Attention Networks for Weakly Supervised Object Localization

Eu Wern Teh
umteht@cs.umanitoba.ca
Mrigank Rochan
mrochan@cs.umanitoba.ca
Yang Wang
ywang@cs.umanitoba.ca

Department of Computer Science
University of Manitoba
Winnipeg, MB, Canada

Abstract

We consider the problem of weakly supervised learning for object localization. Given a collection of images with image-level annotations indicating the presence/absence of an object, our goal is to localize the object in each image. We propose a neural network architecture called the *attention network* for this problem. Given a set of candidate regions in an image, the attention network first computes an attention score on each candidate region in the image. Then these candidate regions are combined together with their attention scores to form a whole-image feature vector. This feature vector is used for classifying the image. The object localization is implicitly achieved via the attention scores on candidate regions. We demonstrate that our approach achieves superior performance on several benchmark datasets.

1 Introduction

We consider the problem of localizing objects from weakly labeled images. For an object category (e.g. "dog"), we have a collection of images, where the labels are only given at the image level. If an image has a positive label, we know there is an object of interest (i.e. "dog") somewhere in the image. But we do not know the exact location of the object in the image. If an image has a negative label, we know that this object is not in the image. From such weakly labeled data, we would like to localize the object of interest in the positive images. Note that there might be multiple instances of the object in a positive image. Our goal is to localize one of those instances. See Fig. 1 for an illustration of our problem.

The field of visual recognition (especially image classification) has witnessed significant success in the past few years. An important enabling force of this success is the availability of large amount of training data, e.g. ImageNet [4]. Most successful image classification systems (e.g. AlexNet [10]) is based on learning a deep convolutional neural network (CNN) from large amount of labeled data.

In addition to image classification, another important task in computer vision is object detection, where the goal is to find the exact location of objects in an image. In order to learn object detectors, we usually need training images where the object locations are manually

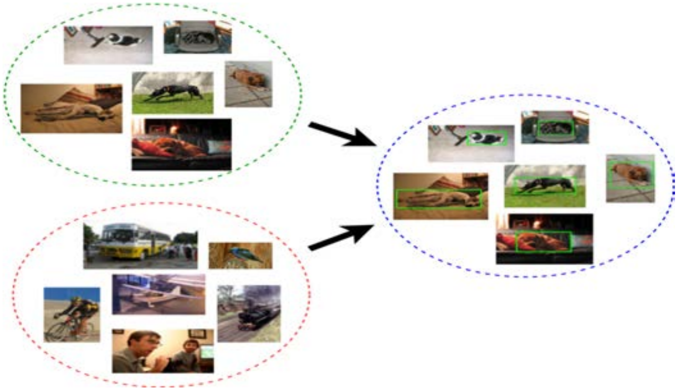

Figure 1: Our aim is to localize the object of interest in weakly labeled data. For a given object category (e.g. "dog"), we have a collection of positive images (top left) where we know there is a dog somewhere in each image, but we do not know the exact location of the dog in an image. In addition, we have a collection of negative images (bottom left) where there are no dogs in them. From such weakly labeled data, our goal is to localize the object in positive images (right).

annotated, e.g. in terms of bounding boxes. Collecting images with bounding box annotations is often expensive and time-consuming. Currently, most of the datasets for object detection are limited to a small number of object categories, e.g. 20 object classes in the PASCAL VOC dataset [7] and 80 object categories in the MS COCO dataset [12]. It is not clear how to scale up the labeled images to cover hundred thousands of object categories. In contrast, it is relatively easy to collect "weakly labeled" data, where images are labeled only at the whole image level. These labels indicate whether an object appears in the image or not, but they do not provide the exact object locations in the image. For example, these weak labels can be obtained from user-generated tags from Flickr images and YouTube videos. A reliable solution to weakly supervised object localization will provide an inexpensive way of collecting datasets for learning object detectors.

In this paper, we propose a new approach for localizing objects in weakly labeled data. The novelty of our method is to introduce the concept of "attention" in weakly supervised learning. Our approach starts with generating a set of candidate object regions in each image using standard object proposal techniques. For each object proposal, instead of directly predicting its class label, we first compute an "attention score". This attention score indicates the importance of each object proposal. We then combine the object proposals in the image using their respective attention scores to form a whole image feature vector. This feature vector is then used to classify this image. Since the feature vector for whole image classification is obtained from candidate regions using their attention scores, this will focus the model to learn to assign high attention scores to regions that contain the object of interest.

The main contribution of this paper is to incorporate the attention mechanism in object localization from weakly labeled data. Although our method is simple, our experimental results demonstrate that it achieves superior performance on several benchmark datasets.

2 Related Work

There has been a line of work on weakly supervised learning (WSL) of visual concepts, such as objects, actions. Many existing approaches [2, 11, 20, 21, 24] formulate WSL as multiple instance learning (MIL) [6]. In the MIL framework, each image is treated as a bag of instances, where each instance corresponds to a candidate region in the image. Each image is labeled as positive or negative to indicate whether the object of interest appears anywhere in the image. If an image is labeled as positive, at least one of the bag instances (i.e. regions) is assumed to be the object of interest. If an image is labeled as negative, no regions contain the object of interest. Using this weak supervision at the bag level, the MIL alternates between estimating the object appearance model and selecting the bag instances from positive bags that correspond to the object. Since the MIL formulation results in a non-convex optimization problem, the learning algorithm tends to get stuck in local optimum. This issue has been addressed by several works. Kumar *et al.* [11] propose a self-page learning method that starts with easy training examples and progressively adds hard examples. Song *et al.* [21] discover discriminative configurations of multiple patches to overcome the issue of mislocalizations. Song *et al.* [20] and Bilen *et al.* [3] propose smoothed version of MIL that softly label instances instead of choosing the best labels. Deselaers *et al.* [5] and Tang *et al.* [23] enforce the appearance similarity of selected positive regions across images. Prest *et al.* [15] uses a model similar to [5] for learning object detectors from videos.

Convolutional neural networks (CNN) [10] have achieved great success in computer vision in recent years. There has been some work on adapting CNNs for object localization in weakly labeled data. Oquab *et al.* [13] adapt a CNN trained for object classification to predict the approximate locations of objects in images. Bilen *et al.* [2] propose a two-stream architecture that combines object classification and detection in a single network.

Our work is also related to a line of research on incorporating attention mechanism in standard deep learning models. The attention mechanism allow the deep neural network to focus on a small part of the input images for high-level recognition tasks. It has been successfully applied in machine translation [1], action recognition [18], image captioning [25], visual question answering [19], etc.

3 Our Approach

In this section, we introduce our attentional network for object localization in weakly labeled images. The overview of our approach is illustrated in Fig. 2.

For a given object class (say “dog”), we assume that we have a collection of weakly labeled images. For each image in the collection, we only have the label at the image level. If an image is labeled as positive (+1), we know that there is at least a dog somewhere in the image. If an image is labeled as negative (-1), there is no dog anywhere in the image. Our goal is to localize the dog in each of the positive images. If there are multiple instances of the object (i.e. multiple dogs in an image), most previous work usually only tries to localize one of them. We will operate on the same assumption. If we can reliably localize the object in weakly labeled data, the localization results can potentially be used as supervision to learn an object detector.

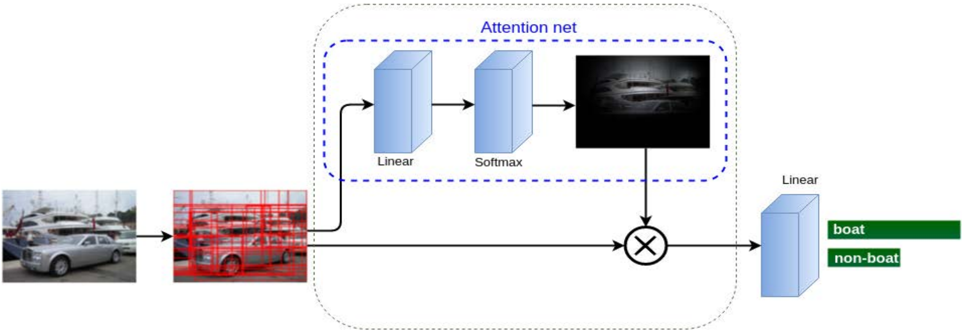

Figure 2: An overview of our architecture. Given an image, we extract proposals that are likely to contain *any* object. Each proposal is passed to a linear layer to obtain its attention score. We then apply the softmax operation to the attention scores before multiplying it with its corresponding proposal features. This gives a whole image feature vector that is the weighted average of proposals. Finally, we use the whole image feature to classify the image.

3.1 Attention networks

Object proposals: Given a collection of weakly labeled images, the first step of our approach is to generate a shortlist of object proposals in each image. We use the edge boxes method [26], which is a commonly used technique for generating object proposals. Each proposal is a bounding box that may contain any object. This method is based on a simple observation – the number of contours contained in a bounding box is a good indication of how likely this box contains an object. Given a candidate bounding box in an image, the edge boxes algorithm assigns an objectness score by examining the number of edges in the box and those that overlap the box's boundary.

Let $\mathbf{x}$ be the input image and K be the number of object proposals generated on the image $\mathbf{x}$. To simplify the notation, we assume that we get the same number of object proposals on each image, although this is not a requirement of our method. We represent each proposal as a fixed length feature vector $\mathbf{x}_i$ ($i = 1, 2, ..., K$). We use the 4096 dimensional CNN feature implemented in Caffe [8] to extract the feature vector from each proposal. This feature has been proved to be effective for a wide variety of computer vision tasks.

Proposal attention: For each object proposal $\mathbf{x}_i$, we then compute an *attention score* s_i indicating how likely this object proposal contains the object of interest. This is achieved by applying a linear mapping on $\mathbf{x}_i$ followed by a softmax operation. Let $\mathbf{w}_a$ denote a vector of parameters for the linear mapping, the attention score s_i is calculated as:

$$g_i = \mathbf{w}_a^\top \mathbf{x}_i \tag{1a}$$

$$s_i = \frac{\exp(g_i)}{\sum_{j=1}^{K} \exp(g_j)}, \quad i = 1, 2, ..., K \tag{1b}$$

Without loss of generality and to simplify the notation, we use a linear mapping without the bias term in Eq. 1 by assuming that the feature vector $\mathbf{x}$ already has 1 appended to the end.

If we ignore the softmax operator in Eq. 1, the linear mapping in Eq. 1 alone can be loosely interpreted as a "detection score". In the ideal case, if we have access to fully supervised data where the ground-truth bounding boxes are provided, we can learn $\mathbf{w}_a$ directly using standard supervised learning. However, since we only have weakly supervised data where the labels are provided only at the whole image level, we can not learn $\mathbf{w}_a$ directly. Instead, the attention score in Eq. 1 simply provides an indication on how likely this object proposal contains an informative image region.

The softmax operator in Eq. 1 is a crucial part of our model. First of all, it introduces nonlinearity in the overall model. Second, it makes sure that the attention scores s_i ($i = 1, 2, ..., K$) of all the object proposals in an image sum to 1.

Image-level classification: Since our data are labeled only at the image-level, we need to use a learning method where the loss function is based on image-level labels. In our work, we use the attention scores to combine the object proposals to get an image-level feature vector $\mathbf{z}$ as $\mathbf{z} = \sum_{i=1}^{K} s_i \mathbf{x}_i$. This image-level feature $\mathbf{z}$ is then used to classify the whole image by a linear classifier with parameters $\mathbf{w}_c$:

$$f(\mathbf{x}; \{\mathbf{w}_a, \mathbf{w}_c\}) = \mathbf{w}_c^\top \mathbf{z} \tag{2}$$

where $f(\mathbf{x}; \{\mathbf{w}_a, \mathbf{w}_c\})$ is the score of classifying $\mathbf{z}$ to be a positive class.

3.2 Learning and localization

Our model has two sets of parameters $\{\mathbf{w}_a, \mathbf{w}_c\}$. Here we explain how to learn these parameters from weakly labeled data. Given a set of N training images $\{(\mathbf{x}^{(n)}, y^{(n)})\}$, where $\mathbf{x}^{(n)}$ represents the image and $y^{(n)} \in \{-1, +1\}$ represents the image-level label. We use the logistic loss as our loss function:

$$\ell(\{\mathbf{w}_a, \mathbf{w}_c\}) = \frac{1}{N} \sum_{n=1}^{N} \log\left(1 + \exp\left(-y^{(n)} f\left(\mathbf{x}^{(n)}; \{\mathbf{w}_a, \mathbf{w}_c\}\right)\right)\right) \tag{3}$$

This loss function is differentiable with respective to its parameters and can be optimized using the stochastic gradient descent. As a regularization, we add a dropout layer with a dropout rate of 0.8 before each of the linear layers in Fig. 2.

Once the learning is done, we localize the object in weakly labeled data directly using the attention scores. For example, suppose the object of interest is "dog". For each positive "dog" image, we simply choose the object proposal that has the highest attention score s_i as the localized dog instance in this image.

4 Experiments

We evaluate our method on one image dataset (Sec. 4.1) and two video datasets (Sec. 4.2 and Sec. 4.3). On the videos, we ignore the temporal information and treat frames in a video as images. We use the CorLoc measurement defined in [5] to evaluate the performance of the object localization results. For each image in the dataset, we compute the area of intersection over union (IoU) according to the PASCAL criterion $\frac{area(B_p \cap B_{gt})}{area(B_p \cup B_{gt})}$, where B_p is the localized bounding box and B_{gt} is the ground-truth bounding box. An image is considered as being correctly localized if the IoU is great than 0.5. Finally, CorLoc is computed as the percentage of images where the object of interest is correctly localized.

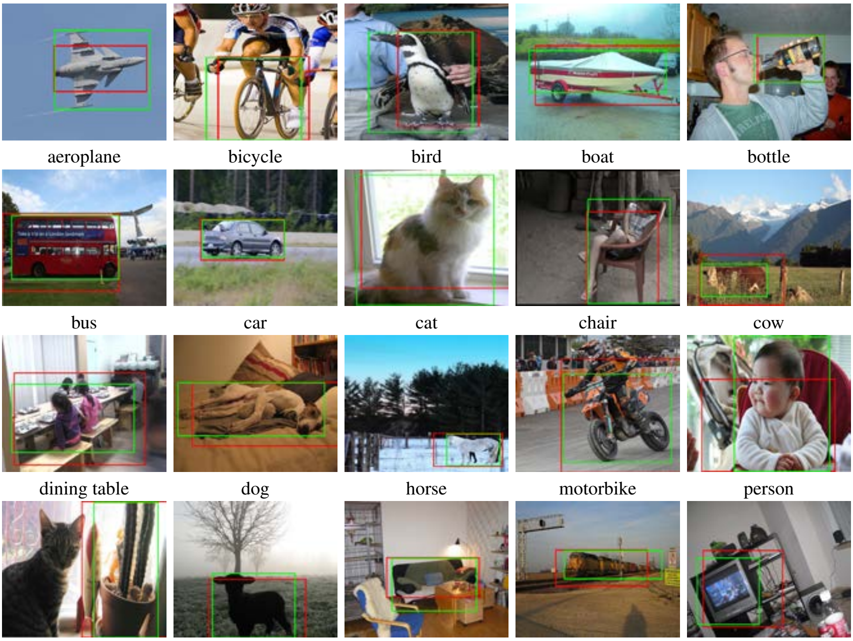
aeroplane bicycle bird boat bottle
bus car cat chair cow
dining table dog horse motorbike person
plant sheep sofa train tv monitor

Figure 3: 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.

4.1 PASCAL VOC 2007

The PASCAL VOC 2007 dataset [7] consists of images of 20 object classes. Similar to previous work, we use trainval subset of PASCAL VOC 2007 to train and evaluate our model. We train our model as a binary classifier for each class in the dataset. We use all the images corresponding to a class as positive examples. Then we randomly select an equal number of negative images that do not contain this object from the dataset.

Table 1 shows the CorLoc performance of our model on this dataset. We compare our method with the several state-of-the-art approaches [2, 16] on weakly supervised object localization. Our method outperforms other approaches. Fig. 3 shows qualitative examples of localization produced by our method on this dataset. Fig. 4 visualizes the attention scores on some examples.

method	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	mbike	person	plant	sheep	sofa	train	tv	avg
[24]	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
[2]	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
[16]	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
ours	**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**

Table 1: CorLoc results on *positive* trainval subset of the PASCAL VOC 2007 dataset. We also compare our approach with several state-of-the-art approaches [2, 16, 24].

The main goal of our work is to localize the objects on weakly labeled data. Once the objects can be reliably localized, the localization results can be used to train object detectors. Our model is not designed to directly work as a detector. In order to compare with previous work that reports detection results on this dataset, we use a simple strategy to obtain detec-

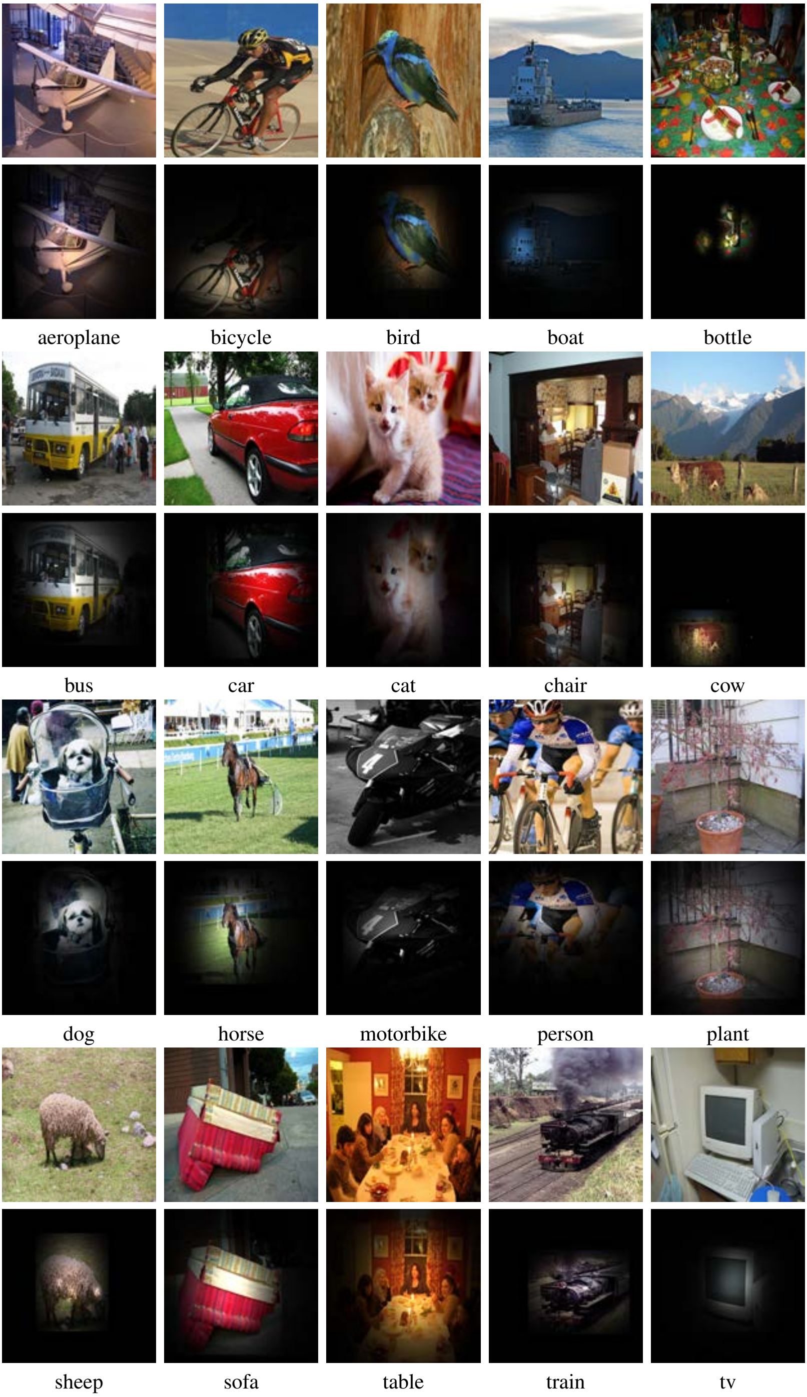

Figure 4: Visualization of attention scores on some sample images on PASCAL VOC 2007.

tion results. Suppose we want to detect instances of an object class (e.g. dog) in an unseen test image, we first run our dog model to obtain the attention scores for all the object proposals in this image. Then we consider the multiplication of softmaxed attention scores and the classification score as the detector scores for dogs and run non-maximum suppression to obtain the final detection results. Using this simple trick, we apply our model as object detectors on the PASCAL VOC 2007 test set. The results in terms of mean average precision are shown in Table 2. Even though our approach is not designed to be a detector, the performance of object detection is still comparable to other state-of-the-art methods. Also note that the best performance in [2] (5th row in Table 2) uses an ensemble of several CNN models, so it is difficult to directly compare with the results.

method	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	mbike	person	plant	sheep	sofa	train	tv	avg
[22]	**48.90**	42.30	26.10	11.30	11.90	41.30	40.90	34.70	**10.80**	34.70	18.80	34.40	35.40	52.70	19.10	**17.40**	35.90	33.30	34.80	46.5	31.60
[2] WSDNN S	42.90	56.00	32.00	17.60	10.20	61.80	50.20	29.00	3.80	36.20	18.50	31.10	45.80	54.50	10.20	15.40	36.30	45.20	50.10	43.80	34.50
[2] WSDDN M	43.60	50.40	32.20	26.00	9.80	58.50	50.40	30.90	7.90	36.10	18.20	31.70	41.40	52.60	8.80	14.00	37.80	46.90	53.40	47.90	34.90
[2] WSDDN L	39.40	50.10	31.50	16.30	12.60	64.50	42.80	42.60	10.10	35.70	24.90	38.20	34.40	55.60	9.40	14.70	30.20	40.70	54.70	46.90	34.80
[2] WSDDN Ensemble	46.40	**58.30**	35.50	25.90	**14.00**	**66.70**	**53.00**	39.20	8.90	**41.80**	26.60	38.60	44.70	**59.00**	10.80	17.30	**40.70**	**49.60**	**56.90**	**50.80**	**39.30**
ours	48.82	45.92	**37.43**	**26.92**	9.23	50.69	43.36	**43.62**	10.62	35.90	**27.02**	**38.62**	**48.51**	43.77	**24.71**	12.09	29.04	23.18	48.84	41.88	34.51

Table 2: Comparison of detection results (mAP%) on the PASCAL VOC 2007 test dataset.

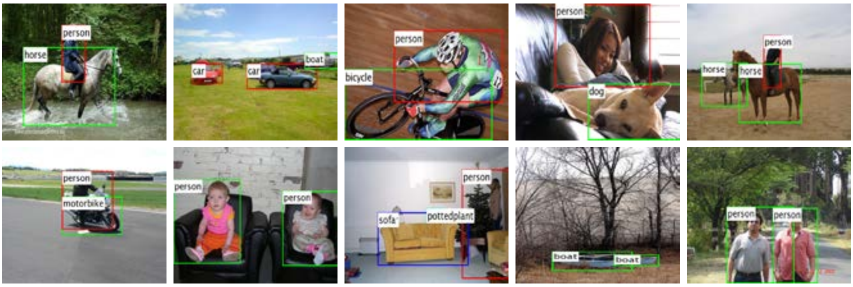

Figure 5: Qualitative examples of detection produced by our model on the PASCAL VOC 2007 test dataset.

4.2 YouTube-Objects

The YouTube-Objects dataset [15] contains videos of 10 different object classes. For each object class, ground-truth bounding box is available for one frame per shot. We train and evaluate our model on all the frames with the bounding box annotation.

Table 3 shows the results on this dataset. We compare the performance of our model with the other baseline approaches. Again, our approach outperforms other baseline methods. Note that the method in [16] exploits detectors for other related objects via transfer learning. Even though our approach does not use detectors for other objects, we still outperform the best reported performance (4th row in Table 3) in [16]. If we compare with the results in [16] that do not use external object detectors (3rd row in Table 3), the performance gain is even larger. Fig. 6 shows qualitative examples of localization on this dataset.

4.3 YouTube-Objects-Subset

Finally, we evaluate our model on the subset of the YouTube-Objects dataset [15] introduced by Tang *et al*. [22]. The advantage of this dataset is that it has more videos with ground-truth annotation. The ground-truth is available in the form of segmentation mask. For evaluation,

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

Table 3: CorLoc results on the YouTube-Objects dataset [15]. We compare with several state-of-the-art approaches on this dataset. Note that "[16] proposal + transfer" uses external object detectors in their framework, so it is not directly comparable to other methods in the table. See the text description for details.

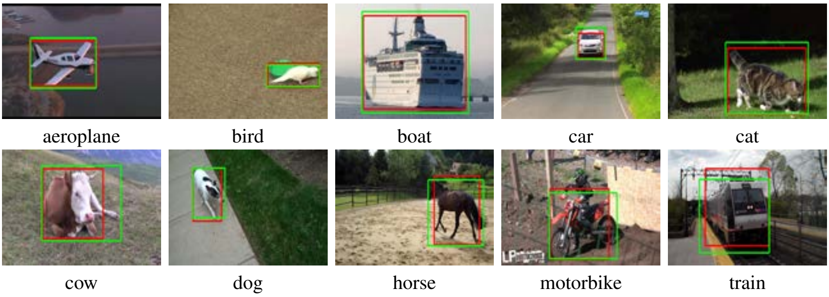
aeroplane bird boat car cat
cow dog horse motorbike train

Figure 6: Qualitative examples of our approach on the YouTube-Objects dataset. Ground-truth localization marked in Green whereas predicted localization is marked in Red.

we take the localized object bounding box and apply grabCut [17] to generate the segmentation mask for the object. Then, we compare the result using the ground-truth segmentation mask.

Table 4 shows the results on this dataset. Our model outperforms [16] even though the best performance of [16] uses external object detectors.

method	aero	bird	boat	car	cat	cow	dog	horse	bike	train	avg
[16] proposal only	42.23	51.24	29.54	67.76	14.75	50.20	47.02	22.18	16.44	18.84	36.02
[16] proposal + transfer	45.74	**55.47**	39.51	58.75	**26.51**	55.00	43.51	33.71	32.76	**25.63**	41.66
ours	**49.19**	45.52	**43.94**	**69.32**	26.43	**60.24**	**56.03**	**40.39**	**40.39**	19.91	**45.10**

Table 4: CorLoc results on the YouTube-Objects-Subset dataset.

5 Conclusion

We have proposed an attention network, a new network architecture for object localization from weakly labeled images. The novelty of our attention network is that it combines the attention mechanism in object localization from weakly labeled data. Our experimental results demonstrate that our attention network outperforms other state-of-the-art approaches on object localization.

There are several avenues for future work. First, we would like to combine the object proposals as part of the network, so the entire pipeline can be trained end-to-end. Second, our attention network is not designed for detection. Therefore, an interesting future work is to combine our work with [2] to make the model amenable to be directly used as a detector.

Acknowledgement

This work was supported by NSERC. We thank Neil Bruce for helpful discussions and NVIDIA for donating the GPUs used in this work.

References

[1] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. In *International Conference on Learning Representations*, 2015.

[2] Hakan Bilen and Andrea Vedaldi. Weakly supervised deep detection networks. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

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

[4] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2009.

[5] Thomas Deselaers, Bogdan Alexe, and Vittorio Ferrari. Weakly supervised localization and learning with generic knowledge. *International Journal of Computer Vision*, 100 (3):257–293, 2012.

[6] Thomas G. Dietterich, Richard H. Lathrop, and Tomas Lozano-Perez. Solving the multiple instance problem with axis-parallel rectangles. *Artificial Intelligence*, 1997.

[7] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL visual object classes (VOC) challenge. *International Journal of Computer Vision*, 88(2):303–338, 2010.

[8] Yangqing Jia, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long, Ross Girshick, Sergio Guadarrama, and Trevor Darrell. Caffe: Convolutional architecture for fast feature embedding. *arXiv:1408.5093*, 2014.

[9] Armand Joulin, Kevin Tang, and Li Fei-Fei. Efficient image and video co-localization with frank-wolfe algorithm. In *European Conference on Computer Vision*, 2014.

[10] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. ImageNet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, 2012.

[11] M. Pawan Kumar, Ben Packer, and Daphne Koller. Self-paced learning for latent variable models. In *Advances in Neural Information Processing Systems*, 2010.

[12] Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hayes, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. Microsoft COCO: Common objects in context. In *European Conference on Computer Vision*, 2014.

[13] Maxime Oquab, Leon Bottou, Ivan Laptev, and Josef Sivic. Is object localization for free? – weakly-supervised learning with convolutional neural networks. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2015.

[14] Anestis Papazoglou and Vittorio Ferrari. Fast object segmentation in unconstrained video. In *IEEE International Conference on Computer Vision*, 2013.

[15] Alessandro Prest, Christian Leistner, Javier Civera, Cordelia Schmid, and Vittorio Ferrari. Learning object class detectors from weakly annotated videos. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2012.

[16] Mrigank Rochan and Yang Wang. Weakly supervised localization of novel objects using appearance transfer. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2015.

[17] Carston Rother, Vladimir Kolmogorov, and Andrew Blake. Grabcut: Interactive foreground extraction using iterative graph cuts. In *SIGGRAPH*, 2004.

[18] Shikhar Sharma, Ryan Kiros, and Ruslan Salakhutdinov. Action recognition using visual attention. In *ICLR Workshop*, 2016.

[19] Kevin J. Shih, Saurabh Singh, and Derek Hoiem. Where to look: Focus regions for visual question answering. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

[20] Hyun Oh Song, Ross Girshick, Stefanie Jegelka, Julien Mairal, Zaid Harchaoui, and Trevor Darrell. On learning to localize objects with minimal supervision. In *International Confernece on Machine Learning*, 2014.

[21] 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.

[22] Kevin Tang, Rahul Sukthankar, Jay Yagnik, and Li Fei-Fei. Discriminative segment annotation in weakly labeled video. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2013.

[23] Kevin Tang, Armand Joulin, Li-Jia Li, and Li Fei-Fei. Co-localization in real-world images. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2014.

[24] Chong Wang, Weiqiang Ren, Kaiqi Huang, and Tieniu Tan. Weakly supervised object localization with latent category learning. In *European Conference on Computer Vision*, 2014.

[25] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhutdinov, Richard Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *International Conference on Machine Learning*, 2015.

[26] C. Lawrence Zitnick and Piotr Dollar. Edge boxes: Locating object proposals from edges. In *European Conference on Computer Vision*, 2014.