

Weakly Supervised Cascaded Convolutional Networks

Ali Diba¹, Vivek Sharma^{2,*}, Ali Pazandeh³, Hamed Pirsiavash⁴ and Luc Van Gool^{1,5}

¹ESAT-PSI, KU Leuven, ²CV:HCI, Karlsruhe Institute of Technology

³Sharif University, ⁴University of Maryland Baltimore County, ⁵CVL, 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 super-

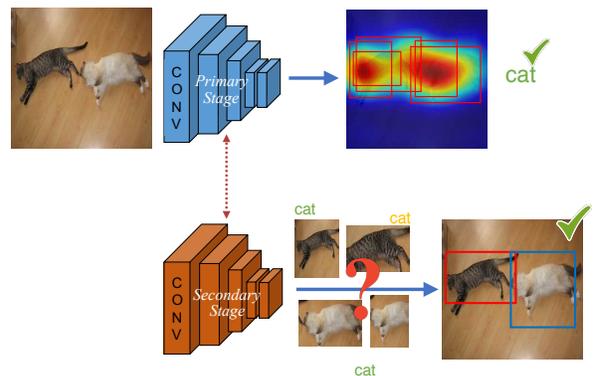


Figure 1. **Weakly Supervised Cascaded Deep CNN:** Overview of the proposed cascaded weakly supervised object detection and classification method. Our cascaded networks take images and existing object labels to find the best location of objects samples in each of images. Trained networks based on these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

vision as a means of producing those has only been embraced to a limited degree.

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 data 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

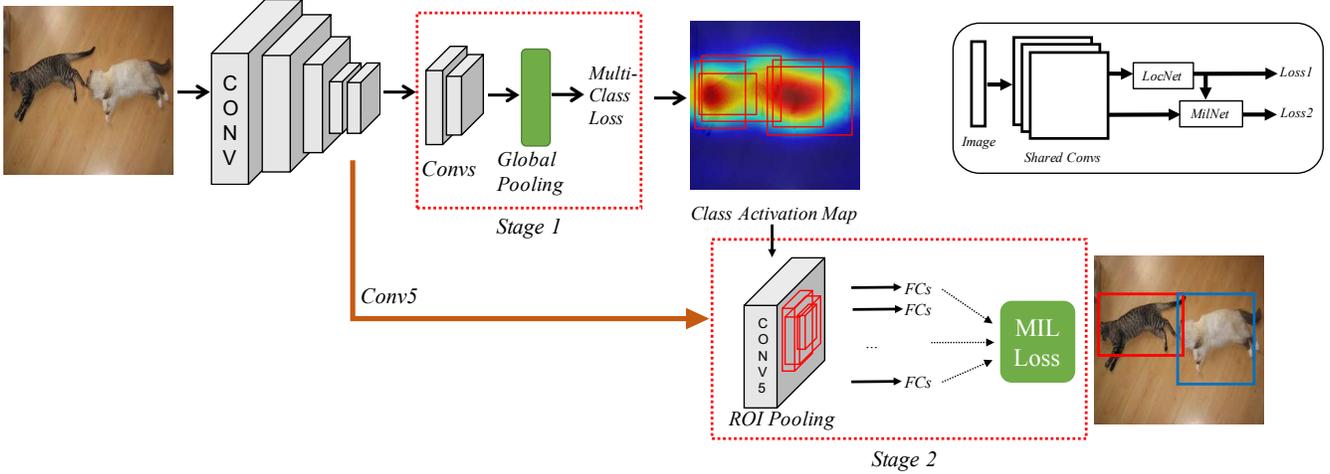


Figure 2. **WCCN (2stage)**: The pipeline of end-to-end 2-stage cascaded CNN for weakly supervised object detection. Inputs to the network are images, labels and unsupervised object proposals. First stage learns to create a class activation map based on object categories to make some candidate boxes 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 $\mathcal{I} = \{(I^1, \mathbf{y}^1), \dots, (I^N, \mathbf{y}^N)\}$ where I^k is an image and $\mathbf{y}^k = [y_1, \dots, y_C] \in \{0, 1\}^C$ 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 $\mathbf{x} = \{x^j | j = 1, 2, \dots, n\}$ is a bag of instances for image I where x^j is a candidate box, and assume $f_{cj} \in \mathbb{R}^{C \times n}$ 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:

$$P_c(\mathbf{x}, I) = \frac{\exp(\max_j f_{cj})}{\sum_{k=1}^C \exp(\max_j f_{kj})} \quad (1)$$

$$L_{MIL}(\mathbf{y}, \mathbf{x}, I) = - \sum_{c=1}^C y_c \log(P_c(\mathbf{x}, I))$$

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-

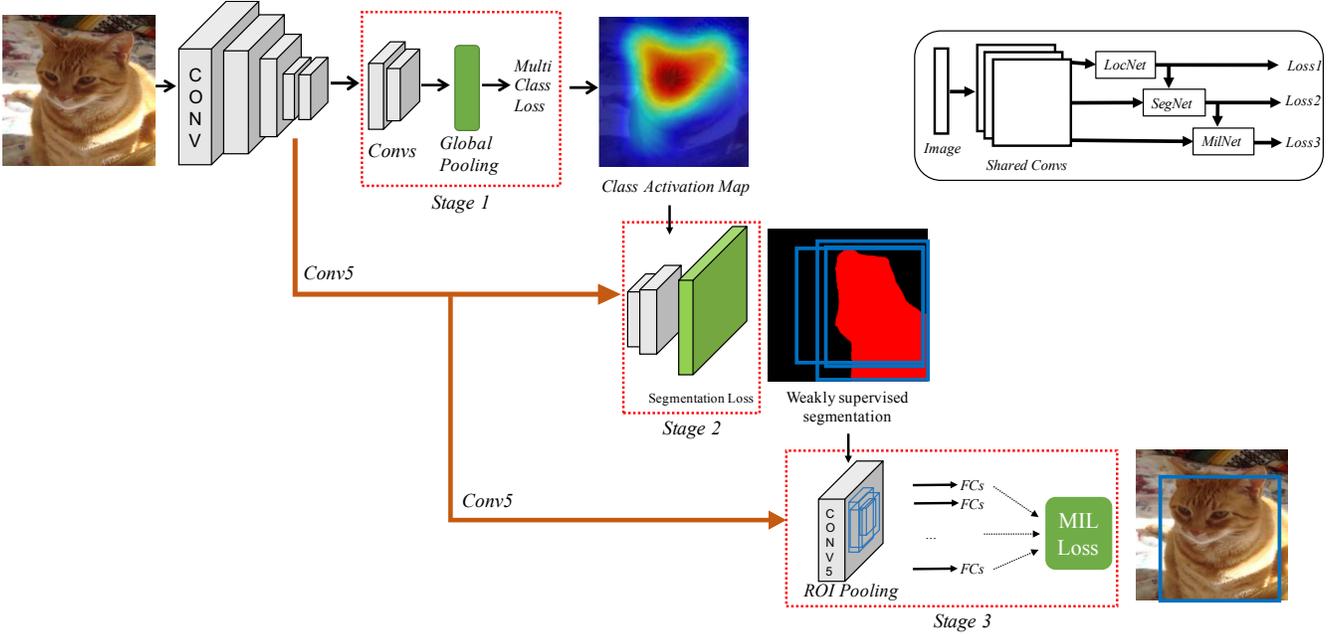


Figure 3. **WCCN (3stage)**: The pipeline of end-to-end 3-stage cascaded CNN for weakly supervised object detection. For this cascaded network, we designed new architecture to have weakly supervised segmentation as last stage, so first and last stages are identical to the stages of the previous cascade. The new stage will improve the selecting candidate bounding boxes by providing more accurate object regions.

caded network is:

$$L_{Total} = L_{GAP}(y, I) + \lambda L_{MIL}(y, x, I) \quad (2)$$

where λ is the hyper-parameter balancing two loss functions. In the experiments, we set $\lambda = 1$. We suspect cross-validation on this hyper-parameter can improve the results.

Generating bag of instances: We use Edgeboxs [37] to generate an initial set of object proposals. Then we threshold the class activation map [36] to come up with a mask. Finally, we choose the initial boxes with largest overlap with the mask.

3.2. Three-stage Cascade

In this section, we extend our 2-stage cascaded model by another stage that adds object segmentation as another task. We believe more information about the objects' boundary learned in a segmentation task can lead to acquisition of a better appearance model and then better object localization. For this purpose, our new stage uses another form of weak supervision to learn a segmentation model, embedded in the cascaded network and trained along with other stages. This extra stage will help the multi-loss CNN to have better initial locations for choosing candidate bounding boxes to pass to the next stage. So this new cascade has three stages: **first stage**, similar to previous cascade is a CNN with global pooling layer; **middle stage**, fully convolutional network

with segmentation loss; **last stage**, multiple instance learning with corresponding loss.

Middle stage (Segmentation Loss): Inspired by [3, 24], we propose to use a weakly supervised segmentation network which uses an object point of location and also label as supervisory signals. Incorporation of initial location of object from previous stage (location network) in the segmentation stage can obtain more meaningful object location map. The weak segmentation network uses the results of the first stage as supervision signal (i.e., pseudo ground truth) and learns jointly with the MIL stage to further improve the object localization results.

In the middle stage, we add a fully convolutional CNN similar to the one in [3] to our network. The final layer is a pixel-wise softmax that outputs $S \in \mathcal{R}^{C \times m}$ where m is the number of pixels in the image. Assuming H_c for the heatmap for class c , we define $\alpha_c = \max(H_c)$ across the whole image and I_c to be the neighborhood around $\text{argmax}(H_c)$. In the experiments, we use a neighborhood of 3×3 pixels. Note that our formulation is closely following the one in [3] except that our point-wise annotation is provided by the automatically generated heatmap rather than manual annotation.

Considering y as the label set for image I , the loss function for the weakly supervised segmentation network

is given by:

$$L_{Seg}(S, H, y) = - \sum_{c=1}^C y_c \left(\log(S_{t_c}) + \sum_{i \in I_c} \alpha_c \log(S_{i_c}) \right) \quad (3)$$

where $t_c = \underset{i \in I}{\operatorname{argmax}} S_{i_c}$. The first term is used for image-level label supervision and the second term is for the set of pixels that the heatmap confidently predicted to be a point on the object. Note that α_c is the second term is emphasizing on more confident categories.

Due to more supervision using pseudo-groundtruth provided by the heatmap, the middle stage provides a better segmentation map compared to the original heatmap. Hence, we pass the resulting segmentation map to the final MIL stage to find candidate boxes with overlapping and then calculate the MIL loss.

Output of this middle stage is a set of candidate bounding boxes of objects for pushing to next stage of the CNN cascade which uses multiple instance learning to choose the most accurate box as the representative of object category. In the experiments, we show that learning this extra task as another stage of cascade can improve performance of the whole network as a weakly supervised classifier.

End-to-End Training: Similar to the last cascade, the total loss in Eq.4 is calculated by simply adding all three loss terms. We learn all parameters of the network jointly in an end-to-end fashion.

$$L_{Total} = L_{GAP}(\mathbf{y}, I) + \gamma L_{Seg}(\mathbf{y}, I) + \lambda L_{MIL}(\mathbf{y}, \mathbf{x}, I) \quad (4)$$

In the experiments, we set $\lambda = 1$ and $\gamma = 1$.

3.3. Object Detection Training

Since we are interested in weakly supervised object detection, we propose to use the output of our network as pseudo-groundtruth in a standard object detection framework e.g., Fast-RCNN [12]. There are two ways of doing this: we can either train a standard Fast-RCNN without our trained model or we can transfer our learned model into the Fast-RCNN framework and finetune it. For the later case, we use the shared early convolutional layers along with the fully connected layers in the last stage of our model. In both cases, at the testing time, we extract object proposals with EdgeBoxes [37], use the trained Fast-RCNN to detect objects among the pool of proposals, and perform non-max-suppression.

4. Experiments

In the following section, we discuss details of our methods and experiments which we applied on object detection

and classification in weakly supervised manner. We introduce datasets and also analyze performance of our approaches on them in different aspects of evaluation.

4.1. Datasets and metrics

The experiments for our proposed methods are extensively done on the PASCAL VOC 2007, 2010, 2012 datasets and also ILSVRC 2013, 2014 which are large scale datasets for objects. The PASCAL VOC is more common dataset to evaluate weakly supervised object detection approaches. The VOC datasets have 20 categories of objects, while ILSVRC dataset has 200 categories which we targeted also for weakly supervised object classification and localization. In all of the mentioned datasets, we incorporate the standard train, validation and test set.

Experimental metrics: To measure the object detection performance, average precision (AP) and correct localization (CorLoc) is used. Average precision is the standard metric from PASCAL VOC which takes a bounding box as a true detection where it has intersection-over-union (IoU) of more than 50% with ground-truth box. The *Corloc* is the fraction of positive images that the method obtained correct location by most confident detection box for at least one object instance per target category in an image. For the object classification, also we use PASCAL VOC standard average precision.

4.2. Experimental and implementation details

We have evaluated both of our proposed cascaded CNN with two architectures: Alexnet [16] and VGG-16 [27]. In each case, the network has been pre-trained on ImageNet dataset [8]. Since the multiple stages of cascades contain different CNN networks losses, in the following we explain details of each part separately to have better overview of the implementation.

CNN architectures:

1. Loc Net: Inspired by [36], we removed fully-connected layers from each of Alexnet or VGG-16 and replaced them by two convolutional layers and one global pooling layer. So for the Alexnet, the layers after *conv5* layer have been removed and for VGG-16 after *conv5-3*. For global pooling layer, we have tested average and max pooling methods and we found that global average pooling performs better than maximum pooling. For the training loss criteria of this part of network, we use a simple sum of C (number of classes) binary logistic regression losses, similar to [22].

2. Seg Net: This part of network is middle stage in the 3-stage cascaded network and is well-known fully convolutional network for segmentation task [3]. The convolutional part is shared with the other stages which comes from the first stage and additional fully-connected layers and a deconvolutional layer is used to produce segmentation map.

Method	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	mbike	person	plant	sheep	sofa	train	tv	mAP
Bilen et al. [4]	42.2	43.9	23.1	9.2	12.5	44.9	45.1	24.9	8.3	24.0	13.9	18.6	31.6	43.6	7.6	20.9	26.6	20.6	35.9	29.6	26.4
Bilen et al. [5]	46.2	46.9	24.1	16.4	12.2	42.2	47.1	35.2	7.8	28.3	12.7	21.5	30.1	42.4	7.8	20.0	26.8	20.8	35.8	29.6	27.7
Cinbis et al. [7]	39.3	43.0	28.8	20.4	8.0	45.5	47.9	22.1	8.4	33.5	23.6	29.2	38.5	47.9	20.3	20.0	35.8	30.8	41.0	20.1	30.2
Wang et al. [35]	48.8	41.0	23.6	12.1	11.1	42.7	40.9	35.5	11.1	36.6	18.4	35.3	34.8	51.3	17.2	17.4	26.8	32.8	35.1	45.6	30.9
Li et al., Alexnet [18]	49.7	33.6	30.8	19.9	13	40.5	54.3	37.4	14.8	39.8	9.4	28.8	38.1	49.8	14.5	24.0	27.1	12.1	42.3	39.7	31.0
Li et al., VGG16 [18]	54.5	47.4	41.3	20.8	17.7	51.9	63.5	46.1	21.8	57.1	22.1	34.4	50.5	61.8	16.2	29.9	40.7	15.9	55.3	40.2	39.5
WSDDN [6]	46.4	58.3	35.5	25.9	14.0	66.7	53.0	39.2	8.9	41.8	26.6	38.6	44.7	59.0	10.8	17.3	40.7	49.6	56.9	50.8	39.3
WCCN_2stage_Alexnet	43.5	56.8	34.1	19.2	13.4	63.1	51.5	33.1	5.8	39.3	19.6	32.9	46.2	56.1	11.2	17.5	38.5	45.7	52.6	43.3	36.2
WCCN_2stage_VGG16	48.2	58.9	37.3	27.8	15.3	69.8	55.2	41.1	10.1	42.7	28.6	40.4	47.3	62.3	12.9	21.2	44.3	52.2	59.1	53.1	41.4
WCCN_3stage_Alexnet	43.9	57.6	34.9	21.3	14.7	64.7	52.8	34.2	6.5	41.2	20.5	33.8	47.6	56.8	12.7	18.8	39.6	46.9	52.9	45.1	37.3
WCCN_3stage_VGG16	49.5	60.6	38.6	29.2	16.2	70.8	56.9	42.5	10.9	44.1	29.9	42.2	47.9	64.1	13.8	23.5	45.9	54.1	60.8	54.5	42.8

Table 1. Detection average precision (%) on the PASCAL VOC 2007 dataset test set.

Method	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	mbike	person	plant	sheep	sofa	train	tv	mAP
WSDDN [6]	95.0	92.6	91.2	90.4	79.0	89.2	92.8	92.4	78.5	90.5	80.4	95.1	91.6	92.5	94.7	82.2	89.9	80.3	93.1	89.1	89.0
Oquab et al. [21]	88.5	81.5	87.9	82.0	47.5	75.5	90.1	87.2	61.6	75.7	67.3	85.5	83.5	80.0	95.6	60.8	76.8	58.0	90.4	77.9	77.7
SPNet [13]	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	82.4
Alexnet [6]	95.3	90.4	92.5	89.6	54.4	81.9	91.5	91.9	64.1	76.3	74.9	89.7	92.2	86.9	95.2	60.7	82.9	68.0	95.5	74.4	82.4
VGG16-net [27]	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	89.3
WCCN_2stage_Alexnet	92.8	90.3	89.3	88.2	80.4	89.4	90	90.4	75.3	88.1	80.1	91.3	89.1	88.3	91.2	80.6	88.5	77.8	92.2	88.7	87.1
WCCN_2stage_VGG16	93.4	93.7	92	91	83.1	91.5	92.7	93.5	79.3	90.7	83.1	96.9	92.9	91.2	95.9	82.4	90.3	81.3	95.1	88.3	89.9
WCCN_3stage_Alexnet	93.1	91.1	89.6	88.9	81	89.6	90.7	91.2	76.4	89.2	80.8	92.2	90.1	89	92.7	82	89.3	78.1	92.8	89.1	87.8
WCCN_3stage_VGG16	94.2	94.8	92.8	91.7	84.1	93	93.5	93.9	80.7	91.9	85.3	97.5	93.4	92.6	96.1	84.2	91.1	83.3	95.5	89.6	90.9

Table 2. Classification average precision (%) on the PASCAL VOC 2007 test set.

Method	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	mbike	person	plant	sheep	sofa	train	tv	mAP
Bilen et al. [5]	66.4	59.3	42.7	20.4	21.3	63.4	74.3	59.6	21.1	58.2	14.0	38.5	49.5	60.0	19.8	39.2	41.7	30.1	50.2	44.1	43.7
Cinbis et al. [7]	65.3	55.0	52.4	48.3	18.2	66.4	77.8	35.6	26.5	67.0	46.9	48.4	70.5	69.1	35.2	35.2	69.6	43.4	64.6	43.7	52.0
Wang et al. [35]	80.1	63.9	51.5	14.9	21.0	55.7	74.2	43.5	26.2	53.4	16.3	56.7	58.3	69.5	14.1	38.3	58.8	47.2	49.1	60.9	48.5
Li et al., Alexnet [18]	77.3	62.6	53.3	41.4	28.7	58.6	76.2	61.1	24.5	59.6	18.0	49.9	56.8	71.4	20.9	44.5	59.4	22.3	60.9	48.8	49.8
Li et al., VGG16 [18]	78.2	67.1	61.8	38.1	36.1	61.8	78.8	55.2	28.5	68.8	18.5	49.2	64.1	73.5	21.4	47.4	64.6	22.3	60.9	52.3	52.4
WSDDN [6]	65.1	63.4	59.7	45.9	38.5	69.4	77.0	50.7	30.1	68.8	34.0	37.3	61.0	82.9	25.1	42.9	79.2	59.4	68.2	64.1	56.1
WCCN_2stage_Alexnet	78.4	66.4	58.2	38.1	34.9	60.1	77.8	53.8	26.6	66.5	18.7	47.3	62.8	73.5	20.4	45.2	64	21.6	59.9	51.6	51.3
WCCN_2stage_VGG16	81.2	70	62.5	41.7	38.2	63.4	81.1	57.7	30.4	70.3	21.7	51	65.9	75.7	23.9	47.9	67.5	25.6	62.4	53.9	54.6
WCCN_3stage_Alexnet	79.7	68.1	60.4	38.9	36.8	61.1	78.6	56.7	27.8	67.7	20.3	48.1	63.9	75.1	21.5	46.9	64.8	23.4	60.2	52.4	52.6
WCCN_3stage_VGG16	83.9	72.8	64.5	44.1	40.1	65.7	82.5	58.9	33.7	72.5	25.6	53.7	67.4	77.4	26.8	49.1	68.1	27.9	64.5	55.7	56.7

Table 3. Correct localization (%) on PASCAL VOC 2007 on positive (CorLoc) trainval set.

The loss function is explained in section 3. Since this loss is provided by weak supervision, part of the supervision is obtained from the last stage in form of best initial regions of object instances.

3. MIL Net: This last stage uses the shared convolutional feature maps as initial layers to train two fully-connected layers with size of 4096 and a label prediction layer. Using the the selected candidate bounding boxes from previous stage, it trains the multiple instance learning loss to select the best sample for each object presented in an image.

Implementation details: We use MatConvNet [34] as CNN toolkit and all the networks are trained on one Titan X GPU. During the training time, images have been re-sized to multiple scale of images ($\{480, 576, 688, 84, 1200\}$) with respect to the original aspect ratio. The learning rate for the CNN networks is 0.0001 for 20 epochs and batch size of 100. For each image, we use 2000 object proposals generated by EdgeBox or SelectiveSearch algorithms. At the last stage, we select 10 boxes for each object instance in

each iteration for training multiple instance learning. To use Fast-RCNN detection with the ground-truths that are obtained by our methods, we set the number of iterations to 40K. For selecting the candidate boxes in our pipelines, we use a thresholding method like [36] for weakly localization.

4.3. Detection performance

Comparison with the state-of-the-art: We evaluate the detection performance of our method in this section. To compare our approach, methods which use deep learning pipelines [6, 18] or multiple instance learning algorithms [7] or clustering based approaches [5] are studied.

Tables 1, 4, 5 present results on PASCAL VOC 2007, 2010, 2012 for object detection on test sets with average precision measurement. It can be observed that by using the weakly supervision setup, we achieved the best performance among of all other recent methods. Our approaches do not incorporate any sophisticated clustering or optimized initialization step, and all the steps are trained together via an end-to-end learning of deep neural networks. There is

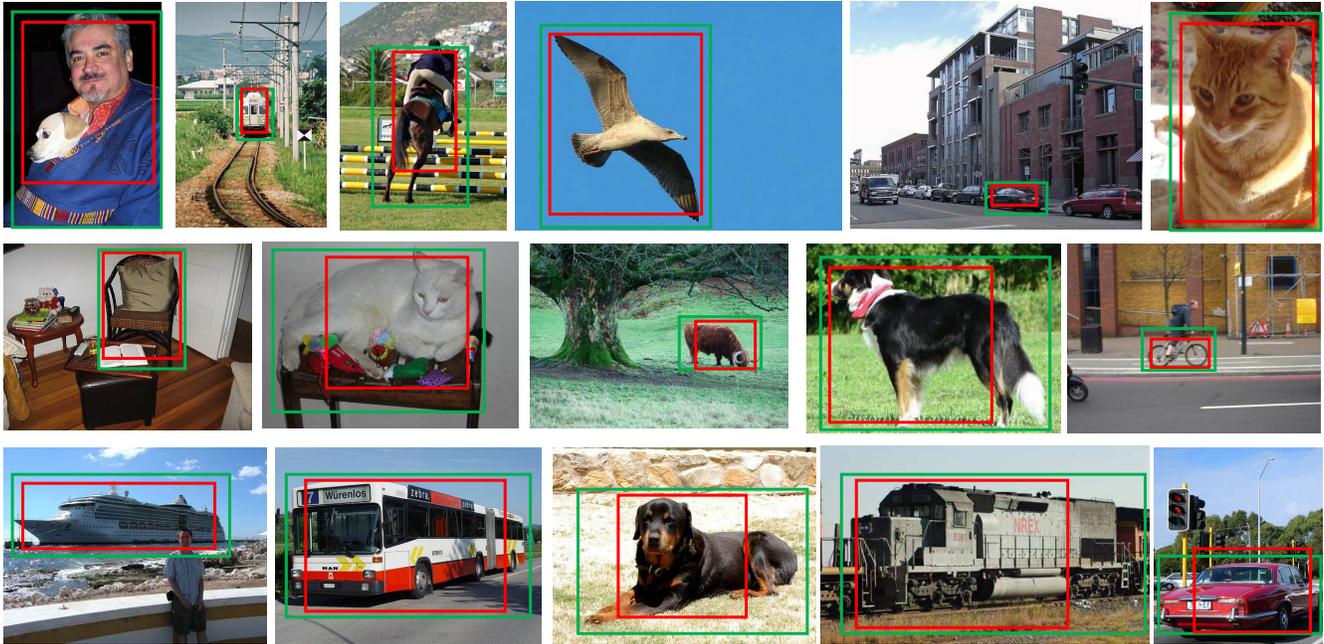


Figure 4. Examples of our object detection results. Green bounding boxes are ground-truth annotations and red boxes are positive detection. Images are sampled from PASCAL VOC 2007 test set.

a semantic relationship between improvement gains using different CNN architectures in our networks in comparison with using the same CNNs in other methods. We have almost the same improvement with two different architectures over other methods.

The localization performance with CorLoc metric is also shown in Table 3 on PASCAL VOC 2007. Our best performance is 56.7% which is achieved by 3stage cascade network using VGG-16 architecture. However, our network with the Alexnet outperformed the other methods using similar network architectures with same number of layers and other non deep learning methods. Most of the other works use CNNs as some part of their pipeline, not in an end-to-end scheme or use it simply as a feature extractor. Differently, our cascaded deep networks bring multiple concepts together in a single training method, learn better appearance model and feature representation for objects under weakly supervision circumstances.

We also compared our object detector results on ILSVRC'13 only with [18, 35], since no other weakly supervised object detector methods have been tried on this dataset. Results are shown in Table 4 and similar to our other tests, we have achieved better number in performance. Since, some part of our work is inspired by GAP networks from [36], we compared our weakly supervised localization on the ILSVRC'14 dataset following their experimental setups and the results are in Table 5.

Object detection training: We compared two different approaches of training object detection using Fast-RCNN,

Method	VOC2010	VOC2012	ILSVRC 2013
Cinbis et al. [7]	27.4	—	—
Wang et al. [35]	—	—	6.0
Li et al., Alexnet [18]	21.4	22.4	7.7
Li et al., VGG16 [18]	30.7	29.1	10.8
WSDDN [6]	36.2	—	—
WCCN_2stage_Alexnet	27.6	27.3	9.1
WCCN_2stage_VGG16	37.8	36.4	14.6
WCCN_3stage_Alexnet	28.8	28.4	9.8
WCCN_3stage_VGG16	39.5	37.9	16.3

Table 4. Detection performance (%) comparison on VOC 2010, 2012 test set and ILSVRC 2013 validation set.

implemented in Caffe [14] which both cases use our generated pseudo ground-truth. Since the Fast-RCNN [12] is a supervised method, we use the pseudo ground-truth (GT) bounding boxes which are generated by our cascaded networks. Proved by our experiments, in the Fig.5, it is shown that the Fast-RCNN can also perform with good results using our input bounding boxes. Fast-RCNN trained by our generated GT performs slightly better than our transfered model the average precision of PASCAL VOC 2007 test set (0.3%). The main goal of this work is to find the most representative and discriminative samples that signify the existing categories in each image.

Object proposals: In our work, we evaluated the effect of different unsupervised object proposals generator. Edge-Box [37] and SelectiveSearch [33] are compared based on the detector trained by our networks. According to the re-

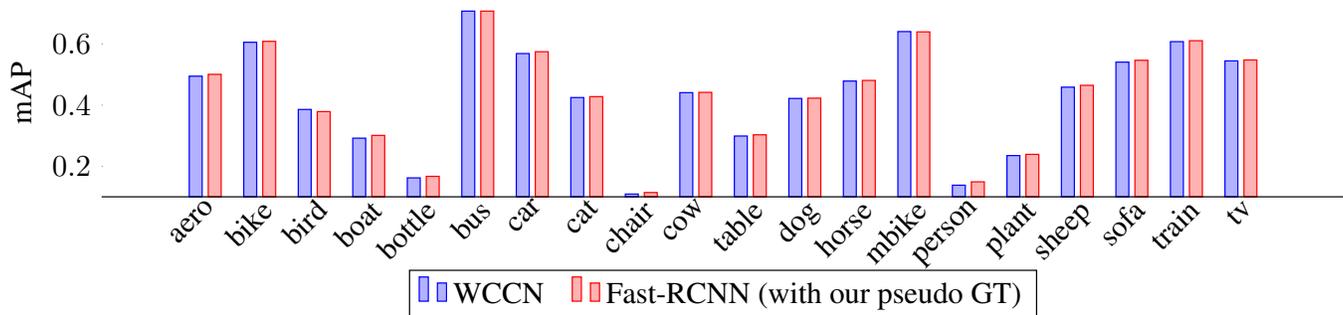


Figure 5. Comparison between our detection full pipeline and training Fast-RCNN using pseudo ground-truth bounding boxes extracted by our method.

Method	Detection top-1	Classification top-1
Alexnet	65.17	42.6
VGG16	61.12	31.2
Alexnet-GAP [36]	63.75	44.9
VGG16-GAP [36]	57.20	33.4
WCCN_2stage_Alexnet	62.2	41.2
WCCN_2stage_VGG16	55.6	30.4

Table 5. Detection and classification top-1 error (%) on ILSVRC’14 validation set

sults on the VOC 2007 detection test set, by training 2stage cascade using Alexnet with Edgebox, approximately 1.5% improvement can be obtained over SelectiveSearch. Similar to the other works like [6, 13], EdgeBox performs better with CNN based object detectors.

4.4. Classification performance

Our proposed network designs has dual purposes: object detection and classification in a weakly supervision manner. Obviously the structure of our cascade is helpful for training classification pipeline on images with multiple objects and minimum supervision of labels. We evaluated our method on PASCAL VOC 2007 and ILSVRC 2014. The performance is compared with other approaches which use novel methods in deep learning for classification on these datasets.

Table 2 presents the comparison on VOC 2007 with different CNN architectures for all of the methods. Since first stage of our cascade is similar to [36], we show the result of classification on ILSVRC’14, the large scale dataset for classification, in Table 5.

4.5. Cascade Architecture Study

To do ablation study over the performance of different stages of proposed cascades, it can be noticed that all of the results show how each of the proposed cascades can affect the performance in detection or classification. Each stage in our multi-stage cascaded CNN can be analyzed by comparison with the CNN-based methods in same context. Training

the stage with multiple instance loss can improve learning the best sample of each category over other works [36, 6]. It can be observed that adding the stage of segmentation to exploit better regions can outperform the two-stage cascade. Adding segmentation stage has impact on finding more accurate initial guess of object locations. For an instance of using the segmentation stage by Alexnet architecture, cascaded network improves almost 2.5% on detection and 2% on classification in PASCAL VOC 2007.

5. Conclusion

Our idea of weak cascaded convolutional networks (WCCN) is about the approaches of cascaded CNNs for weakly supervised visual learning tasks like object detection, localization and classification. In this work, we proposed two multi-stage cascaded networks with different loss functions in each stage to conclude a better pipeline of deep convolutional neural network learning with weak supervision of object labels on images. Our insight was a paradigm of multi-task learning effectiveness using deep neural networks. We proved that our multi-task learning approaches that incorporate localization, multiple instance learning and weakly supervised segmentation of object regions achieve the state-of-the-art performance in weakly supervised object detection and classification. The extensive experiments for object detection and classification tasks on various datasets like PASCAL VOC 2007, 2010, 2012 and also large scale datasets, ILSVRC 2013, 2014 present the full capability of the proposed method.

Acknowledgements

This work was supported by DBOF PhD scholarship, KU Leuven CAMETRON project. The authors would like to thank Nvidia for GPU donation.

References

- [1] B. Alexe, T. Deselaers, and V. Ferrari. What is an object? In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, 2010. 2

- [2] L. Bazzani, A. Bergamo, D. Anguelov, and L. Torresani. Self-taught object localization with deep networks. In *WACV*, 2016. 2
- [3] A. Bearman, O. Russakovsky, V. Ferrari, and L. Fei-Fei. What’s the Point: Semantic Segmentation with Point Supervision. *ECCV*, 2016. 4, 5
- [4] H. Bilen, M. Pedersoli, and T. Tuytelaars. Weakly supervised object detection with posterior regularization. In *BMVC*, 2014. 2, 6
- [5] H. Bilen, M. Pedersoli, and T. Tuytelaars. Weakly supervised object detection with convex clustering. In *CVPR*, 2015. 2, 6
- [6] H. Bilen and A. Vedaldi. Weakly supervised deep detection networks. In *CVPR*, 2016. 6, 7, 8
- [7] R. Cinbis, J. Verbeek, and C. Schmid. Weakly supervised object localization with multi-fold multiple instance learning. *IEEE transactions on pattern analysis and machine intelligence*, 2016. 1, 2, 6, 7
- [8] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009. 5
- [9] T. Deselaers, B. Alexe, and V. Ferrari. Localizing objects while learning their appearance. In *European conference on computer vision*, 2010. 2
- [10] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part-based models. *IEEE transactions on pattern analysis and machine intelligence*, 2010. 2
- [11] A. Ghodrati, A. Diba, M. Pedersoli, T. Tuytelaars, and L. Van Gool. Deepproposal: Hunting objects by cascading deep convolutional layers. In *Proceedings of the IEEE International Conference on Computer Vision*, 2015. 2
- [12] R. Girshick. Fast r-cnn. In *IEEE International Conference on Computer Vision (ICCV)*, 2015. 1, 3, 5, 7
- [13] K. He, X. Zhang, S. Ren, and J. Sun. Spatial pyramid pooling in deep convolutional networks for visual recognition. In *ECCV*, 2014. 6, 8
- [14] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. In *ACM MM*, 2014. 7
- [15] A. Joulin and F. Bach. A convex relaxation for weakly supervised classifiers. In *ICML*, 2012. 2
- [16] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012. 5
- [17] M. P. Kumar, B. Packer, and D. Koller. Self-paced learning for latent variable models. In *Advances in Neural Information Processing Systems*, 2010. 2
- [18] D. Li, J.-B. Huang, Y. Li, S. Wang, and M.-H. Yang. Weakly supervised object localization with progressive domain adaptation. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2016. 2, 6, 7
- [19] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, and S. Reed. Ssd: Single shot multibox detector. In *ECCV*, 2016. 1
- [20] M. H. Nguyen, L. Torresani, F. de la Torre, and C. Rother. Weakly supervised discriminative localization and classification: a joint learning process. In *IEEE International Conference on Computer Vision*, 2009. 2
- [21] M. Oquab, L. Bottou, I. Laptev, and J. Sivic. Learning and transferring mid-level image representations using convolutional neural networks. In *CVPR*, 2014. 2, 6
- [22] M. Oquab, L. Bottou, I. Laptev, and J. Sivic. Is object localization for free?-weakly-supervised learning with convolutional neural networks. In *CVPR*, 2015. 2, 3, 5
- [23] M. Pandey and S. Lazebnik. Scene recognition and weakly supervised object localization with deformable part-based models. In *2011 International Conference on Computer Vision*, 2011. 2
- [24] D. Pathak, P. Krahenbuhl, and T. Darrell. Constrained convolutional neural networks for weakly supervised segmentation. In *CVPR*, 2015. 4
- [25] S. Reed, H. Lee, D. Anguelov, C. Szegedy, D. Erhan, and A. Rabinovich. Training deep neural networks on noisy labels with bootstrapping. In *ICML*, 2014. 2
- [26] S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, 2015. 1
- [27] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In *ICLR*, 2015. 5, 6
- [28] P. Siva, C. Russell, and T. Xiang. In defence of negative mining for annotating weakly labelled data. In *European Conference on Computer Vision*, 2012. 2
- [29] P. Siva and T. Xiang. Weakly supervised object detector learning with model drift detection. In *International Conference on Computer Vision*, 2011. 2
- [30] H. O. Song, R. B. Girshick, S. Jegelka, J. Mairal, Z. Harchaoui, T. Darrell, et al. On learning to localize objects with minimal supervision. 2
- [31] 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. 2
- [32] S. Sukhbaatar, J. Bruna, M. Paluri, L. Bourdev, and R. Fergus. Training convolutional networks with noisy labels. *arXiv preprint arXiv:1406.2080*, 2014. 2
- [33] J. R. Uijlings, K. E. van de Sande, T. Gevers, and A. W. Smeulders. Selective search for object recognition. *International journal of computer vision*, 2013. 2, 7
- [34] A. Vedaldi and K. Lenc. Matconvnet: Convolutional neural networks for matlab. In *ACM MM*, 2015. 6
- [35] C. Wang, W. Ren, K. Huang, and T. Tan. Weakly supervised object localization with latent category learning. In *European Conference on Computer Vision*, 2014. 2, 6, 7
- [36] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba. Learning deep features for discriminative localization. In *CVPR*, 2016. 3, 4, 5, 6, 7, 8
- [37] C. L. Zitnick and P. Dollár. Edge boxes: Locating object proposals from edges. In *European Conference on Computer Vision*, 2014. 4, 5, 7