P. Kumar, B. Packer, and D. Koller, "Self-paced learning for latent variable models," in Proc. Int. Conf. Neural Inf. Process. Syst., 2010, pp. 1189–1197

AUTHORS PROFILE

Ms. Jyoti G. Wadmare, is Ph.D. scholar in the university of Mumbai and an Assistant Professor in the Department of Computer Engineering, KJSIEIT, university of Mumbai. She has nearly 13 years of teaching experience. She has publications in International and national conferences.

Dr. Sunita R. Patil, is a Vice Principal, Professors at KJSIEIT, Mumbai, University of Mumbai (UoM), India. She is a member, Board of Studies in Computer Engineering, UoM. She received Ph.D. in Computer Engineering in the domain Data Mining, Big Data & Data Science. She is having around 20 years of teaching & administrative experience. She has publications her research work in various recognized National/International Journals and conferences. She has visited various international universities and organizations for attending conferences, knowledge sharing & exchange of information. Her passion is to bring in various outcome based reforms in the field of academics contributing to the growth of society, nation and world at large.