Towards Precise End-to-end Weakly Supervised Object Detection Network

Ke Yang  Dongsheng Li  Yong Dou
National University of Defense Technology
yangke13@nudt.edu.cn

Abstract

It is challenging for weakly supervised object detection network to precisely predict the positions of the objects, since there are no instance-level category annotations. Most existing methods tend to solve this problem by using a two-phase learning procedure, i.e., multiple instance learning detector followed by a fully supervised learning detector with bounding-box regression. Based on our observation, this procedure may lead to local minima for some object categories. In this paper, we propose to jointly train the two phases in an end-to-end manner to tackle this problem. Specifically, we design a single network with both multiple instance learning and bounding-box regression branches that share the same backbone. Meanwhile, a guided attention module using classification loss is added to the backbone for effectively extracting the implicit location information in the features. Experimental results on public datasets show that our method achieves state-of-the-art performance.

1. Introduction

In recent years, Convolutional Neural Networks (CNN) approaches have achieved great success in computer vision field, due to its ability to learn generic visual features that can be applied in many tasks such as image classification [20, 31, 12], object detection [10, 9, 26] and semantic segmentation [23, 2]. Fully supervised object detection has been widely studied and achieved promising results. There are also plenty of public datasets which provide precise location and category annotations of the objects. However, precise object-level annotations are always expensive in human resource and huge data volume is required by training accurate object detection models. In this paper, we focus on Weakly Supervised Object Detection (WSOD) problem, which uses only image-level category labels so that significant cost of preparing training data can be saved. Due to the lack of accurate annotations, this problem has not been well handled and the performance is still far from the fully supervised methods.

Recent WSOD methods [5, 1, 34, 22, 18] usually follows a two-phase learning procedure as shown in the top part of Figure 1. In the first phase, the Multiple Instance Learning (MIL) [4, 18, 34, 1] like weakly learning pipeline is used, which trains a MIL detector by using CNN as feature extractor. In the second phase, a fully supervised detector, e.g. Fast R-CNN [9] or Faster R-CNN [26], is trained to further refine object location by using the selected proposals of the first phase as supervision. The main functionality of the second phase is to regress the object locations more precisely. However, we observed that the two-phase learning is easy to get stuck into local minima if the selected proposals of the first phase are too far from real Ground Truth (GT). As shown in the top part of Figure 1, in some categories, the MIL detector tends to focus on the local discriminative parts of the objects, such as the head of a cat,
so that the wrong proposals are used as pseudo GT for the second phase. In this case, the accurate location of the object can hardly be learned in the regression process of the second phase, as the MIL detector has already over-fitted seriously to the discriminate parts, as shown in the middle part of Figure 2.

We further observed that the MIL detector does not select the most discriminative parts at the beginning of the training, but gradually over-fits to these parts, as shown in the left part of Figure 2.

Taking into account the above observations, we propose to jointly train the MIL detector and the bounding-box regressor together in an end-to-end manner, as shown in the bottom part of Figure 1. In this manner, the regressor is able to start to adjust the predicted boxes before the MIL detector focuses seriously to small discriminative parts, as shown in the right part of Figure 2. Specifically, we use MIL detection scheme [1, 34] as baseline and integrate fully supervised RoI-based classification and bounding-box regression branch similar to Fast R-CNN, which shares the same backbone with MIL detector. MIL detector is a weakly learning process, which selects object predictions from the region proposals, e.g. generated by Selective Search Windows (SSW) [36] method, according to classification scores. These selected proposals are then used as the pseudo GT supervision of the classification and regression branch.

In order to further enhance the localization ability of the proposed network, we propose to use a guided attention module using image-level classification loss in the backbone. To our best knowledge, the well trained classification network contains rich object location information. Therefore, we add this attention branch which is guided by image-level classification loss. Fully considering the global characteristics of the objects, the attention branch can improve the discriminative ability of the network as well as detection accuracy.

It is worth noting that though jointly learning of classification and boxes regression has already been shown to be beneficial for fully supervised object detection, for weakly supervised object detection it is still non-trivial and needs innovative idea and insight on this task. Although Our method is conceptually simple in form, it significantly alleviates the weak detector over-fitting to discriminate parts and substantially surpasses previous methods. Our contributions can be summarized as follows.

- We design a single end-to-end weakly supervised object detection network that can jointly optimize the region classification and regression, which boosts performance significantly.
- We design a classification guided attention module to enhance the localization ability of feature learning, which also leads to a noteworthy improvement.
- Our proposed network significantly outperforms previous state-of-the-art weakly supervised object detection approaches on PASCAL VOC 2007 and 2012.

2. Related Work
2.1. Convolutional Feature Extraction

After the success of using CNNs for image classification task[20], a research stream based on CNNs [10, 29] shows significant improvements in detection performance. These methods use convolutional layers to extract features from each region proposal. To speed up the detection, SPP-Net [11] and Fast R-CNN [9] firstly extract region-independent feature maps at the full-image level, and then pool region-wise features via spatial extents of proposals.

2.2. Weakly Supervised Object Detection

Most existing methods formulate weakly-supervised detection as a multiple instance learning problem [1, 32, 13, 18, 22, 27]. These approaches divided training images into positive and negative parts, where each image is con-
sidered as a bag of candidate object instances. If an image is annotated as a positive sample of a specific object class, at least one proposal instance of the image belongs to this class. The main task of MIL-based detectors is to learn the discriminative representation of the object instances and then select them from positive images to train a detector. Previous works on applying MIL to WSOD can be roughly categorized into multi-phase learning approach [18, 4, 22, 38, 30, 42, 43, 41] and end-to-end learning approach [1, 39, 19, 33].

End-to-end learning approaches combine CNNs and MIL into a unified network to address weakly supervised object detection task. Diba et al. [5] proposed an end-to-end cascaded convolutional network to perform weakly supervised object detection and segmentation in cascaded manner. Bilen et al. [1] developed a two-stream weakly supervised deep detection network (WSDNN), which selected the positive samples by aggregating the score of classification stream and detection stream. Based on WSDNN, Kantorov et al. [19] proposed to learn a context-aware CNN with contrast-based contextual modeling. Also based on WSDNN, Tang et al. [34] designed an online instance classifier refinement (OICR) algorithm to alleviate the local optimum problem. Tang et al. [33] also proposed Proposal Cluster Learning (PCL) to improve the performance of OICR. Following the inspiration of [19] and [5], Wei et al. [39] proposed a tight box mining method that leverages surrounding segmentation context derived from weakly-supervised segmentation to suppress low quality distracting candidates and boost the high-quality ones. Recently, Tang et al. [35] proposed a weakly supervised region proposal network to generate more precise proposals for detection. Positive object instances often focus on the most discriminative parts of an object (e.g. the head of a cat, etc.) but not the whole object, which leads to inferior performance of weakly supervised detectors.

Multi-phase learning approaches first employ MIL to select the best object candidate proposals, then use these selected proposals as pseudo GT annotations for learning the fully supervised object detector such as R-CNN [10] or Fast(er) R-CNN [9, 26]. Li et al. [22] proposed classification adaptation to fine-tune the network to collect class specific object proposals, and detection adaptation was used to optimize the representations for the target domain by the confident object candidates. Cinbis et al. [4] proposed a multi-fold MIL detector by re-labeling proposals and retraining the object classifier iteratively to prevent the detector from being locked into wrong object locations. Jie et al. [18] proposed a self-taught learning approach to progressively harvest high-quality positive instances. Zhang et al. [43] proposed pseudo ground-truth excavation (PGE) algorithm and pseudo groundtruth adaptation (PGA) algorithm to refine the pseudo ground-truth obtained by [34]. Wan et al. [38] proposed a min-entropy latent model (MELM) and recurrent learning algorithm for weakly supervised object detection. Ge et al. [8] proposed to fuse and filter object instances from different techniques and perform pixel labeling with uncertainty and they used the resulting pixelwise labels to generate groundtruth bounding boxes for object detection and attention maps for multi-label classification. Zhang et al. [42] proposed a Multi-view Learning Localization Network (ML-LocNet) by incorporating multiview learning into a two-phase WSOD model. However, multi-phase learning WSOD is a non-convex optimization problem, which makes such approaches trapped in local optima.

In this paper, we consider the MIL (positive object candidates mining) and regression (object candidates localization refinement) problems simultaneously. We follow the MIL pipeline and combine the two-stream WSDDN [1] and OICR/PCL algorithms [34, 33] to implement our basic MIL branch and refine the detected boxes with a regression branch in an online manner.

2.3. Attention Module

Attention modules were first used in the natural language processing field and then introduced to the computer vision area. Attention can be seen as a method of biasing the allocation of available computational resources towards the most informative components of a signal [15, 16, 25, 21, 37, 24, 14].

The current attention modules can be divided into two categories: spatial attention and channel-wise attention. Spatial attention is to assign different weights to different spatial regions depending on their feature content. It automatically predicts the weighted heat map to enhance the relevant features and suppress the irrelevant features during the training process of a specific task. Spatial attention has been used in image captioning [40], multi-label classification [45], pose estimation [3] and so on. Hu et al. [14] proposed an Squeeze-and-Excitation block which models channel-wise attention in a computationally efficient manner. In this paper, we use a combination of spatial and channel-wise attention, and our attention module is guided by object category.

3. Method

In this section we introduce proposed weakly supervised object detection network, which consists of three major components: guided attention module (GAM), MIL branch and regression branch. The overall architecture of proposed network is shown in Figure 3. Given an input image, an enhanced feature map is first extracted from the CNN network with GAM. Region features generated by ROI pooling are then sent to MIL branch and regression branch. The object locations and categories proposed by MIL branch are taken as pseudo GT of the regression branch for location.
regression and classification. The remainder of this section discusses the three components in detail.

### 3.1. Guided Attention Module

First, we describe the conventional spatial neural attention structure. Given a feature map \( \mathbf{X} \in \mathbb{R}^{H \times W \times D} \) extracted from a ConvNet, the attention module takes it as input and outputs a spatial-normalized attention weight map \( \mathbf{A} \in \mathbb{R}^{H \times W} \) via a \( 1 \times 1 \) convolutional layer. Attention map is then multiplied to \( \mathbf{X} \) to get attended feature \( \mathbf{X}_a \in \mathbb{R}^{H \times W \times D} \). \( \mathbf{X}_a \) is added to \( \mathbf{X} \) to get the enhanced feature map \( \hat{\mathbf{X}} \). After that, \( \hat{\mathbf{X}} \) is fed to subsequent modules. Attention map \( \mathbf{A} \) acts as a spatial regularizer to enhance the relevant regions and suppress the non-relevant regions for feature \( \mathbf{X} \).

Formally, attention module consists of a convolutional layer, a non-linear activation layer and a spatial normalization as follows:

\[
z_{i,j} = F \left( w^T \mathbf{x}_{i,j} + b \right),
\]

\[
a_{i,j} = \frac{z_{i,j}}{\sum_{i,j} z_{i,j}},
\]

where \( F \) is non-linear activation function, \( w \) and \( b \) are the parameters of the attention module, which is a \( 1 \times 1 \) convolutional layer. The attended feature \( \hat{\mathbf{x}}_{i,j} \) can be calculated by:

\[
\hat{\mathbf{x}}_{i,j} = (1 + a_{i,j}) \mathbf{x}_{i,j}.
\]

The conventional attention map is class-agnostic. We hope it can learn some foreground/background information to help figure out the position of the objects, because it has been proved that CNNs are not only effective at predicting the class label of an image, but also localizing the image regions relevant to this label [44].

We add the classification loss to guide the learning of the attention weights. To achieve this, we expand spatial attention to both spatial and channel attention. Specifically, attention map are changed from \( \mathbf{A} \in \mathbb{R}^{1 \times 1 \times D} \) to \( \mathbf{A} \in \mathbb{R}^{H \times W \times D} \). The attention module can be formalized as:

\[
z_{i,j}^c = F \left( w^T \mathbf{x}_{i,j}^c + b \right),
\]

\[
a_{i,j}^c = \frac{z_{i,j}^c}{1 + \exp \left( -z_{i,j}^c \right)},
\]

where \( c \) denotes the value of the \( c \)-th channel. The attended feature \( \hat{\mathbf{x}}_{i,j}^c \) can be calculated by:

\[
\hat{\mathbf{x}}_{i,j}^c = (1 + a_{i,j}^c) \mathbf{x}_{i,j}^c.
\]

To introduce classification supervision to attention weights learning, attention map \( \mathbf{A} \) is also fed to another convolutional layer and a Global Average Pooling (GAP) layer to get the classification score vector. Then the attention map can be supervised by the standard multi-label classification loss. The enhanced feature map \( \hat{\mathbf{X}} \) is fed to subsequent components for detection.

### 3.2. MIL Branch

We only have image-level labels indicating whether an object category appears. To train a standard object detector with regression, it is necessary to mine instance-level supervision such as bounding-box annotations. Therefore, we need to introduce a MIL branch to initialize the pseudo GT annotations. There are a couple of possible choices such as [1, 4, 34]. We choose to adopt OICR network...
3.3. Multi-Task Branch

After pseudo GT annotations are generated, a multi-task branch can operate fully supervised classification and regression as Fast R-CNN [9]. The detection branch has two sibling branches. The first branch predicts a discrete probability distribution (per RoI), \( p \in \mathbb{R}^{(C+1) \times 1} \), over \( C+1 \) categories, which is computed by a softmax over the \( C+1 \) outputs of a FC layer. The second sibling branch outputs bounding-box regression offsets, \( t^r = (t_{cx}^r, t_{cy}^r, t_{cw}^r, t_{ch}^r) \) for each of the \( C \) object classes, indexed by \( c \).

Since we get the instance annotations from MIL branch as introduced in Section 3.2, each RoI now has a GT classification target \( v \) and GT classification target \( u \). We use a multi-task loss \( L_{det} \) of all labeled RoIs for classification and bounding-box regression:

\[
L_{det} = L_{cls} + \lambda L_{loc},
\]

where \( L_{cls} \) is classification loss, and \( L_{loc} \) is regression loss. \( \lambda \) controls the balance between two losses. For \( L_{loc} \), smooth \( L_1 \) loss is used. For \( L_{cls} \), since the pseudo GT annotations are noisy, we add a weight \( w^r \) with respect to RoI \( r \):

\[
L_{cls} = \frac{1}{|R|} \sum_{r=1}^{C+1} \sum_{c=1}^{C} w^r u^r_c \log p^r_c,
\]

where \( |R| \) is the number of proposals. The weight \( w^r \) is calculated following the weights calculation method in [34] when refining the classifiers.

The overall network is trained by optimizing the following composite loss functions from the four components using stochastic gradient descent:

\[
L = L_{img,cls} + L_{mit} + L_{refine} + L_{det},
\]

where \( L_{img,cls} \) is the multi-label classification loss of GAM; \( L_{mit} \) is the multi-label classification loss of WSDNN; \( L_{refine} \) is the classifier refinement loss; and \( L_{det} \) is multi-task loss of the detection sub-network.

4. Experiments

In this section, we first introduce the evaluation datasets and the implementation details of our approach. Then we
Table 3: Comparison of AP performance (%) on PASCAL VOC 2007 test. The upper part shows results by single-end-to-end model. The lower part shows results by multi-phase approaches or ensemble model.

<table>
<thead>
<tr>
<th>Methods</th>
<th>mAP</th>
</tr>
</thead>
<tbody>
<tr>
<td>ours</td>
<td>45.3</td>
</tr>
<tr>
<td>MIL-oicr-gam+res(Ours)</td>
<td>40.6</td>
</tr>
<tr>
<td>MIL-ocl-gam+res(ours)</td>
<td>39.5</td>
</tr>
<tr>
<td>MIL-occl-gam+res(ours)</td>
<td>37.5</td>
</tr>
<tr>
<td>MIL-occr-gam+res(ours)</td>
<td>35.5</td>
</tr>
<tr>
<td>MIL-occr-gam+res(ours)</td>
<td>33.5</td>
</tr>
<tr>
<td>MIL-occr-gam+res(ours)</td>
<td>31.5</td>
</tr>
<tr>
<td>MIL-occr-gam+res(ours)</td>
<td>29.5</td>
</tr>
<tr>
<td>MIL-occr-gam+res(ours)</td>
<td>27.5</td>
</tr>
<tr>
<td>MIL-occr-gam+res(ours)</td>
<td>25.5</td>
</tr>
<tr>
<td>MIL-occr-gam+res(ours)</td>
<td>23.5</td>
</tr>
<tr>
<td>MIL-occr-gam+res(ours)</td>
<td>21.5</td>
</tr>
<tr>
<td>MIL-occr-gam+res(ours)</td>
<td>19.5</td>
</tr>
<tr>
<td>MIL-occr-gam+res(ours)</td>
<td>17.5</td>
</tr>
<tr>
<td>MIL-occr-gam+res(ours)</td>
<td>15.5</td>
</tr>
<tr>
<td>MIL-occr-gam+res(ours)</td>
<td>13.5</td>
</tr>
<tr>
<td>MIL-occr-gam+res(ours)</td>
<td>11.5</td>
</tr>
<tr>
<td>MIL-occr-gam+res(ours)</td>
<td>9.5</td>
</tr>
<tr>
<td>MIL-occr-gam+res(ours)</td>
<td>7.5</td>
</tr>
<tr>
<td>MIL-occr-gam+res(ours)</td>
<td>5.5</td>
</tr>
<tr>
<td>MIL-occr-gam+res(ours)</td>
<td>3.5</td>
</tr>
<tr>
<td>MIL-occr-gam+res(ours)</td>
<td>1.5</td>
</tr>
<tr>
<td>MIL-occr-gam+res(ours)</td>
<td>0.5</td>
</tr>
<tr>
<td>MIL-occr-gam+res(ours)</td>
<td>0.01</td>
</tr>
</tbody>
</table>

Table 4: Comparison of AP performance (%) on PASCAL VOC 2012 test. The upper part shows results by single-end-to-end model. The lower part shows results by multi-phase approaches or ensemble model.

<table>
<thead>
<tr>
<th>Methods</th>
<th>mAP</th>
</tr>
</thead>
<tbody>
<tr>
<td>ours</td>
<td>45.3</td>
</tr>
<tr>
<td>MIL-oicr-gam+res(Ours)</td>
<td>40.6</td>
</tr>
<tr>
<td>MIL-ocl-gam+res(ours)</td>
<td>39.5</td>
</tr>
<tr>
<td>MIL-occl-gam+res(ours)</td>
<td>37.5</td>
</tr>
<tr>
<td>MIL-occr-gam+res(ours)</td>
<td>35.5</td>
</tr>
<tr>
<td>MIL-occr-gam+res(ours)</td>
<td>33.5</td>
</tr>
<tr>
<td>MIL-occr-gam+res(ours)</td>
<td>31.5</td>
</tr>
<tr>
<td>MIL-occr-gam+res(ours)</td>
<td>29.5</td>
</tr>
<tr>
<td>MIL-occr-gam+res(ours)</td>
<td>27.5</td>
</tr>
<tr>
<td>MIL-occr-gam+res(ours)</td>
<td>25.5</td>
</tr>
<tr>
<td>MIL-occr-gam+res(ours)</td>
<td>23.5</td>
</tr>
<tr>
<td>MIL-occr-gam+res(ours)</td>
<td>21.5</td>
</tr>
<tr>
<td>MIL-occr-gam+res(ours)</td>
<td>19.5</td>
</tr>
<tr>
<td>MIL-occr-gam+res(ours)</td>
<td>17.5</td>
</tr>
<tr>
<td>MIL-occr-gam+res(ours)</td>
<td>15.5</td>
</tr>
<tr>
<td>MIL-occr-gam+res(ours)</td>
<td>13.5</td>
</tr>
<tr>
<td>MIL-occr-gam+res(ours)</td>
<td>11.5</td>
</tr>
<tr>
<td>MIL-occr-gam+res(ours)</td>
<td>9.5</td>
</tr>
<tr>
<td>MIL-occr-gam+res(ours)</td>
<td>7.5</td>
</tr>
<tr>
<td>MIL-occr-gam+res(ours)</td>
<td>5.5</td>
</tr>
<tr>
<td>MIL-occr-gam+res(ours)</td>
<td>3.5</td>
</tr>
<tr>
<td>MIL-occr-gam+res(ours)</td>
<td>1.5</td>
</tr>
<tr>
<td>MIL-occr-gam+res(ours)</td>
<td>0.5</td>
</tr>
<tr>
<td>MIL-occr-gam+res(ours)</td>
<td>0.01</td>
</tr>
</tbody>
</table>

4.1. Datasets and Evaluation Metrics

We evaluate our method on the popular PASCAL VOC 2007 and 2012 datasets [6] which have 9966 and 22531 images for 20 object classes, respectively. These two datasets are split into train, validation, and test sets. We use the trainval set (5011 images for 2007 and 11540 for 2012) for training. As we focus on weakly supervised detection, only image-level labels are utilized during training. Average Precision (AP) and the mean of AP (mAP) are taken as the evaluation metrics to test our model on the testing set. Corect localization (CorLoc) is also used to evaluate our model on the test set to measure the localization accuracy [1]. Both metrics are evaluated on the PASCAL criteria, i.e., IoU > 0.5 between ground truth boxes and predicted boxes.

4.2. Implementation Details

We use the object proposals generated by selective search windows [36] and adopt VGG16 [31] pre-trained on ImageNet [28] as the backbone of our proposed network. For the newly added layers, the parameters are randomly initialized with a Gaussian distribution $\mathcal{N}(\mu, \delta)$ ($\mu = 0, \delta = 0.01$) and 10 times learning rate. During training, we adopt a mini-batch size of 2 images, and set the learning rate to 0.001 for the first 40K iterations and then decrease it to
Table 5: Comparison of correct localization (CorLoc) (%) of single end-to-end model on PASCAL VOC 2007 trainval.

<table>
<thead>
<tr>
<th>Methods</th>
<th>milAP</th>
<th>mAP</th>
</tr>
</thead>
<tbody>
<tr>
<td>MIL-OICR+GAM+REG(Ours)</td>
<td>81.7</td>
<td>81.2</td>
</tr>
<tr>
<td>MIL-PCL+GAM+REG(Ours)</td>
<td>80.0</td>
<td>80.3</td>
</tr>
</tbody>
</table>

Table 6: Comparison of correct localization (CorLoc) (%) of single end-to-end model on PASCAL VOC 2012 trainval.

<table>
<thead>
<tr>
<th>Methods</th>
<th>milAP</th>
<th>mAP</th>
</tr>
</thead>
<tbody>
<tr>
<td>MIL-OICR+GAM+REG(Ours)</td>
<td>82.4</td>
<td>82.0</td>
</tr>
<tr>
<td>MIL-PCL+GAM+REG(Ours)</td>
<td>80.2</td>
<td>80.1</td>
</tr>
</tbody>
</table>

0.0001 in the following 30K iterations. The momentum and weight decay are set to 0.9 and 0.0005, respectively. We use five image scales, i.e., \{480, 576, 688, 864, 1200\}, and horizontal flips for both training and testing data augmentation. During testing, we use the mean output of the regression branch, including classification scores and bounding boxes, as the final results. Our experiments are based on the deep learning framework of Caffe [17]. All of the experiments run on NVIDIA GTX 1080 Ti GPUs.

4.3. Ablation Studies

We conduct ablation experiments on PASCAL VOC 2007 to prove the effectiveness of our proposed network. We validate the contribution of each component including GAM and regression branch.

4.3.1 Baseline

The baseline is the MIL detector without GAM and regression branch that we introduced in Section 3.1, which is the same as OICR [34]. We re-run the experiment and get a slightly higher result of 41.3\% mAP (41.2\% mAP in [34]).

4.3.2 Guided Attention Module

To verify the effect of GAM, we conduct experiments with and w/o GAM. We denote the network with GAM as MIL+GAM, which does not include regression branch. From Table 1, we can conclude that GAM does help the detector learn better features and improves the accuracy of MIL detector by 2.0\%.

4.3.3 Joint Optimization

To optimize proposal classification and regression jointly, we propose to use bounding-box regression in an online manner together with MIL detection. To verify the effect of online regression, we conduct control experiments under two setting: 1) our joint optimization of MIL detector and regressor, which we denote as MIL+REG; 2) we train a MIL detector first, then use the pseudo GT from the MIL detector to train a fully supervised Fast R-CNN [9]. We denote this setting as MIL+FRCN. The experimental results are summarized in Table 1. From the results, we can see the performance of our MIL+REG is much higher than MIL+FRCN. We attribute the improvements to joint optimization. Separate optimization of MIL detector and regressor result in sub-optimal results. It easily gets stuck in local minima if the pseudo GTs are not accurate. This can be seen from the results of the object category cat and dog. The two object classes are much easier to over-fit to the discriminative parts in the MIL detection. Our joint optimization strategy can alleviate this problem as shown in Figure 2. More visualization results are shown in the supplementary file. We also carry the exploration study on the CorLoc metric, as reported in Table 2. From these results, we can draw the same conclusion.

4.4. Comparison with State-of-the-Art

To fully compare with other methods, we report the results for both “single end-to-end network” and “multi-stage approaches or ensemble model”. The results on VOC 2007 and VOC 2012 are shown in Table 3, Table 5, Table 4 and Table 6. From the tables, we can see that our method achieves the highest performance, outperforming...
It is worth noting that our single model results are even much better than the ensemble models results of most methods which ensemble the results of multiple CNN networks. For example, compared with OICR [34], which we use as baseline, our single model outperforms the ensemble models of OICR significantly while keeping much lower complexity (47.0% mAP Versus 48.6% mAP; 60.6% CorLoc Versus 66.8% CorLoc on VOC 2007). In Figure 4, we also illustrate some detection results by our network as compared to those by our baseline method, i.e., OICR+FRCN. It can be concluded from the illustration that our joint training strategy significantly alleviates the detector focusing on the most discriminative parts.

4.5. Discussion

C-WSL [7] also explored bounding box regression in weakly supervised object detection network. We list the relationship and some differences below. Relationship: We both use bounding box regression in an online manner. However, there are key differences in network architecture between the two, which lead to the performance of C-WSL being much lower than ours, even though they use additional object count labels. Differences: The network structure is different. We use bounding box regression after several box classifier refinements and use only once. C-WSL [7] uses a box regressor together with each box classifier refinement after the MIL branch. Their structure brings two problems. First, a single MIL branch’s classification performance is very poor, it is not wise to directly use the box regressor to refine the box location after the MIL branch. The second problem is that the bounding box regression is used in a cascade manner for each refinement without re-extracting features for the RoIs. Specifically, the subsequent box regression branch should take the refined box locations from the previous box regression branch to update RoIs and re-extracting RoIs features for the classifier and regressor. Because of the above problems, after deducting the improvement of extra label information, their network only improves 1.5% compared with OICR as shown in [7] while our network has increased by 6% compared with OICR (Please note that we use the same set of code released by the authors of OICR). In addition, [7] does not solve the problem of local minima. On the two categories that most affected by the local minima problem, [7] drops 4% in the dog category and improves 3% in the cat category while our method improves 16.3% and 38.6% respectively.

5. Conclusion

In this paper, we present a novel framework for weakly supervised object detection. Different from traditional approaches in this field, our method jointly optimize the MIL detection and regression in an end-to-end manner. Meanwhile, a guided attention module is also added for better feature learning. Experiments show substantial and consistent improvements by our method. Our learning algorithm is potential to be applied in many other weakly supervised visual learning tasks.

Acknowledgements

This work is supported by the National Key Research and Development Program of China under Grant No.2018YFB2101100 and the National Natural Science Foundation of China under Grants 61732018, U1435219 and 61802419.
References


