Object Localization in Weakly Labeled Data Using Regularized Attention Networks

Eu Wern Teh*, Zhenyu Guo†, Yang Wang*

* *Department of Computer Science, University of Manitoba, Winnipeg, MB R3T 2N2, Canada*

† *Sengled Canada*

* {umteht, ywang}@cs.umanitoba.ca, † zhenyu.guo@sengled.com

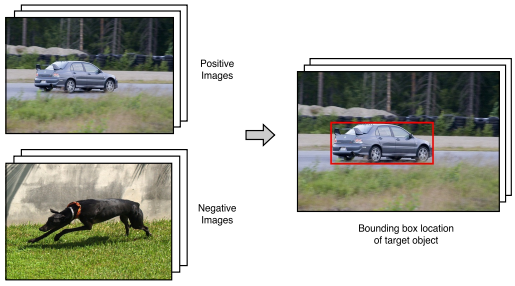


Fig. 1. For an object of interest (e.g. "car"), we have a set of positive images containing this object, and another set of negative without this object. Our goal is to localize the bounding box location of the target object in each positive image.

***Abstract*—We consider the problem of weakly supervised object localization. For an object of interest (e.g. "car"), an image is weakly labeled when its label only indicates the presence/absence of this object, but not the exact location of the object in the image. Given a collection of weakly labeled images for an object, our goal is to localize the object of interest in each image. We propose a novel architecture called the *regularized attention network* for this problem. Our work builds upon the *attention network* proposed in [1]. We extend the standard attention network by incorporating a regularization term that encourages the attention scores of object proposals to mimic the scoring distribution of a strong fully supervised object detector. Despite of the simplicity of our approach, our proposed architecture achieves the state-of-the-art results on several benchmark datasets.**

object localization, weakly supervised learning, attention-based neural network, attention network, weakly supervised object localization

I. Introduction

We address the problem of weakly supervised object localization. For an object of interest (e.g. "car"), we have a collection of weakly labeled images for this object. Each image is annotated with the presence/absence of this object, but not the exact bounding box location in the image. Given a set of such weakly labeled images with only image-level label, our goal is to localize the object of interest in each image. See Fig. 1 for an illustration.

The traditional pipeline of learning object detectors require a large amount of training images annotated with object bounding boxes. It usually requires human annotations in order to collect training data in order to learn object detectors. However, human annotation is often expensive and time-consuming. In contrast, it is very easy to collect classification labels for images, thanks to the social media sites such like Flickr. If we can successfully solve the weakly supervised object localization problem, it can provide an affordable alternative for collecting training data for learning object detectors.

Our work is based on the attention network in [1] for weakly supervised object localization. The attention network first extracts object proposals in each image. It then assigns an *attention score* to each object proposal. The object proposal with the highest attention score is selected as the final localization result.

A limitation of the original attention network is that the attention scores of object proposals in an image often have lots of uncertainty. In [2], it is observed that more than 90% of proposals have very low detection confidence scores when we apply a strong object detector on them. In other words, the score distribution of a strong object detector should have high peak in a small portion of the proposals. In this paper, we introduce the *regularized attention network*. Our proposed model adds an additional regularization term in attention network, so that the attention scores of the proposals mimic the distribution of the detection scores of a strong object detector. Despite of the simplicity of our approach, our proposed method outperforms other state-of-the-art methods on benchmark datasets.

II. Related Work

There has been a series of work on weakly supervised object localization [3], [4], [5], [6], [7], [15]. Many of them use some form of multiple instance learning to solve the weakly supervised learning problem. Bilen [3] propose an end-to-end architecture that combines object classification and detection in a single network. In [4], Bilen use latent SVM by treating bounding boxes as latent variables.

Our work is closely related to a line of work on using attention in deep neural networks in various applications, e.g. machine translation [8], action recognition [9], image caption [10], etc. The attention network in [1] is closed to ours.

978-1-5386-0462-5/17/$31.00

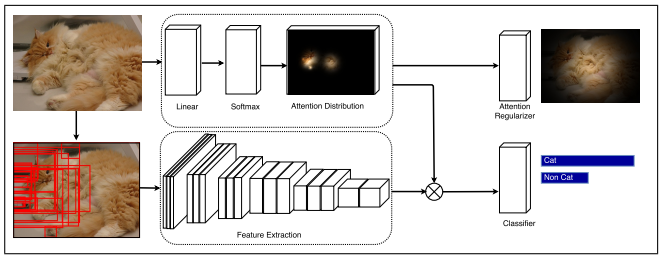


Fig. 2. An overview of our approach. We introduce an attention distribution regularizer to the attention network [1]. We first extract features from each proposals (note that this is a preprocessing step). The proposal features are then forwarded to a linear layer followed by a SoftMax layer to generate the attention scores. After that, the attention scores are multiplied with the corresponding proposal features to generate a whole image feature. This whole image feature is then used to generate a classification score. Finally, we regularize the distribution of attention scores to mimic detection scores of a strong object detector and classifying the image using the whole image feature.

III. Our Approach

Our proposed approach is based on the attention network in [1]. In this section, we first give a brief introduction to the background on attention networks (Sec III-A). We then describe how to modify the original attention network by introducing regularization on the attention scores (Sec. III-B). See Fig. 2 for the overall architecture of our approach.

A. Background: Attention Networks

Our approach is based on the attention networks for weakly supervised object localization proposed in [1]. The attention network consists of three components: object proposals, proposal attention and classification.

Object proposals: For a given object of interest (e.g. "car"), we have a collection of weakly labeled images where each image has a binary label indicating the presence/absence of the object in the image. We first generate K object proposals in each image by using the edge boxes method [11]. Each proposal is a bounding box that may contain any object. Next, we use an existing CNN model implemented in Caffe [12] to extract a 4096 dimensional feature for each proposal.

Proposal attention: For an image $\mathbf{x}$, we use $\mathbf{x}_i$ $(i = 1, 2, ..., K)$ to denote the i-th proposal in the image. We then compute an attention score s_i to indicate the likelihood of the proposal to contain the object of interest. We make the attention scores to form a probability distribution by applying a softmax function to all the proposal attentions in a each image. I.e., the attention score s_i is calculated as:

$$s_i = \frac{\exp(\mathbf{w}_a^\top \mathbf{x}_i)}{\sum_{j=1}^{K} \exp(\mathbf{w}_a^\top \mathbf{x}_j)} \tag{1}$$

Classification: We combine the proposal attention scores and the proposal features to obtain an image level feature. We then use the image level feature to perform a binary classification using the logistic loss:

$$\mathbf{z} = \sum_{i=1}^{K} s_i \mathbf{x}_i, \quad f(\mathbf{x}; \{\mathbf{w}_a, \mathbf{w}_c\}) = \mathbf{w}_c^\top \mathbf{z} \tag{2}$$

$$\ell(\mathbf{x}, y; \{\mathbf{w}_a, \mathbf{w}_c\}) = \log(1 + \exp(-y \cdot f(\mathbf{x}; \{\mathbf{w}_a, \mathbf{w}_c\}))) \tag{3}$$

where $y \in \{+1, -1\}$ indicates the ground-truth label of the image $\mathbf{x}$.

Given a set of N training images and their weak labels $\{\mathbf{x}^n, y^n\}_{n=1}^N$, the model parameters $\{\mathbf{w}_a, \mathbf{w}_c\}$ are learned by minimizing the logistic losses of the all training images:

$$\min_{\mathbf{w}_a, \mathbf{w}_c} L(\mathbf{w}_a, \mathbf{w}_c) = \sum_{n=1}^{N} \ell(\mathbf{x}^n, y^n; \{\mathbf{w}_a, \mathbf{w}_c\}) \tag{4}$$

In the end, we consider the proposal $\mathbf{x}_i$ with the highest attention score as the localized object in this image $\mathbf{x}$.

B. Regularizing Attention Networks

In Eq. 1, the attention score of one proposal will depend on other proposals in this image due to the softmax operation. In the case where proposals are equally likely to contain the object of interest, it will cause the model to have high degree of uncertainty. For example, Figure 3 shows an example to illustrate this. In Fig. 3(left), we show three object proposals with the highest attention scores produced by the original attention network [1]. We can see that the top two proposals (denoted by red and blue colors) almost have the same attention scores. Since the attention scores are used later to localize the object, this uncertainty can cause problem in the end. It is very easy for the model to pick the wrong proposal as the final localization, when the model is presented with two equally promising proposals.

To resolve this uncertainty, we need a way for the original model [1] to be "focused", i.e. the attention score should ideally concentrate on one proposal. Our contribution to this paper is to come out with a way to regularize the proposal attention distribution to mimic the detection score distribution of a strong object detector. Our approach is partly motivated by [2]. A key observation in [2] is that if we apply a strong object detector on all object proposals, more than 90% of the object proposals will have very low detection scores. Only a small portion of the object proposals will have high detection scores. Base on this observation. we want the attention scores of our model to mimic the behavior of a strong object detector. In particular, we would like to encourage the attention scores to form a "peak" distribution where most of the high attention scores are concentrated on a small number of object proposals.

We implement this idea by introducing an additional regularization term on the distribution of attention scores. This regularization term will encourage the distribution of attention scores to be far away from the uniform distribution. Equivalently, this will force the distribution of attention scores to be peaked around a small number of object proposals.

We use the Kullback-Leibler divergence [13] as a way to measure the difference between the current attention distribution s and the uniform distribution. For a discrete distribution with K bins, this regularization is defined as:

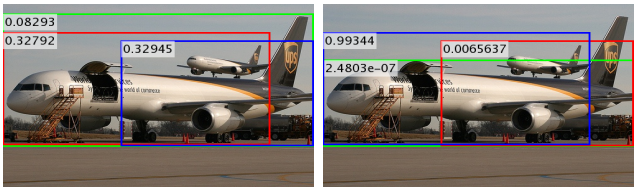
$$R(s) = \sum_{i=1}^{K} s_i \log(\frac{s_i}{1/K}) \quad (5)$$

and our objective is to maximize this regularization term, because we want the attention distribution to be as different as possible from the uniform distribution.

Combing Eq. 3 and Eq. 5, we learn the parameters in our model by minimizing the following regularized loss function:

$$\min_{\mathbf{w}_a, \mathbf{w}_c} L(\mathbf{w}_a, \mathbf{w}_c) - \lambda \cdot R(s) \quad (6)$$

The hyperparameter λ can be used to control the relative contributions of the two terms in Eq. 6. In our experiments, we simply set $\lambda = 1$. The problem in Eq. 6 can be solved using standard software frameworks (e.g. Torch).



before after

Fig. 3. A visualization of the top 3 proposals with their attention scores before and after applying our method. The model (left) has lots of uncertainty since the top two proposals have similar attention scores. By regularizing the attention score distribution, we force the model to concentrate the attention on one of them. This helps resolve the ambiguity and produce the correct localization results. See Fig 4 for the change in attention distribution.

Figure 4 visualizes the distribution of attention scores on the object proposals shown in Fig. 3 before and after we regularize the attention distribution.

IV. Experiments

We first present the experiment setup and some implementation details IV-A. Then we evaluate the proposed approach on three datasets: PASCAL VOC 2007 dataset (Sec. IV-B), YouTube-Objects dataset(Sec. IV-C) and YouTube-Objects-Subset dataset (Sec. IV-D).

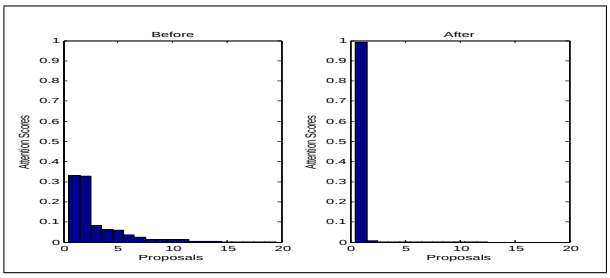


Fig. 4. A comparison of attention score distribution before (left) and after (right) regularizing the attention distribution. With the regularization, the distribution becomes more peaked and the attention score concentrates more on the first proposal.

A. Experiment Details

We train our model as a binary classifier for each class in the dataset. All images consist of a given class is considered as positive examples. To maintain class balance, we randomly select an equal number of negative examples from each of the remaining classes.

Following [1] and [14], we generate 40 proposals for each image in Pascal VOC 2007 dataset and 20 proposals for each image in YouTube-Objects and YouTube-Objects-Subset dataset. A dropout rate of 0.8 is used to regularize our network.

B. PASCAL VOC 2007

The PASCAL VOC 2007 dataset consists of images of 20 object classes. Similar to previous work [1], [3], [14], we train our model and evaluate CorLoc by using all of the training and validation images. Table I shows the result of our CorLoc performance. We compare our method with several state-of-the-art methods [1], [3], [14] on weakly supervised object localization. Our method outperforms all of them.

Fig. 5. A visualization on how our method affects the attention scores of the proposals (1st column: attention network [1]; 2nd column: ours). The brightness indicates the attention scores of each proposals. We use a Gaussian mask with sigma proportional to attentions scores, width and height of each proposals.

Figure 5 shows the visualization of the attention scores of our method compared with [1]. We can see that the attention scores produced by our model tend to be peaked. We show some examples of the final localization in Fig. **??**.

C. YouTube-Objects

The YouTube-Objects dataset [5] contains videos of 10 different object classes. We train and evaluate our model by using all video frames with bounding box annotation. Table II shows the result of our CorLoc performance. Our method outperforms the state-of-art methods in this dataset.

D. YouTube-Objects-Subset

The YouTube-Objects-Subset dataset [18] is a subset of YouTube-Objects dataset [5]. This dataset is unique because it provides segmentation mask as ground truth rather than

TABLE I
CORLOC RESULTS ON *positive* TRAINVAL SUBSET OF THE PASCAL VOC 2007 DATASET. WE ALSO COMPARE OUR APPROACH WITH SEVERAL STATE-OF-THE-ART APPROACHES [15], [14], [3], [1].

method	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	mbike	person	plant	sheep	sofa	train	tv	avg
[15]	80.10	63.90	51.50	14.90	21.00	55.70	74.20	43.50	26.20	53.40	16.30	56.70	58.30	69.50	14.10	38.30	58.80	47.20	49.10	60.90	48.50
[3]	68.90	**68.70**	65.20	42.50	**40.60**	72.60	**75.20**	53.70	29.70	68.10	33.50	45.60	65.90	**86.10**	27.50	**44.90**	76.0	**62.40**	66.30	66.80	58.00
[14]	78.57	63.37	66.36	56.35	19.67	82.26	74.75	69.13	22.47	72.34	31	62.95	74.91	78.37	48.61	29.39	64.58	36.24	75.86	69.53	58.84
[1]	84.03	64.61	70.00	62.43	25.82	80.65	73.91	71.51	**35.73**	81.56	46.50	71.26	**79.09**	78.78	56.72	34.29	69.79	56.72	77.01	72.66	64.59
ours	**84.87**	66.26	**71.82**	**65.75**	25.41	**83.87**	74.89	**74.48**	34.38	**85.11**	**55.00**	**73.63**	**79.09**	79.59	**62.05**	34.69	**77.08**	58.08	**77.39**	**75.39**	**66.94**

TABLE II
CORLOC RESULTS ON THE YOUTUBE-OBJECTS DATASET [5].

method	aero	bird	boat	car	cat	cow	dog	horse	bike	train	avg
[16] (video)	25.12	31.18	27.78	38.46	41.18	28.38	33.91	35.62	23.08	25	30.97
[17]	**65.4**	**67.30**	38.9	65.2	**46.3**	40.2	**65.3**	48.4	39	25	50.1
[14] proposal only	51.69	54.84	32.54	**85.71**	14.53	75.68	55.65	53.42	51.69	39.29	51.50
[14] proposal + transfer	56.04	30.11	39.68	**85.71**	24.79	**87.83**	55.65	**60.27**	61.8	**51.79**	55.37
[1]	55.07	62.37	43.65	84.62	28.21	66.22	58.26	53.42	**62.92**	39.29	55.40
ours	57.00	60.22	**45.24**	84.62	28.21	67.57	62.61	57.53	**62.92**	41.07	**56.70**

bounding boxes. We train and evaluate our model on all the frames in this dataset.

During evaluation, our model will first produce bounding box. We then use grab-cut algorithm [19] on the bounding box to generate segmentation mask. CorLoc is measured based on pixel level IoU region. Table III shows the result of our CorLoc performance. Again, our method outperforms the state-of-art methods in this dataset.

TABLE III
CORLOC RESULTS ON THE YOUTUBE-OBJECTS-SUBSET DATASET.

method	aero	bird	boat	car	cat	cow	dog	horse	bike	train	avg
[14] proposal only	42.23	51.24	29.54	67.76	14.75	50.20	47.02	22.18	16.44	18.84	36.02
[14] proposal + transfer	45.74	55.47	39.51	58.75	26.51	55.00	43.51	33.71	32.76	**25.63**	41.66
[1]	49.19	45.52	43.94	69.32	26.43	60.24	56.03	40.39	**40.39**	19.91	45.10
ours	**51.72**	**69.24**	**46.92**	**70.54**	**29.41**	**61.48**	**62.53**	**43.08**	36.40	14.46	**48.58**

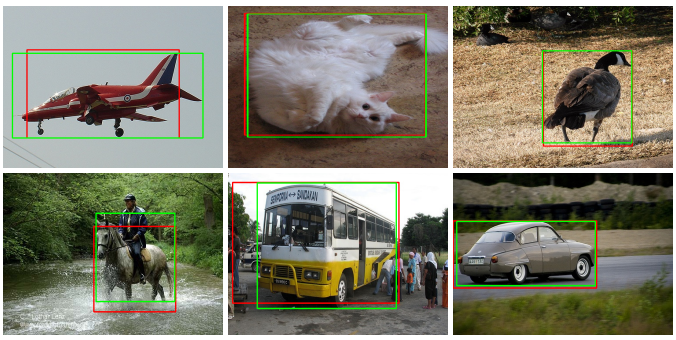


Fig. 6. Qualitative examples of our approach on the PASCAL VOC 2007 trainval dataset. Green boxes show the ground-truth localization. Red boxes show the localization produced by our model.

V. CONCLUSION

We have proposed a new method to regularize the attention distribution for an attention network [1]. Our experiments demonstrate that adding this regularizer does improve the performance of the existing model and at the same time outperform other state-of-the-art methods on weakly supervised object localization.

Acknowledgement: This work was supported by NSERC Discovery and Engage grants. We thank NVIDIA for donating some of the GPUs used in this work.

REFERENCES

[1] E. W. Teh, M. Rochan, and Y. Wang, "Attention networks for weakly supervised object localization," in *British Machine Vision Conference*, 2016.
[2] Y. Li, L. Liu, C. Shen, and A. van den Hengel, "Image co-localization by mimicking a good detector's confidence score distribution," in *European Conference on Computer Vision*, 2016.
[3] H. Bilen and A. Vedaldi, "Weakly supervised deep detection networks," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016.
[4] H. Bilen, M. Pedersoli, and T. Tuytelaars, "Weakly supervised object detection with posterior regularization," in *British Machine Vision Conference*, 2014.
[5] A. Prest, C. Leistner, J. Civera, C. Schmid, and V. Ferrari, "Learning object class detectors from weakly annotated videos," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2012.
[6] H. O. Song, Y. J. Lee, S. Jegelka, and T. Darrell, "Weakly-supervised discovery of visual pattern configurations," in *Advances in Neural Information Processing Systems*, 2014.
[7] K. Tang, A. Joulin, L.-J. Li, and L. Fei-Fei, "Co-localization in real-world images," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2014.
[8] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," in *International Conference on Learning Representations*, 2015.
[9] S. Sharma, R. Kiros, and R. Salakhutdinov, "Action recognition using visual attention," in *ICLR Workshop*, 2016.
[10] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhutdinov, R. Zemel, and Y. Bengio, "Show, attend and tell: Neural image caption generation with visual attention," in *International Conference on Machine Learning*, 2015.
[11] C. L. Zitnick and P. Dollar, "Edge boxes: Locating object proposals from edges," in *European Conference on Computer Vision*, 2014.
[12] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell, "Caffe: Convolutional architecture for fast feature embedding," *arXiv:1408.5093*, 2014.
[13] "Kullback-leibler divergence," https://en.wikipedia.org/wiki/Kullback-Leibler_divergence, accessed: 2016-11-10.
[14] M. Rochan and Y. Wang, "Weakly supervised localization of novel objects using appearance transfer," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2015.
[15] C. Wang, W. Ren, K. Huang, and T. Tan, "Weakly supervised object localization with latent category learning," in *European Conference on Computer Vision*, 2014.
[16] A. Joulin, K. Tang, and L. Fei-Fei, "Efficient image and video co-localization with frank-wolfe algorithm," in *European Conference on Computer Vision*, 2014.
[17] A. Papazoglou and V. Ferrari, "Fast object segmentation in unconstrained video," in *IEEE International Conference on Computer Vision*, 2013.
[18] K. Tang, R. Sukthankar, J. Yagnik, and L. Fei-Fei, "Discriminative segment annotation in weakly labeled video," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2013.
[19] C. Rother, V. Kolmogorov, and A. Blake, "Grabcut: Interactive foreground extraction using iterative graph cuts," in *SIGGRAPH*, 2004.