Weakly Supervised Cascaded Convolutional Networks

Ali Diba[1], Vivek Sharma[2,⋆], Ali Pazandeh[3], Hamed Pirsiavash[4] and Luc Van Gool[1,5]
[1]ESAT-PSI, KU Leuven, [2]CV:HCI, Karlsruhe Institute of Technology
[3]Sharif University, [4]University of Maryland Baltimore County, [5]CVL, ETH Zürich
ali.diba@kuleuven.be, vivek.sharma@kit.edu, pazandeh@ee.sharif.edu, hpirsiav@umbc.edu

Abstract

Object detection is a challenging task in visual understanding domain, and even more so if the supervision is to be weak. Recently, few efforts to handle the task without expensive human annotations is established by promising deep neural network. A new architecture of cascaded networks is proposed to learn a convolutional neural network (CNN) under such conditions. We introduce two such architectures, with either two cascade stages or three which are trained in an end-to-end pipeline. The first stage of both architectures extracts best candidate of class specific region proposals by training a fully convolutional network. In the case of the three stage architecture, the middle stage provides object segmentation, using the output of the activation maps of first stage. The final stage of both architectures is a part of a convolutional neural network that performs multiple instance learning on proposals extracted in the previous stage(s). Our experiments on the PASCAL VOC 2007, 2010, 2012 and large scale object datasets, ILSVRC 2013, 2014 datasets show improvements in the areas of weakly-supervised object detection, classification and localization.

1. Introduction

The ability to train a system that detects objects in cluttered scenes by only naming the objects in the training images, without specifying their number or their bounding boxes, is understood to be of major importance. Then it becomes possible to annotate very large datasets or to automatically collect them from the web.

Most current methods to train object detection systems assume strong supervision [12, 26, 19]. Providing both the bounding boxes and their labels as annotations for each object, still renders such methods more powerful than their weakly supervised counterparts. Although the availability of larger sets of training data is advantageous for the training of convolutional neural networks (CNNs), weak supervision as a means of producing those has only been embraced to a limited degree.

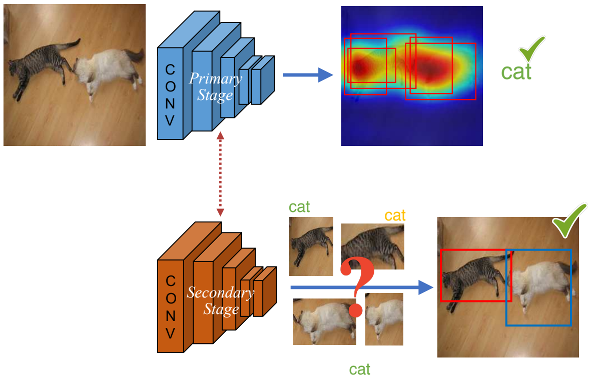

Figure 1. **Weakly Supervised Cascaded Deep CNN:** Overview of the proposed cascaded weakly supervised object detection and classification method. Our cascaded networks take images and existing object labels to find the best location of objects samples in each of images. Trained networks based on these location is capable of detecting and classifying objects in images, under weakly supervision circumstances.

The proposed weak supervision methods have come in some different flavors. One of the most common approaches [7] consists of the following steps. The first step generates object proposals. The last stage extracts features from the proposals. And the final stage applies multiple instance learning (MIL) to the features and finds the box labels from the weak bag (image) labels. This approach can thus be improved by enhancing any of its steps. For instance, it would be advantageous if the first stage were to produce more reliable - and therefore fewer - object proposals.

It is the aforementioned approach that our weak supervision algorithm also follows. To improve the detection performance, object proposal generation, feature extraction, and MIL are trained in a cascaded manner, in an end-to-end way. We propose two architectures. The first is a two stage network. The first stage extracts class specific object proposals using a fully convolutional network followed by a

[⋆]This work was carried out while he was at ESAT-PSI, KU Leuven.

global average (max) pooling layer. The last stage extracts features from the object proposals by a ROI pooling layer and performs MIL. Given the importance of getting better object proposals we added a middle stage to the previous architecture in our three stage network. This middle stage performs a class specific segmentation using the input images and the extracted objectness of the first stage. This results in more reliable object proposals and a better detection.

The proposed architecture improves both initial object proposal extraction and final object detection. In the forward sense, less noisy proposals indeed lead to improved object detection, due to the non-convexity of the cost function. In the reverse, backward sense, due the weight sharing between the first layers of both stages, training the MIL on the extracted proposals will improve the performance of feature extraction in the first convolutional layers and as a result will produce more reliable proposals.

Next, we review related works in section 2 and discuss our proposed method in section 3. In section 4 we explain the details of our experiments, including the dataset and complete set of experiments and results.

2. Related works

Weakly supervised detection: In the last decade, several weakly supervised object detection methods have been studied using multiple instance learning algorithms [4, 5, 29, 30]. To do so they define images as the bag of regions, wherein they assume the image labeled positive contains at least one object instance of a certain category and an image labeled negative do not contain an object from the category of interest. The most common way of weakly supervised learning methods often work by selecting the candidate positive object instances in the positive bags, and then learning a model of the object appearance using appearance model. Due to the training phase of the MIL problem alternating between out of bag object extraction and training classifiers, the solutions are non-convex and as a result is sensitive to the initialization. In practice, a bad initialization is prone to getting the solution stuck in a local optima, instead of global optima. To alleviate this shortcoming, several methods try to improve the initialization [31, 9, 28, 29] as the solution strongly depends on the initialization, while some others focus on regularizing the optimization strategies [4, 5, 7]. Kumar et al. [17] employ an iterative self-learning strategy to employ harder samples to a small set of initial samples at training stage. Joulin et al. [15] use a convex relaxation of soft-max loss in order to minimize the prone to get stuck in the local minima. Deselaers et al. [9] initialize the object locations via the objectness score. Cinbis et al. [7] split the training date in a multi-fold manner for escaping from getting trapped into the local minima. In order to have more robustness from poor initialization, Song et al. [30] apply Nesterov's smoothing technique to latent SVM formulation [10]. In [31], the same authors initialize the object locations based on sub-modular clustering method. Bilen et al. [4] formulates the MIL to softly label the object instances by regularizing the latent object locations based on penalizing unlikely configurations. Further in [5], the authors extend their work [4] by enforcing similarity between object windows via regularization technique. Wang et al. [35] employ probabilistic latent semantic analysis on the windows of positive samples to select the most discriminative clusters that represents the object category. As a matter of fact, majority of the previous works [25, 32] use a large collection of noisy object proposals to train their object detector. In contrast, our method only focuses on a very few clean collection of object proposals that are far more reliable, robust, computationally efficient, and gives better performance.

Object proposal generation: In [20, 23], Nguyen et al. and Pandey et al. extract dense regions of candidate proposals from an image using an initial bounding box. To handle the problem of not being able to generate enough candidate proposals because of fixed shape and size, object saliency [9, 28, 29] based approaches were proposed to extract region proposals. Following this, generic objectness measure [1] was employed to extract region proposals. Selective search algorithm [33], a segmentation based object proposal generation was proposed, which is currently among the most promising techniques used for proposal generation. Recently, Ghodrati et al. [11] proposed an inverse cascade method using various CNN feature maps to localize object proposals in a coarse to fine manner.

CNN based weakly supervised object detection: In view of the promising results of CNNs for visual recognition, some recent efforts in weakly supervised classification have been based on CNNs. Oquab et al. [21] improved feature discrimination based on a pre-trained CNN. In [22], the same authors improved the performance further by incorporating both localization and classification on a new CNN architecture. Bilen et al. [4] proposed a CNN-based convex optimization method to solve the problem to escape from getting stuck in local minima. Their soft similarity between possible regions and clusters was helpful in improving the optimization. Li et al. [18] introduced a class-specific object proposal generation based on the mask out strategy of [2], in order to have a reliable initialization. They also proposed their two-stage algorithm, classification adaptation and detection adaptation.

3. Proposed Method

This section introduces our weak cascaded convolutional networks (WCCN) for object detection and classification with weak supervision. Our networks are designed to learn multiple different but related tasks all together jointly. The

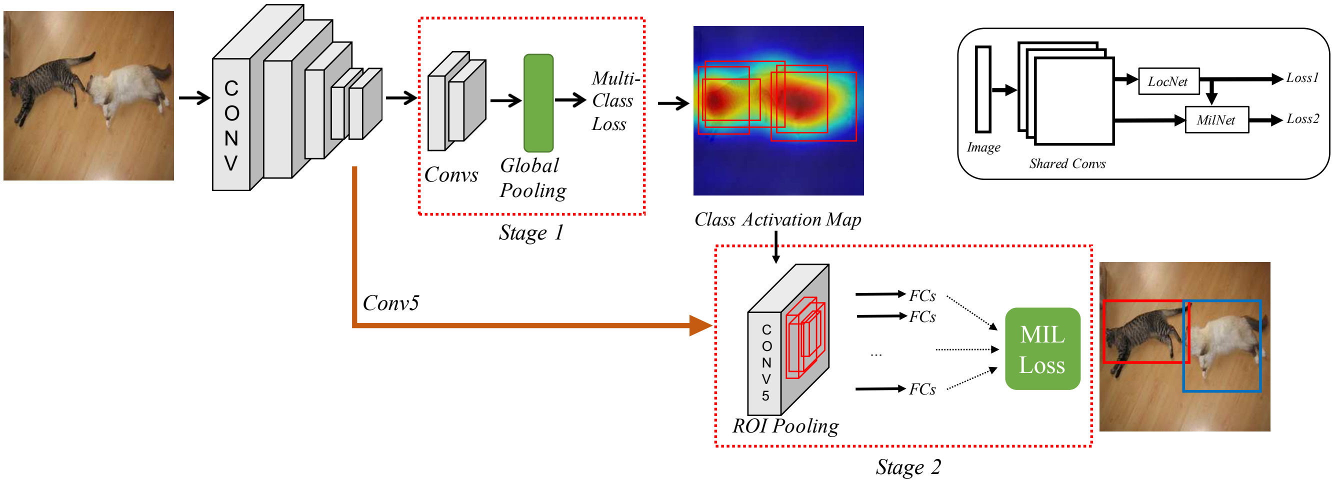

Figure 2. **WCCN (2stage):** The pipeline of end-to-end 2-stage cascaded CNN for weakly supervised object detection. Inputs to the network are images, labels and unsupervised object proposals. First stage learns to create a class activation map based on object categories to make some candidate boxes for each instance of objects. Second stage picks the best bounding box among the candidates to represent the specific category by multiple instance learning loss.

tasks are classification, localization, and multiple instance learning. We show that learning these tasks jointly in an end-to-end fashion results in better object detection and localization. The goal is to learn good appearance models from images with multiple objects where the only manual supervision signal is image-level labels. Our main contribution is improving multiple object detection with such weak annotation. To this end, we propose two different cascaded network architectures. The first one is a 2-stage cascade network that first localizes the objects and then learns to detect them in a multiple instance learning framework. Our second architecture is a 3-stage cascade network where the new middle stage performs semantic segmentation with pseudo ground truth in a weakly supervised setting.

3.1. Two-stage Cascade

As mentioned earlier, there are only a few end-to-end frameworks with deep CNNs for weakly supervised object detection. In particular, there is not much prior art on object localization without supervising in localization level. Suppose we have dataset $\mathcal{I}$ of N training images in C classes. The set is given as $\mathcal{I} = \{(I^1, \mathbf{y}^1), ..., (I^N, \mathbf{y}^N)\}$ where I^k is an image and $\mathbf{y}^k = [y_1, ..., y_C] \in \{0,1\}^C$ is a vector of labels indicating the presence or absence of each class in image I^k.

In the proposed cascaded network, the initial fully-convolutional stage learns to infer object location maps based on the object labels in the given images. This stage produces some candidate boxes of objects as input to the next stage. The last stage selects the best boxes through an end-to-end multiple instance learning.

First stage (Location network): The first stage of our cascaded model is a fully-convolutional CNN with a global average pooling (GAP) or global maximum pooling (GMP) layer, inspired by [36]. The training yields the object location or 'class activation' maps, that provide candidate bounding boxes. Since multiple categories can exist in a single image [22], we use an independent loss function for each class in this branch of the CNN architecture, so the loss function is the sum of C binary logistic regression loss functions.

Last stage (MIL network): The goal of the last stage is to select the best candidate boxes for each class from the outputs of the first stage using multiple instance learning (MIL). To obtain an end-to-end framework, we incorporate an MIL loss function into our network. Assume $\mathbf{x} = \{x^j | j = 1, 2, ..., n\}$ is a bag of instances for image I where x^j is a candidate box, and assume $f_{cj} \in \Re^{C \times n}$ is the score of box x^j belonging to category i. We use ROI-pooling layer [12] to achieve f_{cj}. We define the probabilities and loss as:

$$\begin{gathered} P_c(\mathbf{x}, I) = \frac{exp(\max_j f_{cj})}{\sum_{k=1}^{C} exp(\max_j f_{kj})} \\ L_{MIL}(\mathbf{y}, \mathbf{x}, I) = -\sum_{c=1}^{C} y_c log(P_c(\mathbf{x}, I)) \end{gathered} \qquad (1)$$

The weights for *conv1* till *conv5* are shared between the two stages. For the last stage, we have additional two fully connected layers and a score layer for learning the MIL task.

End-to-End Training: The whole cascade with two loss functions is learned jointly by end-to-end stochastic gradient descent optimization. The total loss function of the cas-