Weakly Supervised Localization Using Deep Feature Maps

Archith John Bency[1(✉)], Heesung Kwon[2], Hyungtae Lee[2,3], S. Karthikeyan[1], and B.S. Manjunath[1]

[1] University of California, Santa Barbara, CA, USA
{archith,karthikeyan,manj}@ece.ucsb.edu
[2] Army Research Laboratory, Adelphi, MD, USA
heesung.kwon.civ@mail.mil, htlee@umiacs.umd.edu
[3] Booz Allen Hamilton Inc., McLean, VA, USA

Abstract. Object localization is an important computer vision problem with a variety of applications. The lack of large scale object-level annotations and the relative abundance of image-level labels makes a compelling case for weak supervision in the object localization task. Deep Convolutional Neural Networks are a class of state-of-the-art methods for the related problem of object recognition. In this paper, we describe a novel object localization algorithm which uses classification networks trained on only image labels. This weakly supervised method leverages local spatial and semantic patterns captured in the convolutional layers of classification networks. We propose an efficient beam search based approach to detect and localize multiple objects in images. The proposed method significantly outperforms the state-of-the-art in standard object localization data-sets.

Keywords: Weakly supervised methods · Object localization · Deep convolutional networks

1 Introduction

Given an image, an object localization method aims to recognize and locate interesting objects within the image. The ability to localize objects in images and videos efficiently and accurately opens up a lot of applications like automated vehicular systems, searching online shopping catalogues, home and health-care automation among others. Objects can occur in images in varying conditions of occlusion, illumination, scale, pose and context. These variations make object detection a challenging problems in the field of computer vision.

The current state of the art in object detection includes methods which involve 'strong' supervision. In the context of object detection, strong supervision entails annotating localization and pose information about present objects of interest. Generating such rich annotations is a time-consuming process and is

B. Leibe et al. (Eds.): ECCV 2016, Part I, LNCS 9905, pp. 714–731, 2016.
DOI: 10.1007/978-3-319-46448-0_43

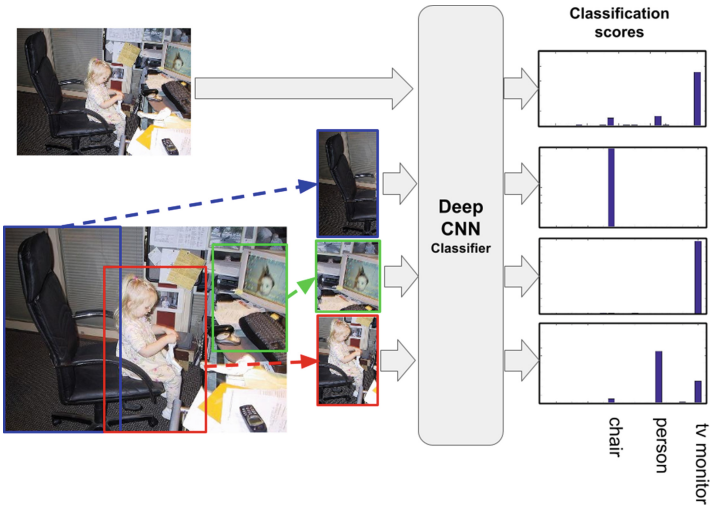

Fig. 1. When localizations centered around objects of interest are classified by Deep CNNs, the corresponding object classes are assigned high scores (Color figure online)

expensive to perform over large data-sets. Weak supervision lends itself to large-scale object detection for data-sets where only image-level labels are available. Effective localization under weak supervision enables extensions to new object classes and modalities without human-generated object bounding box annotations. Also, such methods enable generation of inexpensive training data for training object detectors with strong supervision.

Deep Convolutional Neural Networks (CNNs) [28,46] have created new benchmarks in the object recognition challenge [11]. CNNs for object recognition are trained using image-level labels to predict the presence of objects of interest in new test images. A common paradigm in analyzing CNNs has emerged where the convolutional layers are considered as data-driven feature extractors and the subsequent fully-connected layers constitute hyperplanes which delineate object categories in the learnt feature space. Non-linearities through Rectified Linear Units (ReLU) and sigmoidal transfer functions have helped to learn complex mapping functions which relate images to labels. The convolutional layers encode both semantic and spatial information extracted from training data. This information is represented by activations from the convolutional units in the network which are commonly termed as Feature Maps.

In this paper, we present a method that exploits correlation between semantic information present in Feature Maps and localization of an object of interest within an image. An example of such correlation can be seen in Fig. 1. Note that crudely localized image-patches with the objects of classes,'chair', 'person' and 'tv monitor', generate high classification scores for the corresponding classes. This suggests that one can coarsely localize objects solely by image classification scores in this context.

CNN based classifiers are trained for the task of image recognition on large image classification data-sets [11,14,15]. The learnt convolutional filters compute

spatially localized activations across layers for a given test image [31]. We examine the activation values in the outermost convolutional layer and propose localization candidates (or bounding boxes) which maximize classification scores for a class of interest. Class scores vary across localization candidates because of the aforementioned local nature of the convolutional filters. We then progressively explore smaller and smaller regions of interest till a point is reached where the classifier is no longer able to discriminate amongst the classes of interest. The localization candidates are organized in a search tree, the root node being represented by the entire test image. As we traverse from the root node towards the leaf nodes, we consider finer regions of interest. To approximate the search for optimal localization candidates, we adopt a beam search strategy where the number of candidate bounding boxes are restricted as we progress to finer localizations. This strategy enables efficient localization of multiple objects of multiple classes in images. We outperform the state-of-the-art in localization accuracy by a significant margin on two standard data-sets with complex scenes, PASCAL VOC 2012 [15] and the much larger MS COCO [29].

The main contributions of this paper are:

- We present a method that tackles the problem of object localization for images in a weakly supervised setting using deep convolutional neural networks trained for the simpler task of image-level classification.
- We propose a method where the correlation between spatial and semantic information in the convolutional layers and localization of objects in images is used explicitly for the localization problem.

2 Related Work

The task of object detection is one of the fundamental problems in computer vision with wide applicability. Variability of object appearance in images makes object detection and localization a very challenging task and thus has attracted a large body of work. Surveys of the state-of-the-art are provided in [39,50].

A large selection of relevant work are trained in the strong supervision paradigm with detailed annotated ground truth in the form of bounding boxes [16,48], object masks [4,21,27] and 3D object appearance cues [20,44]. The requirement of rich annotations curb the application of these methods in data-sets and modalities where training data is limited to weaker forms of labeling. Weak supervision for object detection tries to work around this limitation by learning localization cues from large collection of data with in-expensive annotations.

Large data-sets like Imagenet [11] and MS COCO are available with image-level labels. There has been significant work in this direction for object localization and segmentation [3,7,12,17,23,38]. Apart from image-level labels, other kinds of weak supervision include using eye-tracking data [37,43].

Deep convolutional neural networks (CNN) have seen a surge of attention from the computer vision community in the recent years. New benchmarks have been created in diverse tasks such as image classification and recognition

[5,28,45,46], object detection [19,36,42,51,53] and object segmentation [6,30,33] among others by methods building on deep convolutional network architectures. These networks perform tasks using feature representations learnt from training data instead of traditional hand-engineered features [10,16,32]. Typical algorithms of this paradigm perform inference over the last layer of the network. There have been recent works [9,22,24] which exploit semantic information encoded in convolutional feature map activations for semantic segmentation and object detection. A prerequisite for these CNN-based algorithms is strong supervision with systems focused on detection requiring location masks or object bounding boxes for training. [52] studies the presence of object detector characteristics in image-classification CNNs, but does not provide a computational method to carry out object detection.

Oquab et al. [35] has proposed a weakly supervised object localization system which learns from training samples with objects in composite scenes by explicitly searching over candidate object locations and scales during the training phase. While this method performs well on data-sets with complex scenes, the extent of localization is limited with respect to estimating one point in the test image. The extent of the object is not estimated and detecting multiple instances of the same object class is not considered. In our proposed approach, we estimate both the location and extent of objects and are capable of estimating multiple instances of objects in the test image. Also, we use pre-existing classification networks for localization where as [35] proposes training custom adaptation layers.

3 Weakly Supervised Object Localization

3.1 Overview of the Method

We aim to localize and recognize objects in images using CNNs trained for classification. There are two distinct phases. The first phase consists of learning image-level recognition from training image sets using existing Deep CNN architectures. We use the popular Alexnet [28] and VGG-16 [45] networks for our experiments. The next phase involves generating localization candidates in the form of bounding boxes for object classes of interest. These candidates are generated from a spatial grid corresponding to the final convolutional layer of the network and are organized in a search tree. We carry out a beam-search based exploration of these candidates with the image classifier scoring the candidates and reach at a set of final localization candidates for each class of interest.

3.2 Network Architecture and Training

The Alexnet network has five convolutional layers with associated rectification and pooling layers $C_1, C_2, \ldots, C_5$, along with three fully connected layers F_6, F_7, F_8 with $M_6 = \sigma(W_6 M_5 + B_6)$, $M_7 = \sigma(W_7 M_6 + B_7)$ and $M_8 = \gamma(W_8 M_7 + B_8)$. W_n, B_n are learn-able parameters for the n-th layer, M_n is the output of the n-th layer. $\sigma(\mathbf{X}) = \max(\mathbf{0}, \mathbf{X})$ is the rectification function and $\gamma(\mathbf{X}) = [e^{\mathbf{X}[i]}/\Sigma_j e^{\mathbf{X}[j]}]$

is the softmax function. Of particular interest to us is the output of the last convolutional layer C_5, M_5 which we will refer to subsequent sections.

We learn the network parameters through stochastic gradient descent and back-propagation of learning loss error [41] from the classification layer back through the fully connected and convolutional layers. Keeping in mind that objects of multiple classes can be present in the same training image, we use the cross entropy loss function to model error loss J between ground truth class probabilities $\{p_k\}$ and predicted class probabilities $\{\hat{p}_k\}$, where $k \in \{0, 1, ..., K-1\}$ indexes the class labels.

$$J = -\frac{1}{K} \sum_{k=0}^{K-1} [p_k \log \hat{p}_k + (1 - p_k) \log(1 - \hat{p}_k)] \tag{1}$$

As specified in [34], we remove F_8 and add two additional fully connected adaptation layers F_a, F_b. Similar to the Alexnet network, the output of these layers are computed as $M_a = \sigma(W_a M_7 + B_a)$ and $M_b = \gamma(W_b M_a + B_b)$. In order to assess the effectiveness of the proposed method for localization, these additional layers are added to facilitate re-training of the network from the Imagenet data-set to the Pascal VOC or MS COCO object detection data-sets. We initialize network parameters to values trained on the Imagenet data-set and fine-tune them [26] to adapt onto a target data-set. This is achieved by setting the learning rate parameter for the last layer weights to a higher value relative to earlier layer weights. An illustration of the network architecture is presented in Fig. 2 of [34].

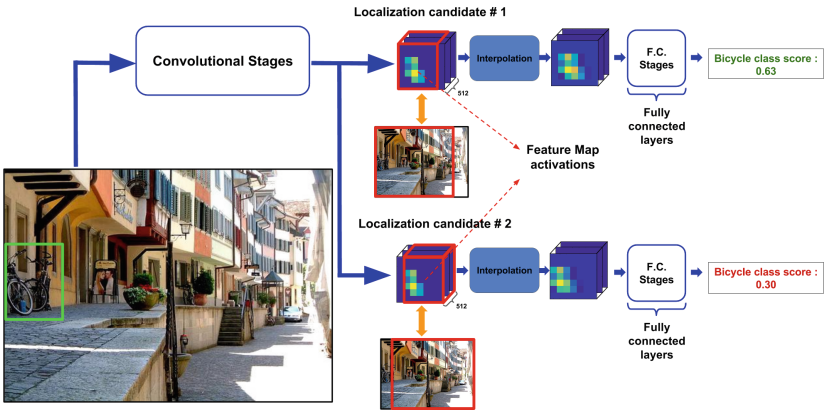

Fig. 2. An illustration of how two different localization candidates are compared in the localization process. Candidate #1 scores higher for the bicycle class than candidate #2. The first candidate is further iterated upon to achieve finer localization. The green box in the left image denotes ground-truth location of the bicycle object (Color figure online)

We train the augmented network on labeled samples from the target data-set. The trained network produces class scores at the final layer which are treated as probability estimates of the presence of a class in the test image.

The VGG-16 network, being similar to the Alexnet network, has thirteen convolutional layers $C_1, C_2, C_3, ...C_{13}$ with associated rectification and pooling layers, along with three fully connected layers F_6, F_7, F_8. Similar to the Alexnet network, the feature map M_{13} is of special interest to us. The increased number of layers and associated learnable parameters provides an improved image recognition performance when compared to the Alexnet network. The improvement however comes at the cost of increased GPU memory (442 MB vs 735 MB) and computations (6 ms vs 26 ms for classifying an image).

In addition to using image-labels to train the deep CNNs, we also use label co-occurrence information to improve classification. Some classes tend to occur together frequently. For example, people and motorbikes or people and chairs tend to share training samples. We treat the class scores from the classifier as unary scores and combine them with the likelihood of co-existence of multiple objects of different classes in the same object. We model the co-existence likelihood by building a co-occurrence matrix for class labels from the training data-set. For the class b_i,

$$s_{comb}(b_i) = s_{unary}(b_i) + \alpha \sum_{i \neq j} s_{pair}(b_i|b_j) \tag{2}$$

$$s_{pair}(b_i|b_j) = p_{pair}(b_i|b_j) s_{unary}(b_j) \tag{3}$$

$$p_{pair}(b_i|b_j) = \frac{|b_i \cap b_j|}{|b_j|} \tag{4}$$

where s_{unary} is the initial classification score for the test image, s_{pair} is the pairwise score, $|b_i \cap b_j|$ denotes the number of training samples containing the labels b_i and b_j and s_{comb} is the combined score which we use to re-score the classes for the test image. The parameter α denotes the importance given for pair-wise information in re-scoring. An optimal value is derived by testing over a randomly sampled validation sub-set from the training set.

3.3 Localization

In deep CNNs trained for classification, feature map activation values are the result of repeated localized convolutions, rectification (or other non-linear operations) and spatial pooling. Hence the structure of the network inherently provides a receptive field for each activation on the input image. The foot-print region becomes progressively coarser as we go deeper in the layers towards the fully connected layers. In a first attempt, we explore ways to exploit the spatial information encoded in the last convolutional layer for object localization.

Also, standard state-of-the-art object recognition data-sets (for e.g. Imagenet) typically have the object of interest represented in the middle of training samples. This gives rise to a bias in the classifier performance where more

centered an object is in the input image, higher the corresponding class score becomes. An example is illustrated in Fig. 1. The correlation between the location of objects and class scores has been observed in other works [19,34].

A naive approach to exploit the correlation would be to carry out a multi-scale sliding window sampling of sub-images from the test sample and spatially pool the classifier scores to generate a heat map of possible object locations for a given object class C. The number of sub-images required for effective localization can be in the order of thousands. Although powerful hardware like GPUs have brought image recognition CNNs into the domain of real-time methods, processing a large number of windows for every test sample is prohibitively expensive. A class of object detection methods [19] try to reduce the number of candidate windows by using object region proposal methods [1,47]. Time taken to detect objects in each image using these methods still range in tens of seconds when using powerful GPUs.

For a more computationally efficient approach, we take advantage of the spatial and semantic information encoded in the final convolutional feature maps to guide the search process. We refer to the maps as M_5 for Alexnet and M_{13} for VGG-16 in the Sect. 3.2. For a general CNN network, the final convolutional layer is of size $L \times L \times T$ which means there are T feature maps of size $L \times L$. For the Alexnet and VGG-16 networks, the feature maps are of size $6 \times 6 \times 256$ and $7 \times 7 \times 512$ respectively.

Given a test image I, we forward propagate the layer responses for the image up-to the final convolutional layer C_{last} and generate the feature map activations M_{last}. We generate localization candidates which are sub-grids of the $L \times L$ grid. In concrete terms, these candidates are parametrised as boxes $b_i = [x_i, y_i, w_i, h_i]$ for $i = 1, 2, \ldots, B$ where x, y, w and h represent the coordinates of the upper-left corner, width and height and B is the total number of possible sub-grids. For each localization candidate, we sample the feature map activations contained within the corresponding boxes and interpolate them over the entire $L \times L$ grid. This is done independently over all T feature maps. For the box b_i,

$$\hat{M}^t_{last}(x, y) = f(M^t_{last}(x', y')) \quad \forall\ x_i \leq x' \leq x_i + w_i - 1,\ y_i \leq y' \leq y_i + h_i - 1,\ t \in 0, 1, \ldots, T - 1$$

where $f(.)$ is an interpolation function which resizes the activation subset of size $w_i \times h_i$ to the size $L \times L$. In the above equation, $x, y \in \{0, \ldots, L-1\}$ and bi-linear interpolation is used. After obtaining the reconstructed feature maps $\hat{M}_{last}$, we forward propagate the activations into the fully connected layers and obtain the class scores. An illustration of this step is presented in Fig. 2.

A limitation of the above approach is related to the fact that interpolating from a smaller subset to the larger grid will introduce interpolation artifacts into the reconstructed feature maps. In order to mitigate the effects of the artifacts, we limit the localization candidates to boxes with $L - 1 \leq w_i \leq L$ and $L - 1 \leq h_i \leq L$. From this limited corpus of localization candidates, we generate the

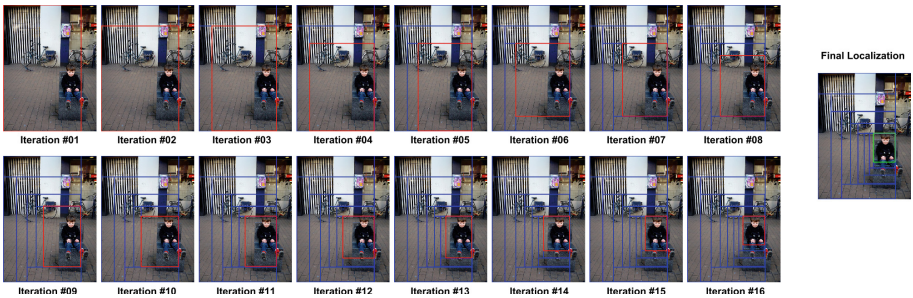

Fig. 3. A visual result of the proposed localization strategy on an image. The class scores for 'person' category are used to progressively localize the object of interest. Blue rectangles represent localization candidates considered in previous iterations and red rectangles represent current candidates (Color figure online)

corresponding $\hat{M}_{last}$ and consequently the object class scores, and choose the candidate with the highest class score. With the resultant localization candidate box b_r, we backproject onto the image space by cropping:

$$\begin{aligned} x_{crop} &= \frac{x_r}{L}W,\ y_{crop} = \frac{y_r}{L}H \\ w_{crop} &= \frac{w_r}{L}W,\ h_{crop} = \frac{h_r}{L}H \end{aligned} \tag{5}$$

$$I_{crop}(x, y) = I(x + x_{crop}, y + y_{crop})\ \forall\ 0 \leq x < w_{crop} \\ 0 \leq y < h_{crop}$$

where x, y indicate pixel locations, and W and H are width and height of the test image respectively. We then repeat the above described localization process on I_{crop} till a predetermined number of iterations. A visual example of progress in the iterative process is shown in Fig. 3.

3.4 Search Strategy

The localization strategy can be visualized as traversing down a search-tree where each node corresponds to a localization candidate b_i. The root node of such a tree would be $b_0 = [0, 0, L, L]$. The children of a node b_i in the tree would be the candidates $\{b_j\}$ which lie within sub-grid corresponding to b_i and whose parameters $\{w_j\}$ and $\{h_j\}$ satisfy the below conditions:

$$w_i - 1 \leq w_j \leq w_i,\ h_i - 1 \leq h_j \leq h_i \tag{6}$$

We consider children nodes whose width or height values, but not both of them differ from the parent node by 1. This restriction is put in place so that we are minimally modifying the feature map activations for discriminating amongst

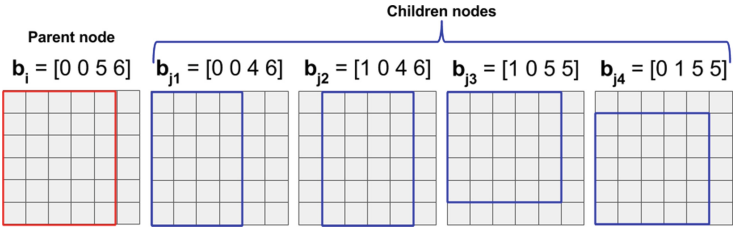

Fig. 4. An example of a parent node (represented in red) and it's children nodes (represented in blue) displayed on a 6×6 grid, as is the case for the Alexnet M_5 feature maps (Color figure online)

candidates. An example of a parent node b_i and the corresponding children node set $\{b_j\}$ is shown in Fig. 4.

During traversal, the child candidate with the highest score for the class C is selected. This approach is a greedy search strategy where we follow one path from the root node to a leaf node which represents the finest localization, and is susceptible to arrival at a locally optimal solution. Alternatively, we could evaluate all the nodes in the entire search-tree and could come up with the localization candidate with the highest score for class C. However, this would be computationally prohibitive.

To address this, we use the widely known beam-search [40] strategy. At each level of the search-tree we generate sets of children nodes from the current set of localization candidates using Eq. (6). We then rank them according to the scores for class C. Only the top M candidates are pursued for further evaluation. An illustration is presented in Fig. 5. In the Figure, we show an example where the two highest candidates are chosen at each level. The children nodes of these candidates are evaluated and ranked. We traverse a total of H levels. This approach helps us achieve a balance between keeping the number of computations to be tractable and avoiding greedy decisions. An additional advantage is the ability to localize multiple instances of the same class as the beam-search increases the set of localization candidates that are evaluated when compared to the greedy search strategy. Regions in the image corresponding to top-ranked candidates from each level are spatially sum-pooled using candidates scores to generate a heat-map. The heat-map is then thresholded. Bounding rectangles for the resulting binary blobs are extracted. The bounding rectangles are presented as detection results of our method. The average value of the heat-map values enclosed within detection boxes are assigned as the score of the boxes. In our experiments, we have set the value of M as 8 and H in the search tree as 10 for all data-sets. Heat-map thresholds for each class were determined by evaluation on a small validation sub-set from the training set.

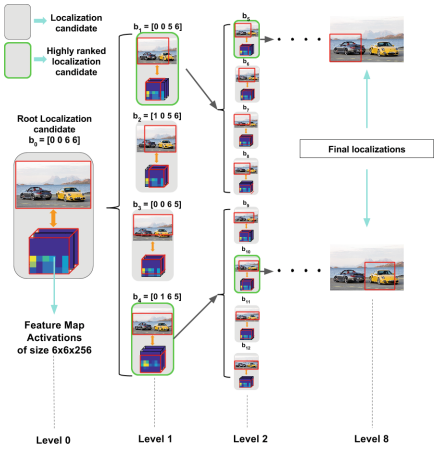

Fig. 5. A visual example of beam-search strategy to navigate the search tree amongst localization candidates. In this specific case, the class C is 'car', M is set to 2 and L is 6 (Color figure online)

4 Experiments

4.1 Data-Sets and Network Training

We evaluate our localization method on two large image data-sets, the PASCAL VOC 2007 [14], 2012 and the MS COCO. The VOC 2012 data-set has labels for 20 object categories and contains 5717 training images, 5823 validation images and 10991 test images. VOC 2007 shares the same class-labels with 2501 training images, 2510 validation images and 4952 test images. For the MS COCO data-set, there are 80000 images for training and 40504 images for validation with 80 object classes being present. These data-sets contain both image-level labels and object location annotations. For weak supervision we use the image-level labels from the training set to train classification networks and use the location annotations in the test and validation sets for evaluation.

We fine-tune the original VGG-16 and Alexnet networks (trained on Imagenet) by re-training the final fully connected layer for the VOC 2007, 2012 and MS COCO data-sets. We set the learning rate parameter to 0.001 which we decrease by a factor 10 for every 20000 training batches. Each training batch consists of 50 samples and the network was trained with 400000 batches. In order to balance the data-sets with respect to number of samples per class, we oversampled training samples from under-represented classes. We generate additional samples by a combination of adding white Gaussian noise and random rotations in the $\pm 30°$ range. We use Caffe [25] as our software platform for training and deploying classification networks on an NVIDIA TITAN X Desktop GPU.

4.2 Metrics

To compare results with the state-of-the-art in weakly supervised localization methods, we use the localization metric suggested by [35]. From the class-specific heat-maps generated by our localization, we extract the region of maximal response. If the center location of the maximal response lies within the ground-truth bounding box of an object of the same class, we label the location prediction as correct. If not, the false positive count is increased as the background was assigned to the class, and the false negative count is increased because object was not detected. The maximal value of the heat-map is assigned as confidence of the localization. The confidence score is then used to rank localizations and associated precision-recall (p-r) curves are generated for each object class. The p-r curves are characterized by an estimate of the area under the curve, which is termed as the Average Precision (AP). The AP score can vary from 0 to 100. An AP score of 100 signifies that all true positives were localized and no false positives were assigned scores. The AP scores for all classes are averaged to derive the Mean Average Precision (mAP), which presents a summarized score for the entire test set. This evaluation metric differs from the traditional Intersection-over-Union (IoU) measures to determine bounding box quality w.r.t the ground-truth, as the extent of the localization is not captured.

In addition to the above metric, we are interested in measuring how effective our method is in capturing the extent of the object of interest. We calculate the standard average precision for our detection results, where true positives are determined when intersection over union (IoU) between the predicted bounding boxes and the corresponding ground-truth box of the same class exceeds 0.5. We also utilize the CorLoc [13] metric which measures the percentage of samples containing the class of interest where the IoU between detected bounding box and ground-truth box exceeds 0.5.

4.3 Results

For obtaining localization results, we fine-tuned the networks using training samples from the *train* set of PASCAL VOC 2012 data-set and tested the trained networks on the *validation* set. As we use the class-scores from the classifiers to drive our localization strategy, good classification performance is essential for robust object localization. We present the classification performance on the PASCAL VOC 2012 *validation* set in Table 1. The VGG-16 network provides improved classification with respect to Alexnet and a consequent improvement can be seen in the localization scores as well.

In Table 1, we also compare the localization results of our method with respect to recent state-of-the-art weakly supervised localization methods on the PASCAL VOC 2012 *validation* set. We achieve a significant improvement of 5 mAP over the localization performance of Oquab et al. [35]. We also compare against the RCNN [19] and Fast RCNN [18] detectors which are trained with object-level bounding boxes. Similar to the way [35] evaluates [19], we select the most confident bounding box proposal per class per image for evaluation. Since

Table 1. Comparison of image classification and object localization scores on the PASCAL VOC 2012 *validation* set. For computing localization scores, responses are labeled as correct when the maximal responses fall within a ground-truth bounding box of the same class. False negatives are counted when no responses overlap with the ground-truth annotations. The class scores of the associated image-level classification are used to rank the responses and generate average precision scores. * RCNN and Fast-RCNN are trained for object detection with object-level bounding box data. We use the most confident bounding box per class in every image for evaluation

	Image classification		Localization				
	Proposed method + VGG-16	Proposed method + Alexnet	Proposed method + VGG-16	Proposed method + Alexnet	Oquab et al. [35]	RCNN* [19]	Fast-RCNN* [18]
airplane	93.0	92.0	90.1	90.0	90.3	92.0	79.2
bike	89.7	82.9	86.4	81.2	77.4	80.8	74.7
bird	91.4	87.2	86.4	81.2	77.4	80.8	74.7
boat	89.6	83.8	77.6	82.2	79.2	73.0	65.8
bottle	69.5	54.1	56.8	47.5	41.1	49.9	39.4
bus	90.9	87.3	90.3	86.7	87.8	86.8	82.3
car	81.6	74.5	68.3	64.9	66.4	77.7	64.8
cat	92.0	87.0	89.9	85.7	91.0	87.6	85.7
chair	69.3	56.4	54.7	53.9	47.3	50.4	54.5
cow	88.9	76.7	86.8	75.8	83.7	72.1	77.2
dining table	80.2	71.1	66.4	67.9	55.1	57.6	58.8
dog	90.4	83.5	88.5	82.2	88.8	82.9	85.1
horse	90.0	85.5	89.0	84.1	93.6	79.1	86.1
motorbike	90.0	84.3	88.1	83.4	85.2	89.8	80.5
person	91.6	88.1	78.5	83.9	87.4	88.1	76.6
plant	85.5	80.1	64.1	71.7	43.5	56.1	46.7
sheep	90.4	83.5	90.0	83.1	86.2	83.5	79.5
sofa	75.5	64.5	67.0	63.7	50.8	50.1	68.3
train	91.4	90.8	89.9	89.4	86.8	82.0	85.0
tv	89.6	81.4	82.6	78.2	66.5	76.6	60.0
mAP	86.5	79.8	**79.7**	**77.1**	74.5	74.8	71.3

deep neural networks are the state-of-the-art in object detection and localization tasks, we have compared with CNN-based methods.

We summarize the localization results for the much larger MS COCO *validation* data-set in Table 2. In-spite of having weaker classification performance (54.1 mAP vs 62.8 mAP) than the network used by [35], we are able to produce stronger localization performance by a margin of 2.5 mAP with the Alexnet network and a larger margin of 8 mAP with the VGG-16 network. This is a

Table 2. Comparison of localization and classification mAP scores for the MS COCO *validation* set

Method	Localization score (mAP)
Oquab et al. [35]	41.2
Proposed method + Alexnet	**43.7**
Proposed method + VGG-16	**49.2**

Table 3. Comparison of detection mean average precision scores and mean Correct Localization (CorLoc) scores on the PASCAL VOC 2007 *test* set

Method	Mean detection mAP	Mean CorLoc
Multi-fold MIL [8]	22.4	38.8
Bilen et al. [2]	27.7	43.7
LCL-pLSA [49]	**30.9**	**48.5**
Proposed method + VGG-16	25.7	46.8

significant improvement in performance over the state-of-the-art method. This is mainly because the proposed method actively seeks out image regions triggering higher classification scores for the class of interest. This form of active learning, where the localizing algorithm is the weak learner and the classifier is the strong teacher, lends us an advantage when trying to localize objects in complex scenes where multiple objects can exist in varying mutual configurations. This is also observed for the PASCAL VOC 2012 data-set. The fine-tuned VGG-16 and Alexnet networks produce classification performance scores of 74.3 mAP and 82.4 mAP respectively on the *test* set, where as the network used by [35] is scored at 86.3 mAP. As noted before, the proposed method outperforms competing methods on the localization task.

We have provided results on object bounding box detection and CorLoc for the PASCAL VOC 2007 *test* set in Table 3. We fine-tuned our network on the VOC 2007 *train* and the *validation* set, where 10 % of this joint group of images was set aside for parameter tuning, and provide test results on the *test* set. We are comparable in performance with respect to other state-of-the-art weakly supervised methods [2,8,49]. Examples of visual results for object detection are provided in Fig. 6.

Re-scoring the class likelihood scores using co-occurrence information referenced in Eq. (3) contributes to an improvement of 1.2 with the VGG-16 network in classification mAP score and 0.8 localization mAP score from Table 1.

Fig. 6. Visual sample results from the proposed method for Pascal VOC 2007 test set. Yellow rectangles overlaid on the images represent location and extent predictions. The locations of objects in the shown images are accurately estimated. Considering that only image-level labels are used for training, extent estimations are a challenging problem in this setting (Color figure online)

5 Discussion and Conclusions

The proposed method requires 2.6 s to localize an object on an image on machine with a 2.3 GHz CPU with a NVIDIA TITAN X desktop GPU. Compared to region proposal-based detection methods like RCNN which take around 20 s to detect objects, we achieve a significant reduction in localization time.

As can be seen from Table 1, an improvement in the classification performance (e.g. from Alexnet to VGG-16) directly leads to an improvement in the localization performance. As the state-of-the-art of the classification CNNs improves, we can expect a similar improvement in localization performance from our proposed method.

In summary, this method directly leverages feature map activations for object localization. This work uses the spatial and semantic information encoded in the convolutional layers and we have explored methods to utilize activations in the last convolutional layer. It would be interesting to see the improvements that could be derived by combining coarser semantic and finer localization information in earlier convolutional layers as well. Another direction to explore would be combining fast super-pixel segmentation and localization candidates from proposed method to improve detection performance.

The proposed method relies on weak supervision, with networks trained for image classification being used for localizing objects in test images with complex scenes and hence opens up possibilities for extending object localization to new object categories and image modalities without requiring expensive object-level annotations.

Acknowledgments. Research was sponsored by the Army Research Laboratory and was accomplished under Cooperative Agreement Number W911NF-09-2-0053 (the ARL Network Science CTA). The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Army Research Laboratory or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation here on.

References

1. Alexe, B., Deselaers, T., Ferrari, V.: Measuring the objectness of image windows. IEEE Trans. Pattern Anal. Mach. Intell. **34**(11), 2189–2202 (2012)
2. Bilen, H., Pedersoli, M., Tuytelaars, T.: Weakly supervised object detection with convex clustering. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1081–1089 (2015)
3. Blaschko, M., Vedaldi, A., Zisserman, A.: Simultaneous object detection and ranking with weak supervision. In: Advances in Neural Information Processing Systems, pp. 235–243 (2010)
4. Brox, T., Bourdev, L., Maji, S., Malik, J.: Object segmentation by alignment of poselet activations to image contours. In: 2011 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2225–2232. IEEE (2011)
5. Chatfield, K., Simonyan, K., Vedaldi, A., Zisserman, A.: Return of the devil in the details: delving deep into convolutional nets. In: British Machine Vision Conference (2014)
6. Chen, L.C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L.: Semantic image segmentation with deep convolutional nets and fully connected CRFs. In: ICLR (2015)
7. Chum, O., Zisserman, A.: An exemplar model for learning object classes. In: IEEE Conference on Computer Vision and Pattern Recognition, 2007. CVPR 2007, pp. 1–8. IEEE (2007)
8. Cinbis, R.G., Verbeek, J., Schmid, C.: Multi-fold MIL training for weakly supervised object localization. In: 2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2409–2416. IEEE (2014)
9. Dai, J., He, K., Sun, J.: Convolutional feature masking for joint object and stuff segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3992–4000 (2015)
10. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2005. CVpPR 2005, vol. 1, pp. 886–893. IEEE (2005)
11. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: ImageNet: a large-scale hierarchical image database. In: IEEE Conference on Computer Vision and Pattern Recognition, 2009. CVpPR 2009, pp. 248–255. IEEE (2009)
12. Deselaers, T., Alexe, B., Ferrari, V.: Localizing objects while learning their appearance. In: Daniilidis, K., Maragos, P., Paragios, N. (eds.) ECCV 2010, Part IV. LNCS, vol. 6314, pp. 452–466. Springer, Heidelberg (2010)
13. Deselaers, T., Alexe, B., Ferrari, V.: Weakly supervised localization and learning with generic knowledge. Int. J. Comput. Vis. **100**(3), 275–293 (2012)
14. Everingham, M., Van Gool, L., Williams, C.K.I., Winn, J., Zisserman, A.: The PASCAL Visual Object Classes Challenge 2007 (VOC 2007) Results (2007). http://www.pascal-network.org/challenges/VOC/voc2007/workshop/index.html

15. Everingham, M., Van Gool, L., Williams, C.K.I., Winn, J., Zisserman, A.: The PASCAL Visual Object Classes Challenge 2012 (VOC 2012) Results (2012). http://www.pascal-network.org/challenges/VOC/voc2012/workshop/index.html
16. Felzenszwalb, P.F., Girshick, R.B., McAllester, D., Ramanan, D.: Object detection with discriminatively trained part-based models. IEEE Trans. Pattern Anal. Mach. Intell. **32**(9), 1627–1645 (2010)
17. Galleguillos, C., Babenko, B., Rabinovich, A., Belongie, S.: Weakly supervised object localization with stable segmentations. In: Forsyth, D., Torr, P., Zisserman, A. (eds.) ECCV 2008, Part I. LNCS, vol. 5302, pp. 193–207. Springer, Heidelberg (2008)
18. Girshick, R.: Fast R-CNN. In: International Conference on Computer Vision (ICCV) (2015)
19. Girshick, R., Donahue, J., Darrell, T., Malik, J.: Rich feature hierarchies for accurate object detection and semantic segmentation. In: 2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 580–587. IEEE (2014)
20. Glasner, D., Galun, M., Alpert, S., Basri, R., Shakhnarovich, G.: Viewpoint-aware object detection and pose estimation. In: 2011 IEEE International Conference on Computer Vision (ICCV), pp. 1275–1282. IEEE (2011)
21. Hariharan, B., Arbeláez, P., Girshick, R., Malik, J.: Simultaneous Detection and Segmentation. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014, Part VII. LNCS, vol. 8695, pp. 297–312. Springer, Heidelberg (2014)
22. Hariharan, B., Arbelaez, P., Girshick, R., Malik, J.: Hypercolumns for object segmentation and fine-grained localization. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2015
23. Hartmann, G., et al.: Weakly supervised learning of object segmentations from web-scale video. In: Fusiello, A., Murino, V., Cucchiara, R. (eds.) ECCV 2012. LNCS, vol. 7583, pp. 198–208. Springer, Heidelberg (2012). doi:10.1007/978-3-642-33863-2_20
24. He, K., Zhang, X., Ren, S., Sun, J.: Spatial pyramid pooling in deep convolutional networks for visual recognition. IEEE Trans. Pattern Anal. Mach. Intell. **37**(9), 1904–1916 (2015)
25. Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., Guadarrama, S., Darrell, T.: Caffe: convolutional architecture for fast feature embedding. In: Proceedings of the ACM International Conference on Multimedia, pp. 675–678. ACM (2014)
26. Karayev, S., Trentacoste, M., Han, H., Agarwala, A., Darrell, T., Hertzmann, A., Winnemoeller, H.: Recognizing image style. In: Proceedings of the British Machine Vision Conference. BMVA Press (2014)
27. Kim, J., Grauman, K.: Shape sharing for object segmentation. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.) ECCV 2012, Part VII. LNCS, vol. 7578, pp. 444–458. Springer, Heidelberg (2012)
28. Krizhevsky, A., Sutskever, I., Hinton, G.E.: ImageNet classification with deep convolutional neural networks. In: Advances in Neural Information Processing Systems, pp. 1097–1105 (2012)
29. Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft COCO: common objects in context. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014, Part V. LNCS, vol. 8693, pp. 740–755. Springer, Heidelberg (2014)
30. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3431–3440 (2015)

31. Mahendran, A., Vedaldi, A.: Understanding deep image representations by inverting them. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2015
32. Murphy, K., Torralba, A., Eaton, D., Freeman, W.T.: Object detection and localization using local and global features. In: Ponce, J., Hebert, M., Schmid, C., Zisserman, A. (eds.) Toward Category-Level Object Recognition. LNCS, vol. 4170, pp. 382–400. Springer, Heidelberg (2006)
33. Noh, H., Hong, S., Han, B.: Learning deconvolution network for semantic segmentation. In: 2015 IEEE International Conference on Computer Vision (ICCV) (2015)
34. Oquab, M., Bottou, L., Laptev, I., Sivic, J.: Learning and transferring mid-level image representations using convolutional neural networks. In: CVPR (2014)
35. Oquab, M., Bottou, L., Laptev, I., Sivic, J.: Is object localization for free? weakly-supervised learning with convolutional neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2015)
36. Ouyang, W., Wang, X., Zeng, X., Qiu, S., Luo, P., Tian, Y., Li, H., Yang, S., Wang, Z., Loy, C.C., Tang, X.: DeepID-Net: deformable deep convolutional neural networks for object detection. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2015
37. Papadopoulos, D.P., Clarke, A.D.F., Keller, F., Ferrari, V.: Training object class detectors from eye tracking data. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014, Part V. LNCS, vol. 8693, pp. 361–376. Springer, Heidelberg (2014)
38. Pourian, N., Vadivel, K.S., Manjunath, B.: Weakly supervised graph based semantic segmentation by learning communities of image-parts. In: 2015 IEEE International Conference on Computer Vision (ICCV). IEEE (2015)
39. Roth, P.S., Winter, M.: Survey of appearance-based methods for object recognition, iCG01/08 (2008)
40. Rubin, S.M., Reddy, R.: The locus model of search and its use in image interpretation. IJCAI **2**, 590–595 (1977)
41. Rumelhart, D.E., Hinton, G.E., Williams, R.J.: Learning representations by back-propagating errors. Cogn. Model. **5**, 3 (1988)
42. Sermanet, P., Eigen, D., Zhang, X., Mathieu, M., Fergus, R., LeCun, Y.: OverFeat: integrated recognition, localization and detection using convolutional networks. In: International Conference on Learning Representations (ICLR 2014). CBLS, April 2014. http://openreview.net/document/d332e77d-459a-4af8-b3ed-55ba
43. Shanmuga Vadivel, K., Ngo, T., Eckstein, M., Manjunath, B.: Eye tracking assisted extraction of attentionally important objects from videos. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2015
44. Shrivastava, A., Gupta, A.: Building part-based object detectors via 3D geometry. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 1745–1752 (2013)
45. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. In: International Conference on Learning Representations (2015)
46. Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A.: Going deeper with convolutions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1–9 (2015)
47. Uijlings, J.R., van de Sande, K.E., Gevers, T., Smeulders, A.W.: Selective search for object recognition. Int. J. Comput. Vis. **104**(2), 154–171 (2013)

48. Viola, P., Jones, M.: Rapid object detection using a boosted cascade of simple features. In: Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2001. CVpPR 2001, vol. 1, pp. I-511. IEEE (2001)
49. Wang, C., Ren, W., Huang, K., Tan, T.: Weakly supervised object localization with latent category learning. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014, Part VI. LNCS, vol. 8694, pp. 431–445. Springer, Heidelberg (2014)
50. Zhang, X., Yang, Y.H., Han, Z., Wang, H., Gao, C.: Object class detection: a survey. ACM Comput. Surv. **46**(1), 10:1–10:53 (2013). http://doi.acm.org/10.1145/2522968.2522978
51. Zhang, Y., Sohn, K., Villegas, R., Pan, G., Lee, H.: Improving object detection with deep convolutional networks via Bayesian optimization and structured prediction. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2015
52. Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., Torralba, A.: Object detectors emerge in deep scene CNNs. In: International Conference on Learning Representations (ICLR) (2015)
53. Zhu, Y., Urtasun, R., Salakhutdinov, R., Fidler, S.: segDeepM: exploiting segmentation and context in deep neural networks for object detection. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2015