Fewer is More: Image Segmentation Based Weakly Supervised Object Detection with Partial Aggregation

Ce Ge
nwlgc@bupt.edu.cn
Jingyu Wang
wangjingyu@bupt.edu.cn
Qi Qi
qiqi@ebupt.com
Haifeng Sun
sunhaifeng_1@ebupt.com
Jianxin Liao
liaojianxin@ebupt.com

State Key Laboratory of Networking and Switching Technology,
Beijing University of Posts and Telecommunications,
Beijing, China

Abstract

We consider addressing the major failures in weakly supervised object detectors. As most weakly supervised object detection methods are based on pre-generated proposals, they often show two false detections: (i) group multiple object instances with one bounding box, and (ii) focus on only parts rather than the whole objects. We propose an image segmentation framework to help correctly detect individual instances. The input images are first segmented into several sub-images based on the proposal overlaps to uncouple the grouping objects. Then the batch of sub-images are fed into the convolutional network to train an object detector. Within each sub-image, a partial aggregation strategy is adopted to dynamically select a portion of the proposal-level scores to produce the sub-image-level output. This regularizes the model to learn context knowledge about the object content. Finally, the outputs of the sub-images are pooled together as the model prediction. The ideas are implemented with VGG-D backbone to be comparable with recent state-of-the-art weakly supervised methods. Extensive experiments on PASCAL VOC datasets show the superiority of our design. The proposed model outperforms other alternatives on detection, localization, and classification tasks.

1 Introduction

Increasing efforts are made to study weakly supervised object detection problem. However, as mentioned in [3], most methods encountered two main types of failures. One failure is to mistake multiple objects as one and mark them with only single bounding box. This situation often occurs in the cases of birds, planes, persons, and other images that contain crowded objects. The appearance of a group of objects in the same category looks similar, and they often overlap each other. They are easily mistakenly discriminated as one object

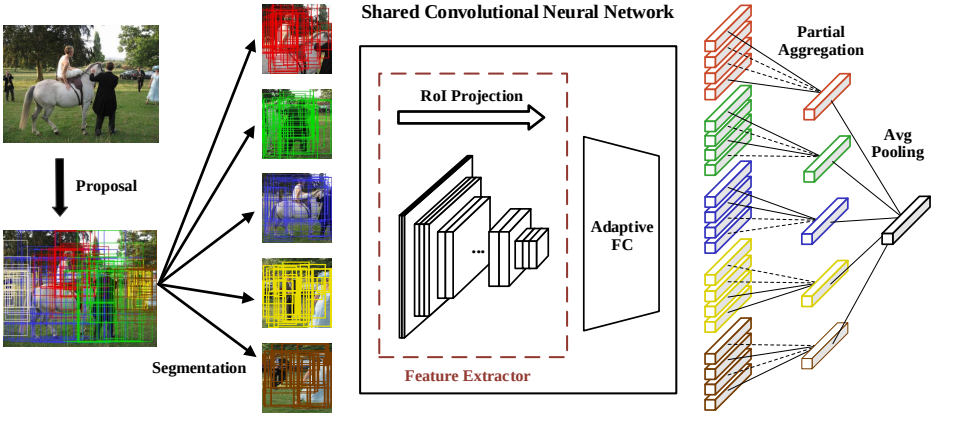

Figure 1: Overview of the architecture. Given an image, a set of proposals are generated by EdgeBoxes [32]. The Ncut algorithm [26] is applied to segment the images based on the proposal overlaps. The convolutional layers are first fine-tuned for multi-label classification. Then they are reserved as the shared feature extractor. The proposals within each sub-image are projected to the corresponding regions of interest (RoI) on the feature maps. An adaptive fully-connected layer is designed to handle the variable-size RoI. The proposal-level scores within each sub-image are dynamically summed by a partial aggregation strategy. Finally, the sub-image-level scores are average-pooled together to produce the prediction.

instance. And another frequent case of false detection is to focus only on the discriminative parts rather than the whole object. For images containing persons or animals, the detected bounding boxes are very likely to cover only the faces. The object detection process essentially depends on the classification scores. As pointed out [25], when training an image classifier, the Convolutional Neural Networks (CNN) are automatically learned to focus on the most discriminative parts, and it has no semantic knowledge about the object content. If these problems are resolved, the performance of many weakly supervised detection methods should be further improved.

We propose an image segmentation based framework consisting of two training phase: sub-image fine-tuning for multi-label classification and detector training with proposals. In order to address the multi-instance grouping issue, for both phases, the input images are segmented into several sub-images before being fed into the CNN. The intention of the image pre-segmentation is to uncoupled the close objects to make individual instances more distinguishable. The idea of this design is twofold:

(1) Each segmented sub-image contains fewer objects than the whole image. We expect that there is only one dominant object in each sub-image. It is obviously hard to achieve before we realize an object detector. However, the segmentation procedure still separate the tightly close objects, *e.g.* persons and birds. The model discrimination will concentrate on the true content of individual instance. Moreover, the feature space of sub-images is closer to that of the original training set (single-object images). The shift between the two feature domains is reduced. The fine-tuning thus becomes more stable and converges faster.

(2) Along with segmentation, some poor proposals could be filtered out, i.e., those cover object parts across multiple sub-images and those cover extra distracting backgrounds. Although such process does not change the segmentation number, it acts as a rectifier to refine the extent of each sub-image area. This is particularly helpful for the detector to determine the accurate object content.

Firstly, as most weakly supervised object detectors are built from networks pre-trained on ImageNet, domain adaptation is commonly used to transfer the feature knowledge to the target dataset. Rather than using the whole image, our image fine-tuning process is performed on the segmented sub-images. They are taken as a minibatch input through the CNN backbone (VGG-D in this case), and the classification scores of each are max-pooled together to form the final classification output.

After multi-label classification fine-tuning, the model is further fine-tuned with the objectness proposals for detector training. The architecture is depicted in Fig. 1. The input images are still segmented to several pieces. The sub-images as well as the proposals are fed into the shared CNN for end-to-end learning. The convolutional layers extract features and project the proposals onto feature maps. To reduce the loss of spatial information, an adaptive fully-connected layer is designed to substitute Region of Interest (RoI) pooling for variable-size input. A partial aggregation strategy is designed to dynamically sum a portion of score vectors in the sub-image level. We incorporate both the maximum and minimum scores in each category. The maximum scores ensure correct classification, while the minimum ones are crucial to determine the object extent. The partial aggregation strategy encourages the model to learn more context information, hence supports more accurate localization.

Overall, in this paper, we propose a novel image segmentation based weakly supervised object detection framework. The core idea is to separate the multi-instance images into sub-images containing fewer objects. With fewer and clearer visual features, more individual objects are correctly localized. The partial aggregation strategy regularizes the model to learn context information by selecting a portion of regions. The idea is succinct but effective. We evaluate our approaches on PASCAL VOC to demonstrate its superiority. Besides, the design is very flexible and can be easily adapted to other data and other backbone networks.

2 Related Work

Most existing methods formalize the weakly supervised learning problem as a Multiple Instance Learning (MIL) paradigm. In MIL paradigm, the image is viewed as a bag of instances (i.e., objects or regions). The optimization of MIL algorithm alternates between selecting positive samples and training an instance detector. Since the objective is none-convex, the optimization is easily trapped in a local optimum. To address the issue, many researches have designed the initialization and optimization elaborately. Deselaers *et al.* [8] proposed a conditional random field approach and initialize the object location based on the objectness measure [1]. Bilen *et al.* [5] involved additional domain-specific knowledge and use soft-max to make the optimization smoother. Song *et al.* [23] proposed a submodular cover algorithm to discover the initial training samples. Bilen *et al.* [4] designed a convex clustering to enforce a soft similarity between the selected regions. And Cinbis *et al.* [7] used a multi-fold split scheme to avoid poor convergence.

Another main line is to design end-to-end CNN models. Oquab *et al.* [24, 25] demonstrated the ability of CNN model to localize objects that is trained using only image-level supervision. Bilen *et al.* [3] further designed a two-stream weakly supervised deep detec-

tion network (WSDDN). Kantorov *et al.* [19] extended the Fast R-CNN[15] framework and proposed a context-aware method to perform weakly supervised object localization. More recently, Li *et al.* [23] proposed a two-step adaptation approach. They first transferred a pre-trained CNN model to performance multi-label classification, then applied a MIL mining strategy. The following work [17] proposed a seed-proposal discovery algorithm and a self-taught learning strategy to further improve the quality of the mined samples. These methods utilized instance discovery (or mining) algorithms of MIL. However, substantially their detector training phase is still an end-to-end CNN framework. Weakly supervised learning can gain more benefits from the combination of traditional machine learning methods and end-to-end CNN models. Our work mainly follows these studies. We propose a two-phase adaptation network, and the object detector is trained end-to-end with only weak supervision.

3 Methods

Almost all weakly supervised detection methods rely on objectness proposals to determine the possible location of objects. The commonly used proposal generating methods are based on low-level cues, *e.g.*, colors, intensities, and edges. The generated bounding boxes lack semantic information, so that some proposals cover parts of multiple objects. During detector training, the model learns to discriminate these wrong proposals as the top-scoring detections. To reduce the wrong cases, we propose to pre-segment the input images to distinguish individual object more clearly. Some improper proposals are filtered out to refine the segmentation. To be comparable with other weakly supervised methods, we implement our approach based on VGG-D net pre-trained on ImageNet.

3.1 Proposal-Based Image Segmentation

A few methods [1, 2, 6, 29, 32] have been proposed to generate high-quality proposal candidates to indicate possible regions of objects. We adopt one of the state-of-the-art methods, EdgeBoxes [32] for our experiments. The normalized cut algorithm (Ncut) [26] is a graph-theoretic partitioning method and is often used for image segmentation. Ncut defines a normalized disassociation measure to partition graph nodes into unbiased groups. Given an image I, denote the generated proposals by EdgeBoxes as $V = \{v_1, v_2, ..., v_l\}$, where l is the number of proposals. Then the proposals are treated as a weighted undirected graph $G = (V, E)$, where the nodes V are proposals and an edge $e \in E$ connecting two nodes indicates that the two proposals are overlapped. The weight of edge e is defined as the intersection over union (IoU) of the two proposals. For a given number m, the Ncut algorithm could partition the nodes (i.e., proposals) into m unbiased groups.

The Ncut algorithm partitions the proposal candidates into m non-empty groups. In such a way, every proposal must fall in one and only one group. It is observed from previous works [3, 22] that there are some improper proposals that may cause significant failures, i.e., those involve multiple object instances or redundant backgrounds. After Ncut grouping, these proposals are reserved but unwanted. But it is noticed that most proposals in each partitioned group overlap each other densely and cover around the correct object, while the improper proposals are sparsely distributed and dilated away from the object content. In order to filter out these improper proposals, we assign a statistical average density to each proposal. The average density of a proposal v_p is defined as the weighted average of the

overlap densities inside its coverage:

$$D(v_p) = \sum_{r \in R(v_p)} \frac{area(r)}{area(v_p)} \cdot d(r), \tag{1}$$

where $R(\cdot)$ returns all the superregions of a proposal. Function $area(\cdot)$ computes the area of the given superregion and $d(\cdot)$ provide its overlap density. In such a way, within each proposal group, the densities of proposal overlap can be viewed as a heatmap. The proportion of low-density proposals in each group is removed to refine the segmentation.

For each filtered proposal group, we take the minimum coordinates $(x_{\min}, y_{\min})$ of all the top left corners and the maximum coordinates $(x_{\max}, y_{\max})$ of all the bottom right corners. The rectangular areas determined by these coordinates are cut out as the sub-images.

3.2 Multi-label Adaptation

State-of-the-art CNN models [16, 21, 27] for visual recognition have well-designed convolutional layers. The stacked convolutional layers act as an extractor for general visual features. Under weakly supervised learning paradigm, the bounding-box annotations are not used during detector training. Hence, the feature extracting ability of the backbone neural network is important. It is helpful and very common to use a pre-trained CNN backbone.

The VGG-D net pre-trained on ImageNet performs single-label image classification. The original training images are provided as containing only one dominant object instance. In this schema, the scales of different objects are at a similar level. But for object detection task, the input images are assumed to contain multiple object instances that belong to different categories. To deal with the difference, many works replace the last fully-connected layer of the backbone network with a new C-way binary output layer (C is the number of object categories). Then the whole images are fed into the network to train a multi-label classifier. However, the instances in multi-object images tend to have different scales. When existing together, they share the same extracted feature maps. There is a big gap between the original single-object feature space and the target multi-instance feature space. The straightforward domain adaptation causes a shift between the two feature domains. Moreover, it becomes harder to extract informative features for small objects due to the low resolution. Scaling images to different sizes may be a viable solution for smaller objects. However, this leads to extra computation for large objects. And the scales are chosen manually and requires an estimation of the size distribution of all object instances.

The designed proposal-based image segmentation procedure transforms the multi-object images into a set of sub-images. By selecting an proper number of segmentation, each sub-image contains much fewer objects. The sub-images are rescaled to the same size and input to the CNN model together. For fine-tuning, the last fully-connected layer is replaced with a new C-way classification layer to match the ground-truth label. For each sub-image, a C-dim score vector $s = (s^{(1)}, s^{(2)}, ..., s^{(C)})$ is produced that represents the possibility distribution of all categories. These vectors are pooled together through an element-wise max operation to output the final scores $\hat{y} \in \mathbb{R}^C$, where each element of class c is calculated as:

$$\hat{y}^{(c)} = \max_{j \in \{1,...,m\}} s_j^{(c)}. \tag{2}$$

We designed this fine-tuning framework independently, but we noticed that it is much similar to the *image-fine-tuning (I-FT)* process in HCP [31]. The principal difference is that we use

sub-images as model input. Since the score vector s produced by the C-way output layer is a probability distribution, after max pooling, each element of output $\hat{y}$ is still in range $(0,1)$, but no longer mutually exclusive. The loss function is defined as the sum of C binary-log losses for category:

$$\mathcal{L}_{cls} = -\sum_{i=1}^{n}\sum_{c=1}^{C}(y_i^{(c)}\log\hat{y}_i^{(c)} + (1-y_i^{(c)})\log(1-\hat{y}_i^{(c)})). \tag{3}$$

3.3 Detector Training with Partial Aggregation

After multi-label classification fine-tuning, the convolutonal layers are reserved as a feature extractor for single objects. Then for proposal fine-tuning (i.e., detector training), many works [3, 10, 18, 20] adopted the implementation of region of interest (RoI) pooling layer [15] to pool the regional features to a fixed size, *e.g.* 7×7. This is necessary when fully-connected layer follows. However, the segmented sub-images have fewer effective pixels, and after $32\times$ downsampling (as the input size is 224×224, and the feature map size of `pool5` layer is 7×7), the resolution will be too low to extract informative features for small objects. Even we upsample the sub-images to bigger size to fit the CNN input, the two sampling steps will cause information loss of key spatial features. Since there is no coordinates regression, the localized information matters much to determine the exact content of objects. For weakly supervised learning, all feature neurons are useful for fine discrimination.

We design a size-insensitive fully-connected layer to handle the variable-size input. We retain the fine-tuned convolutional layers (remove `pool5` layer) and modify the first fully connected layer (`fc6`) for adaptive input size. Each neuron in `fc6` layer is connected to all the feature neurons in each region. Concretely, the adaptive input is implemented by 1×1 convolutions. The features of each region are first reduced to single channel while keeping the resolution, and then summed to a single value. Assume a regional feature volume $\boldsymbol{X}$ of size $h\times w\times d$, the operation can be formulized as

$$\phi(\boldsymbol{X}\mid w) = \sum_{i=1}^{h}\sum_{j=1}^{w}\boldsymbol{X}_{ij}^{\mathsf{T}}\cdot w, \tag{4}$$

where w is the weights of the 1×1 convolutional kernels. The weights are shared over features in each region, thus the fewer parameters also help to reduce overfitting.

Suppose the image I is segmented to m pieces $\{H_1,...,H_m\}$ as illustrated in Fig. 1. For a sub-image H_j, it contains a set of proposals $\mathcal{P}_j=\{p_1,...,p_{|\mathcal{P}_j|}\}$. Through the shared CNN, these proposals are encoded to region-level scores, denoted as a matrix $\boldsymbol{S}_j\in\mathbb{R}^{|\mathcal{P}_j|\times C}$. To match the ground-truth labels, the score matrix need to be suppressed to a C-dim vector. The simple way is to sum score vectors over all proposals [3]. But this causes the neural network to focus on only discriminative parts of objects (*e.g.* heads). From experimental studies, we found that dynamically select a portion of proposals helps to learn more context information. Similar findings are also mentioned in [11, 12]. The sub-image-level score vector $\hat{s}_j$ of H_j is computed by a partial aggregation strategy. Its each element is a partial sum of dynamically selected positive and negative proposals:

$$\hat{s}_j^{(c)} = \sum_{k\in\mathcal{P}^+}\boldsymbol{S}_j^{(k,c)} + \alpha\sum_{l\in\mathcal{P}^-}\boldsymbol{S}_j^{(l,c)}, \tag{5}$$

where $\mathcal{P}^+$ and $\mathcal{P}^-$ are the selected proposals of highest and lowest scores. The weighting factor α trade off their contributions. The positive proposals are crucial for category classification, while the negative proposals are helpful to distinguish backgrounds and the cluttered parts. To Improve localization performance while keeping classification capacity, the reasonable value of α should be in the range of 0 to 1. For simplicity, the value of α is empirically set to 0.5 throughout all experiments. The sub-image-level scores are average-pooled together to produce the final output $\hat{y}$. And the loss function for detector training is defined as the sum of C binary-log-log losses.

4 Experiments

We build our model based on VGG-D net and evaluate the performance on PASCAL VOC 2007 dataset [13, 14]. PASCAL VOC 2007 is the mostly used benchmarks for weakly supervised recognition tasks. It contains images of 20 object categories. As recommended, the *trainval* set (the union of training and validation) is used for training. We report our results in two metrics: AP and CorLoc. Average precision (AP) and the mean of AP (mAP) are standard PASCAL VOC protocols for object recognition challenge [14], and Correct Localization (CorLoc) [9] is defined to measure the localization performance. By convention, the AP metrics are evaluated on the *test* set, and the CorLoc is reported on the *trainval* set.

4.1 Experimental Setup

For multi-label *adaptation*, all the layers are initialized using the parameters pre-trained on ImageNet. The new 20-way output layer is randomly initialized. All the layers are fine-tuned with an initial learning rate 0.01 and decays to one-tenth after every 10 training epochs. For object detector *training*, the learning rate of feature extractor and fully connected layers are initialized to 10^{-4} and 10^{-3} respectively, and also decay to one-tenth. The loss functions of the two training phases are both optimized with weight decay of 0.0005. Horizontal flip is used for data augmentation. The model trainings are carried out for 40 epochs respectively. For *testing*, the input images are pre-segmented. A C-dim score vector is produced for each proposal. A standard non-maximum suppression (NMS) is performed in each category to remove duplicate detections. To avoid improper segmentation (*e.g.*, splitting an object into several sub-images or grouping too many objects in one sub-image), we adopt an ensemble configuration. The detection results of different number of segmentations are assembled and an additional NMS is applied over the whole image.

method	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	mbike	person	plant	sheep	sofa	train	tv	mAP
Cinbis *et al.* [8]	38.1	47.6	28.2	13.9	13.2	45.2	48.0	19.3	17.1	27.7	17.3	19.0	30.1	45.4	13.5	17.0	28.8	24.8	38.2	15.0	27.4
Song *et al.* [23]	27.6	41.9	19.7	9.1	10.4	35.8	39.1	33.6	0.6	20.9	10.0	27.7	29.4	39.2	9.1	19.3	20.5	17.1	35.6	7.1	22.7
Bilen *et al.* [6]	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
Wang *et al.* [30]	48.9	42.3	26.1	11.3	11.9	41.3	40.9	34.7	10.8	34.7	18.8	34.4	35.4	52.7	19.1	17.4	35.9	33.3	34.8	46.5	31.6
Bilen *et al.* [5]	39.4	50.1	31.5	16.3	12.6	64.5	42.8	42.6	10.1	35.7	24.9	38.2	34.4	55.6	9.4	14.7	30.2	40.7	54.7	46.9	34.8
Kantorov *et al.* [14]	57.1	52.0	31.5	7.6	11.5	55.0	53.1	34.1	1.7	33.1	49.2	42.0	47.3	56.6	15.3	12.8	24.8	48.9	44.4	47.8	36.3
Li *et al.* [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
Jie *et al.* [12]	52.2	47.1	35.0	26.7	15.4	61.3	66.0	54.3	3.0	53.6	24.7	43.6	48.4	65.8	6.6	18.8	51.9	43.6	53.6	62.4	41.7
Ours	49.1	53.6	43.5	21.3	18.5	66.9	64.0	55.6	11.9	53.7	26.6	45.6	48.7	64.6	20.4	23.3	50.0	44.7	55.9	60.6	43.9

Table 1: Comparison of detection results (mAP %) on 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	mean
Cinbis *et al.* [9]	57.2	62.2	50.9	37.9	23.9	64.8	74.4	24.8	29.7	64.1	40.8	37.3	55.6	68.1	25.5	38.5	65.2	35.8	56.6	33.5	47.3
Bilen *et al.* [3]	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
Wang *et al.* [30]	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
Bilen *et al.* [3]	65.1	58.8	58.5	33.1	39.8	68.3	60.2	59.6	34.8	64.5	30.5	43.0	56.8	82.4	25.5	41.6	61.5	55.9	65.9	63.7	53.5
Kantorov *et al.* [13]	83.3	68.6	54.7	23.4	18.3	73.6	74.1	54.1	8.6	65.1	47.1	59.5	67.0	83.5	35.3	39.9	67.0	49.7	63.5	65.2	55.1
Li *et al.* [12]	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
Jie *et al.* [11]	72.7	55.3	53.0	27.8	35.2	68.6	81.9	60.7	11.6	71.6	29.7	54.3	64.3	88.2	22.2	53.7	72.2	52.6	68.9	75.5	56.1
Ours	75.9	67.6	62.2	37.3	36.6	71.5	80.2	63.8	19.7	70.6	32.4	56.1	67.8	81.7	35.9	50.9	73.4	50.4	66.0	66.8	58.3

Table 2: Comparison of localization results (CorLoc %) on PASCAL VOC 2007 trainval set.

4.2 PASCAL VOC Results

Detection and Localization. Comparisons of our model with recent state-of-the-art weakly supervised methods are shown in Table 1 and Table 2. The proposed model outperforms all other alternatives in both mAP and CorLoc. Considerable improvements are obtained on categories that often appear in groups, *e.g.*, bird and person. The segmentation procedure makes the detected bounding boxes more accurate for individual instances, especially for small objects. The model also shows robust localization performance. The results of cat, horse, and person are at a high level; the introduction of partial aggregation strategy reduces the false localization of focusing on discriminative parts (i.e., faces).

Classification. Our primary goal is to train an object detector, but the detection performance basically depends on the classification capacity. The backbone VGG-D is originally trained to perform single-object image classification. Our multi-label adaptation phase indeed trains a multi-label image classifier. Table 3 shows the classification results. The original VGG-D achieved mAP of 89.3%, and we further gain an improvement of 4pt. We noticed that the idea of HCP-VGG is a little similar to our framework (while HCP-VGG is trained only for classification). However, during image fine-tuning, we use many sub-images instead of only the original image. And for proposal (or called hypothesis in [31]) fine-tuning, we also adopt much more proposals than HCP, and more importantly, we train an object detector instead of only classifier.

Method	mAP (%)
*VGG-D** [27]	*89.3*
HCP-VGG [31]	90.9
WSDDN [3]	89.7
WELDON [11]	90.2
Ours	91.3

Table 3: Comparison of classification results (mAP) on PASCAL VOC 2007.

	mAP(%)	CorLoc(%)
$m = 1$	38.7	55.7
$m = 5$	40.6	56.3
$m = 10$	38.9	54.9
Ensemble	43.9	58.3

Table 4: The results of our model using different number of sub-images .

4.3 Analysis

Number of sub-images. The main hyperparameter of our model is the number of segmentations m for Ncut algorithm. The poor-proposal filtering procedure is mainly for refinement,

while the Ncut fundamentally determines the segmentation quality. Generally, a desirable segmentation should satisfy two principal requirements:

- *Integrity of object instances:* For image classification, the model learns statistical features of object parts. It gives a correct classification as long as the discriminative parts are identified. The relative positions are not so important. However, for object localization and detection, the goal is to localize the whole content of object instances. It requires the segmented sub-images cover all object content as completely as possible. If the number of sub-images m is too large, the objects are likely to be cut off.
- *High recall for individual object instances:* A multi-label image may contain many objects. If m is too small, multiple object instances will crowd in one sub-image, which affects the subsequent discrimination. Furthermore, the PASCAL VOC dataset contains only 20 categories, hence there are many unlabeled objects, e.g. chandeliers and windows. These objects also tend to be densely covered by plenty of proposals. The number of sub-images m should be large enough to cover all object instances.

We vary the number of segmentation to examine its effect, and the results are shown in Table 4. For single model, the highest mAP of 40.6% is obtained using 5 sub-images. When $m = 1$, the input images are actually not segmented. It can be viewed as a baseline and achieves an mAP of 38.7%. For more sub-images (e.g. 10), the mAP instead drops. This is due to the over-segmentation that causes objects cut off. The best CorLoc of 56.3% is also achieved when $m = 5$. The ensemble item brings together all the detections, hence further boosts both mAP and CorLoc.

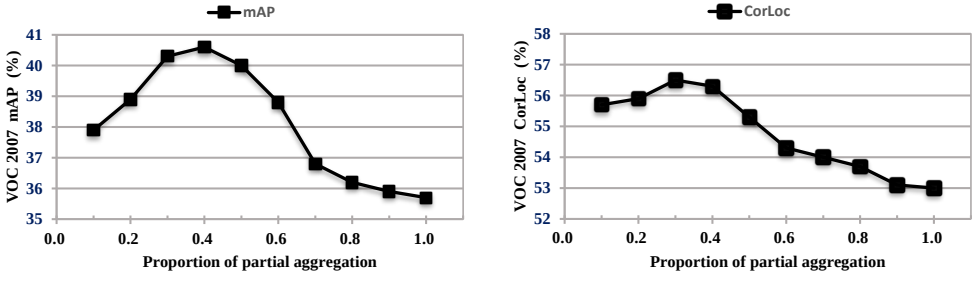

Figure 2: The mAP and CorLoc results using different proportion of partial aggregation.

Proportion of partial aggregation. The partial aggregation strategy is designed to select a portion of scores dynamically. It acts as regularization to encourage the learning of object content. Take the best single model ($m = 5$), we analyze the effect of proportion as shown in Fig. 2. In our experiments, we adopt the setting of $|\mathcal{P}^+| = |\mathcal{P}^-|$. The horizontal axes represent the total proportion of selected scores. Obviously, the mAP and CorLoc are comparatively better when fewer than 50% are adopted. We can figure out a considerable improvement of $\sim$5pt and $\sim$3pt for mAP and CorLoc respectively. So the proportion of $\mathcal{P}^+$ and $\mathcal{P}^-$ are both set to 20% throughout all our experiments, but the performance should be further improved by careful tuning for different number m of sub-images.

4.4 Qualitative Examples

Some Qualitative examples are shown in Fig. 3. It can be observed that our method could localize single objects accurately in complex background. The detections cover the majority

of object content instead of only the most discriminative parts. For images that contain multiple instances, most instances are detected even for small objects. However, there are several notable failure cases. The front-facing plane and the sofa is falsely localized. This can be explained that their outlines are too complex to determine the exact extent. Some visual-similar or scattered objects are mis-detected, e.g., plants. This failures should be reduced by other backbones with stronger discriminative capacity.

Figure 3: Qualitative examples on PASCAL VOC 2007 test set. Green and red bounding boxes represent correct and false detections respectively.

5 Conclusion

In this paper, we proposed two major approaches: image segmentation framework and partial aggregation strategy, aiming at addressing the common failures in weakly supervised object detectors. We implemented a two-phase adaptation CNN model to demonstrate our design. The backbone network is modified with a novel adaptive fully-connected layer to deal with the variable-size input. Experimental results show that our model outperforms other state-of-the-art methods in many visual recognition tasks. Future works include incorporating adaptive image segmentation algorithms for more flexible workflow.

6 Acknowledgment

This work was jointly supported by: (1) National Natural Science Foundation of China (No. 61771068, 61671079, 61471063, 61372120, 61421061); (2) Beijing Municipal Natural Science Foundation (No.4182041, 4152039); (3) the National Basic Research Program of China (No. 2013CB329102).

References

[1] B. Alexe, T. Deselaers, and V. Ferrari. Measuring the objectness of image windows. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(11):2189–2202,

Nov 2012. ISSN 0162-8828. doi: 10.1109/TPAMI.2012.28.

[2] P. Arbeláez, J. Pont-Tuset, J. Barron, F. Marques, and J. Malik. Multiscale combinatorial grouping. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 328–335, June 2014. doi: 10.1109/CVPR.2014.49.

[3] H. Bilen and A. Vedaldi. Weakly supervised deep detection networks. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2846–2854, June 2016. doi: 10.1109/CVPR.2016.311.

[4] H. Bilen, M. Pedersoli, and T. Tuytelaars. Weakly supervised object detection with convex clustering. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1081–1089, June 2015. doi: 10.1109/CVPR.2015.7298711.

[5] Hakan Bilen, Marco Pedersoli, and Tinne Tuytelaars. Weakly supervised detection with posterior regularization. In *Proceedings of the British Machine Vision Conference BMVC*, 2014. doi: 10.5244/c.28.52.

[6] M. M. Cheng, Z. Zhang, W. Y. Lin, and P. Torr. Bing: Binarized normed gradients for objectness estimation at 300fps. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 3286–3293, June 2014. doi: 10.1109/CVPR.2014.414.

[7] R. G. Cinbis, J. Verbeek, and C. Schmid. Weakly supervised object localization with multi-fold multiple instance learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(1):189–203, Jan 2017. ISSN 0162-8828. doi: 10.1109/TPAMI.2016.2535231.

[8] Thomas Deselaers, Bogdan Alexe, and Vittorio Ferrari. Localizing objects while learning their appearance. In *Proceedings of the 11th European Conference on Computer Vision (ECCV)*, volume 6314, pages 452–466, 2010. doi: 10.1007/978-3-642-15561-1_33.

[9] Thomas Deselaers, Bogdan Alexe, and Vittorio Ferrari. Weakly supervised localization and learning with generic knowledge. *International Journal of Computer Vision*, 100: 275–293, 2012.

[10] A. Diba, V. Sharma, A. Pazandeh, H. Pirsiavash, and L. V. Gool. Weakly supervised cascaded convolutional networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5131–5139, July 2017. doi: 10.1109/CVPR.2017.545.

[11] T. Durand, N. Thome, and M. Cord. Weldon: Weakly supervised learning of deep convolutional neural networks. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4743–4752, June 2016. doi: 10.1109/CVPR.2016.513.

[12] T. Durand, T. Mordan, N. Thome, and M. Cord. Wildcat: Weakly supervised learning of deep convnets for image classification, pointwise localization and segmentation. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5957–5966, July 2017. doi: 10.1109/CVPR.2017.631.

[13] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2007 (VOC2007) Results. http://www.pascal-network.org/challenges/VOC/voc2007/workshop/index.html.

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

[15] R. Girshick. Fast r-cnn. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 1440–1448, Dec 2015. doi: 10.1109/ICCV.2015.169.

[16] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, June 2016. doi: 10.1109/CVPR.2016.90.

[17] Z. Jie, Y. Wei, X. Jin, J. Feng, and W. Liu. Deep self-taught learning for weakly supervised object localization. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4294–4302, July 2017. doi: 10.1109/CVPR.2017.457.

[18] Z. Jie, Y. Wei, X. Jin, J. Feng, and W. Liu. Deep self-taught learning for weakly supervised object localization. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4294–4302, July 2017. doi: 10.1109/CVPR.2017.457.

[19] Vadim Kantorov, Maxime Oquab, Minsu Cho, and Ivan Laptev. Contextlocnet: Context-aware deep network models for weakly supervised localization. In *Proceedings of the 14th European Conference on Computer Vision (ECCV)*, volume 9909, pages 350–365, 2016. doi: 10.1007/978-3-319-46454-1_22.

[20] Vadim Kantorov, Maxime Oquab, Minsu Cho, and Ivan Laptev. Contextlocnet: Context-aware deep network models for weakly supervised localization. In *Proceedings of the 14th European Conference on Computer Vision (ECCV)*, volume 9909, pages 350–365, 2016. doi: 10.1007/978-3-319-46454-1_22.

[21] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. ImageNet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6):84–90, may 2017. doi: 10.1145/3065386.

[22] Baisheng Lai and Xiaojin Gong. Saliency guided end-to-end learning for weakly supervised object detection. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence*, aug 2017. doi: 10.24963/ijcai.2017/285.

[23] D. Li, J. B. Huang, Y. Li, S. Wang, and M. H. Yang. Weakly supervised object localization with progressive domain adaptation. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3512–3520, June 2016. doi: 10.1109/CVPR.2016.382.

[24] M. Oquab, L. Bottou, I. Laptev, and J. Sivic. Learning and transferring mid-level image representations using convolutional neural networks. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1717–1724, June 2014. doi: 10.1109/CVPR.2014.222.

[25] M. Oquab, L. Bottou, I. Laptev, and J. Sivic. Is object localization for free? - weakly-supervised learning with convolutional neural networks. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 685–694, June 2015. doi: 10.1109/CVPR.2015.7298668.

[26] Jianbo Shi and J. Malik. Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(8):888–905, Aug 2000. ISSN 0162-8828. doi: 10.1109/34.868688.

[27] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556, 2014.

[28] Hyun Oh Song, Ross B. Girshick, Stefanie Jegelka, Julien Mairal, Zaïd Harchaoui, and Trevor Darrell. On learning to localize objects with minimal supervision. In *Proceedings of the 31th International Conference on Machine Learning ICML*, volume 32, pages 1611–1619, 2014.

[29] Jasper R. R. Uijlings, Koen E. A. van de Sande, Theo Gevers, and Arnold W. M. Smeulders. Selective search for object recognition. *International Journal of Computer Vision*, 104:154–171, 2013.

[30] C. Wang, K. Huang, W. Ren, J. Zhang, and S. Maybank. Large-scale weakly supervised object localization via latent category learning. *IEEE Transactions on Image Processing*, 24(4):1371–1385, April 2015. ISSN 1057-7149. doi: 10.1109/TIP.2015.2396361.

[31] Y. Wei, W. Xia, M. Lin, J. Huang, B. Ni, J. Dong, Y. Zhao, and S. Yan. Hcp: A flexible cnn framework for multi-label image classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(9):1901–1907, Sept 2016. ISSN 0162-8828. doi: 10.1109/TPAMI.2015.2491929.

[32] C. Lawrence Zitnick and Piotr Dollár. Edge boxes: Locating object proposals from edges. In *Proceedings of the 13th European Conference on Computer Vision (ECCV)*, volume 8693, pages 391–405, 2014. doi: 10.1007/978-3-319-10602-1_26.