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

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.

Our proposed method is loosely inspired by the binding problem in human visual systems [2]. It is believed that human brains involve two stages when processing visual information. First, human brains preconsciously segregate visual input into disparate brain regions. Then the brains combine the results of many sensory operations to create coherent visual experience. During this whole process, the ability to selectively focus on certain spatial regions (i.e. *attention*) is crucial for human visual systems.

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 overview of our approach is illustrated in Fig. 1.

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 [3], 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 [1] 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:

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

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

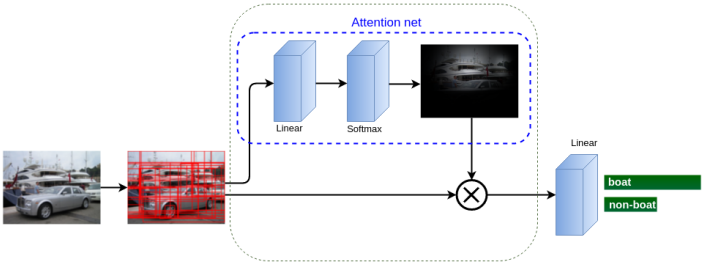

Figure 1: An overview of our architecture. Given an image, we extract proposals that are likely to contain *any* object. Each proposal is passed to a linear layer to obtain its attention score. We then apply the softmax operation to the attention scores before multiplying it with its corresponding proposal 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.

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.

[1] 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.

[2] Hubert Zimmer, Axel Mecklinger, and Ulman Lindenberger. Handbook of binding and memory: Perspectives from cognitive neuroscience. 2006.

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