Human-Like Delicate Region Erasing Strategy for Weakly Supervised Detection

Qing En,[1] Lijuan Duan,[1] Zhaoxiang Zhang,[2*] Xiang Bai,[3] Yundong Zhang[4]
[1]Beijing Key Laboratory of Trusted Computing, Faculty of Information Technology,
Beijing University of Technology, Beijing 100124, China
[2]Center for Research on Intelligent Perception and Computing, National Laboratory of Pattern Reconition,
Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China
[3]School of Electronic Information and Communications, Huazhong University of Science and Technology, Wuhan 430074, China
[4]State Key Lab of Digital Multimedia Chip Technology, Vimicro Corp, Beijing 100191, China
qingen@emails.bjut.edu.cn, ljduan@bjut.edu.cn, zhaoxiang.zhang@ia.ac.cn.
xbai@hust.edu.cn, raymond@vimicro.com

Abstract

We explore a principle method to address the weakly supervised detection problem. Many deep learning methods solve weakly supervised detection by mining various object proposal or pooling strategies, which may cause redundancy and generate a coarse location. To overcome this limitation, we propose a novel human-like active searching strategy that recurrently ignores the background and discovers class-specific objects by erasing undesired pixels from the image. The proposed detector acts as an agent, providing guidance to erase unremarkable regions and eventually concentrating the attention on the foreground. The proposed agents, which are composed of a deep Q-network and are trained by the Q-learning algorithm, analyze the contents of the image features to infer the localization action according to the learned policy. To the best of our knowledge, this is the first attempt to apply reinforcement learning to address weakly supervised localization with only image-level labels. Consequently, the proposed method is validated on the PASCAL VOC 2007 and PASCAL VOC 2012 datasets. The experimental results show that the proposed method is capable of locating a single object within 5 steps and has great significance to the research on weakly supervised localization with a human-like mechanism.

Introduction

Object detection is among the most fundamental and vital research problems in the domain of computer vision. The main issues in object detection lie in constructing a favorable discriminative model and gaining a more compact area. Due to the advancement of convolutional neural networks (He et al. 2016), researchers have made progress on this problem. Supervised learning relies on a large set of training examples with strong supervision. However, acquiring large amounts of strongly supervised labels entails large labor and financial expense. To alleviate the costly annotation requirement, weakly supervised learning methods have been proposed to explore the exploitation of cheaper training labels to complete complex tasks. To address the weakly supervised detection problem, several researchers integrate bottom-up information to acquire class-specific regions of images and treat it as a multiple instance learning (MIL) problem (Song et al. 2014; Oquab et al. 2015). However, although CNN-based models have demonstrated their effectiveness in weakly supervised localization, they do not follow the human visual mechanism, which guides humans to focus on the irregular shapes of objects. Meanwhile, top-down searching cues provide a promising research line, substantially reducing the searching space in detection tasks (Lu, Javidi, and Lazebnik 2016). Attention shifts to the object center when the viewpoint changes, which can be represented by a sequential policy routine. These methods instruct biosystems to integrate feature maps together by combining prior knowledge and the current target.

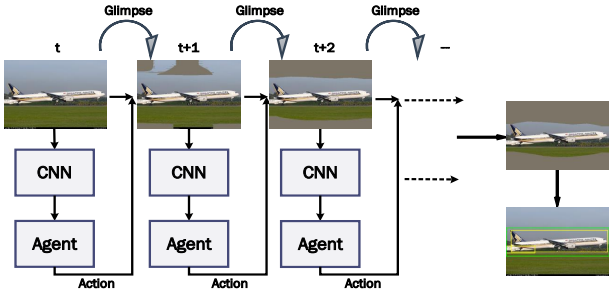

Figure 1: Illustration of the proposed idea. A sequence of glimpsing actions are taken in each time step to locate the object. Undesired regions are erased recurrently. The proposed agents determine which area should be attended in the next time step, and the focusing regions are finally acquired.

Accordingly, exploration of models to facilitate weakly supervised localization in such a manner from both biological and computational perspectives is desirable. Con-

*Corresponding Author (zhaoxiang.zhang@ia.ac.cn).

sequently, three basic observations motivate our research. First, humans search attentional regions according to fovea movement. Second, during visual information processing, the human eye's visual system can select only a few significant information points to process. Third, the visual attentional regions are always shown in an irregular shape, which prevents border confusion with the background.

In this paper, we propose a novel human-like active searching strategy that recurrently ignores the background and discovers class-specific objects by erasing undesired pixels from the image, as is shown in Fig. 1. Our proposed method conforms to the fact that the greater the area of objects that are focused, the better the classification of the image. Specifically, a CNN-based feature extractor is trained by image-level annotation. Then, the features and a discriminating heat map are extracted by the feature extractor. Our proposed agents analyze the contents of the current features to infer the localization action according to the learned policy, which is composed of deep Q-network. As a result, a single object is detected in fewer than 5 steps per image. The experimental results of the proposed method in weakly supervised localization are comparable with those of CNN-based methods and can generate meaningful visualization results. In summary, our work makes the following contributions:

- To the best of our knowledge, this work presents the first deep reinforcement learning solution for weakly supervised object localization.
- We propose a novel human-like active searching strategy to remove undesired regions iteratively by utilizing a deep Q-network.
- The proposed method is validated on two datasets and is shown to be efficient in weakly supervised localization.

Related Work

CNN-Based Weakly Supervised Learning

A beam searching approach is described in (Bency et al. 2016), leveraging local spatial information and semantic patterns to detect multiple objects effectively. Additionally, multiscale fully convolution streams are applied to propose possible object regions in Pronet (Sun et al. 2016). Moreover, a multifold MIL approach is presented in (Gokberk Cinbis, Verbeek, and Schmid 2014) to avoid early determination of the error location. (Zhang et al. 2016) proposes a probabilistic WTA model along with excitation backpropagation to generate top-down attention maps based on a top-down signal. Additionally, (Zhou et al. 2016) analyzes the response value of the neuron instead of the gradient to indicate the class-specific object location. Further efforts (Selvaraju et al. 2017) use the category-aware neuron gradient, backing along convolutional layers to generate a localization map. The above methods represent many inspiring ideas in weakly supervised learning.

Active Top-Down Searching

(Kong et al. 2017a) design several different types of reversion connections to select and integrate the features from different layers. (Fu, Zheng, and Mei 2017) imitate the human principle by proposing a framework called the recurrent attention convolutional neural network that iteratively gains attention proposals that focus on the most distinguished area. Meanwhile, reinforcement learning has shown competitive performance due to its superior decision process (Mnih et al. 2013; Silver et al. 2016). In (Cao et al. 2017), long short-term memory is combined with a policy network to consider the attended region recurrently to solve the face hallucination problem.

Our work has similarities with (Wei et al. 2017; Kumar Singh and Jae Lee 2017), but we do not require the classification network to be retrained after each image is erased. Additionally, instead of setting the size of the erasing area manually, our proposed agents have the ability to learn the erasing degree themselves.

Proposed Approach

Overview

The process of weakly supervised localization is seen as controlling a sequence of actions to change the fixation area and identify the target objects. Hence, we cast the challenge of object localization as a Markov decision process (MDP) because this setting allows the proposed agents to make a decision in continuous steps. As in Fig. 2, a CNN-based feature extractor is trained by image-level annotation and is used to extract the feature maps of input images. Subsequently, actions are made by deep Q-network-based agents to maximize the long-term global reward. The environment is composed of the input image and the actions taken thus far, and an action is defined to guide the next erasing regions. The agent receives the state of the current image status and series history actions to make a decision, which then generates a positive or negative reward and guides the agent to search for as large of a reward as possible during the training phase. Strictly, this MDP contains a set of components. **S** describes the understanding of the current environment. **A** represents a series of operations that help us to achieve the goal. **R** gives agents a reward for optimizing the decision strategy.

CNN-Based Feature Extractor

The CNN-based feature extractor is initialized based on the Resnet-50 model. Global average pooling and class-wise pooling (Durand et al. 2017) are applied on top of the model. We denote the training set with the image-level labels as $\mathcal{D}$=$\{(x_i, y_i)\}_{i=1}^{N}$, where x_i represents the image data, and y_i is a C-dimensional label vector with C categories. The correct loss for multi-label classification is the multi-label cross-entropy loss function:

$$Loss = -\frac{1}{N}\sum_{i=1}^{N}[y_i log(\frac{1}{1+e^{-x_i}}) + (1-y_i) log(\frac{e^{-x_i}}{1+e^{-x_i}})]. \quad (1)$$

Here, $log(\cdot)$ is the logarithmic function.

Actions

The set of actions **A** is different from the traditional actions in the RL detection framework in (Caicedo and Lazebnik

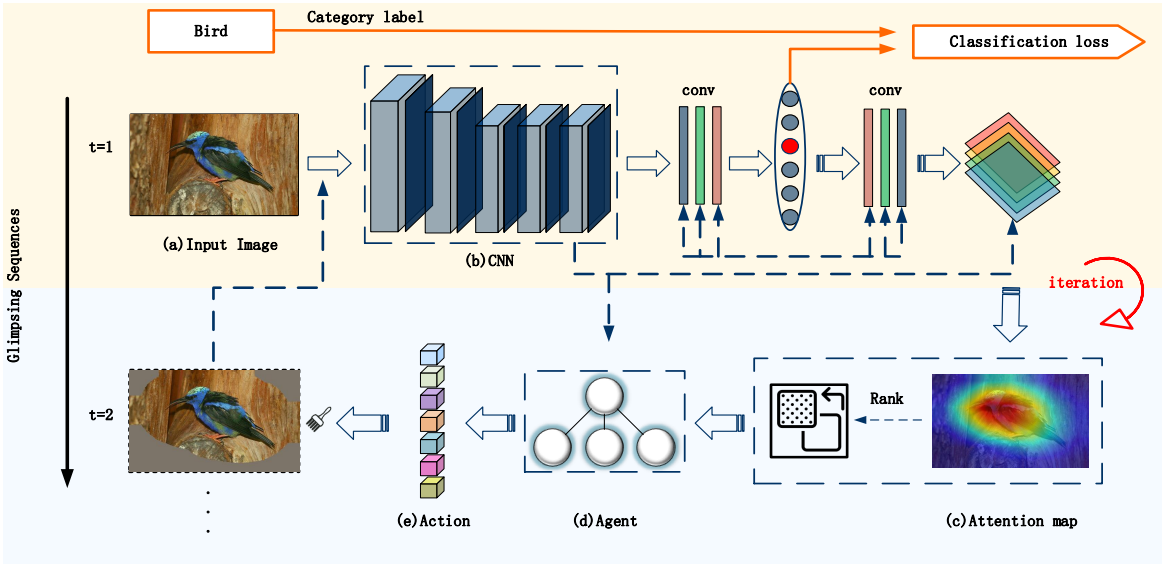

Figure 2: Pipeline of the proposed active searching strategy. We first train the feature extractor (b) using the category label. Then, the CAM method is employed to obtain the class-specific attention map (c). Ranking the attention map (c) reveals both of the most discriminate and undesired regions. The feature vector of the input image (a) is seen as the state of the proposed agents (d). The proposed agents (d) erase the input image and produce the next step image, which is then fed into the feature extractor.

2015; Lan et al. 2017; Kong et al. 2017b), which are composed of several transformations on the bounding box. Furthermore, our actions are close to operating on a pixel-level rather than a bounding box. First, the class activation maps (CAMs) (Zhou et al. 2016) are obtained from the current step image x_t^i with category c. In this way, we define the activation maps $M_c(u, v)$ in the localization (u, v) as category c, which is given by formulation:

$$M_c(u, v) = \sum_{k=1}^{K} w_k^c f_k(u, v), \tag{2}$$

where k is the k^{th} neuron from the last convolutional layer, and w_k^c represents the weights amounting to category c for neuron k. Given a spatial location (u, v), $f_k(u, v)$ indicates the activation of neuron k in the last convolutional layer from the feature extractor. $M_c(u, v)$ represents the importance guiding the classification of image to category c.

Second, the activation values from $M_c(u, v)$ are sorted in descending order to obtain $rank_{M_c(u,v)}$. We define the action set **A** to enable the agent to decide the location and size of the erasing region according to $rank_{M_c(u,v)}$. Formally, the action set is composed of six actions, **A**={5%,10%,15%,20%,25%,terminate}, which represent different degrees of increments for erasing. Therefore, the terminate action is performed when the entire region is erased, which is not expected to occur in our experiments. Compared with moving a fixed degree in each step, our proposed dynamic strategy improves the efficiency in detecting different sizes of objects through a process of "brushing" using different sized brushes.

States

We denote a two-tuple as **S**=(e, h), the feature vectors of the current input image $e_t^i \in \mathbb{R}^{512*7*7}$ and the history vector representing taken actions in last few steps $h \in \mathbb{R}^{24}$. The feature extractor $f : \mathbb{R}^W \mapsto \mathbb{R}^{512*7*7}$ forward propagates the image $x_t^i \in \mathbb{R}^W$ of W pixels at step t and extracts the image information e_t^i of the i^{th} images of the t^{th} step. We propose to erase the input image x_t^i by replacing the pixel values in a given mask of the image $Mask_t^i \in \mathbb{R}^W$ with the 3-dimensional vector g. Vector g is learned from a training set as the mean value of the channels of each image. The function that erases image x_t^i given mask $Mask_t^i$ using vector g as $h_g is : \mathbb{R}^W \mapsto \mathbb{R}^W$. Note that the output of the function is again the image, and the masks $Mask_t^i$ are generated by our proposed actions. We define the state transition of image x_t^i subject to mask $Mask_t^i$ as the value of the function $\delta_f(x_t^i, Mask_t^i) : \mathbb{R}^W \times \mathbb{R}^W \in \mathbb{R}^W$ given by

$$\delta_f(x_t^i, Mask_t^i) = f(h_g(x_t^i, Mask_t^i)). \tag{3}$$

Rewards

The reward function $\mathbf{R}(s, a)$ is the critical factor in guiding which action to encourage in the decision-making process. We define image x_t^i at step t and obtain image x_{t+1}^i after taking action a. The current classification confidence is represented as $cls_t^{x^i}$, and classification confidence of the next step is defined as $cls_{t+1}^{x^i}$. Denote states s and s' in step t and step $t + 1$. The reward function of the classification is then

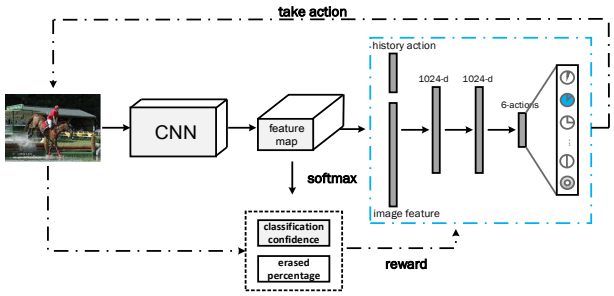

Figure 3: Architecture of the proposed Q-network.

defined as follows:

$$R_a^{cl}(s,s') = \begin{cases} +\sigma,\ if\ cls_{t+1}^{x_t^i} - cls_t^{x_t^i} < \tau\ and\ cls_{t+1}^{x_t^i} > \xi \\ -\sigma,\ otherwise \end{cases}. \quad (4)$$

Eq. 4 shows that the proposed agent will receive the reward if the change in classification confidence is less than τ, and the agent will be penalized otherwise. σ is the classification reward, which we set to 3.2 in our experiments. We constrain the classification confidence to not less than $\xi = 0.4$. Considering some of the situations that may occur in Fig. 9, we set the difference in the classification confidence of the current step and next step τ as -0.1 in Eq. 4. This setting avoids the insufficiency of the CNN model when the objects of focus have irregular borders. The agent is penalized when the discriminate region of the object is erased and is rewarded when the target object is retained in the current state.

We design a reward function to restrict the degree of the erasing region to a feasible range. For ease of explanation, we define $erase_t^{x^i}$ as step t of the accumulative erasing degree. The reward function of the degree of the erased region can be written as follows:

$$R_a^{erase}(s,s') = \begin{cases} +\beta,\ if\ \mu < erase_t^{x^i} < \psi \\ -\beta,\ otherwise \end{cases}, \quad (5)$$

where we set μ to 0.5 and ψ to 0.8. The erasing degree reward β is set to 0.5 in our experiments.

The terminating action acts as a penalty in our setting. Therefore, the termination reward function is presented as follows:

$$R_a^{termi}(s,s') = \zeta, \quad (6)$$

where ζ is -0.5 in our experiments. In other words, the reward function aims to force the proposed agent to not stop until the final step. The total number of steps T is set to 5.

Deep Q-Learning

A policy $\pi(s)$ is specified to learn the optimal action for the current state in each time step. We apply a deep Q-learning framework to build the relation between state and action. In the training phase, the agent randomly selects an action from the action container with probability ϵ and chooses the maximum activation value action with probability $1-\epsilon$. Starting from 0.9, ϵ decreases by 0.1 in each epoch until $\epsilon = 0.1$. To overcome the unstable and useless learning in traditional Q learning, a memory replay scheme is applied in the training stage. Moreover, the Bellman equation is adopted to iteratively update the selection of the action policy, where s and a are the current state and action, respectively. r is the reward, and $max_{a'}Q(s',a')$ represents the future reward. The action a chosen at state s determines the corresponding reward of the agent via the function $Q(s,a)$, where the highest reward is acquired through the selected action:

$$Q(s,a) = r + \gamma max_{a'}Q(s',a'), \quad (7)$$

where γ represents a discount factor to balance future and current rewards. We set γ to 0.9 in our experiments. The weights of the deep Q-network θ^t at time step t with transition quadruple (s,a,r,s') update as follows:

$$\theta_{t+1} = \theta_t + \alpha(r + Q(s',a';\theta_t) - Q(s,a;\theta_t))\nabla_{\theta_t}Q(s,a;\theta_t). \quad (8)$$

Algorithm 1 Training process of the proposed method

Require: Training data $\mathcal{D}=\{(x_i,y_i)\}_{i=1}^N$, replay_counter, class_num, max_epoch, T;
Train the feature extractor f with $\mathcal{D}$;
for $c = 1, class\_num$ **do**
  Initialize agent DQN_c with param θ^c;
  for epoch=1,max_epoch **do**
    for x_i,y_i in $\mathcal{D}$ **and** c in y_i **do**
      Set status=1;
      Get state $s_t^{x_i}$, current step confidence $cls_t^{x_i}$;
      while t $< T$ and status==1 **do**
        Calculate $M_{i,t}^c(u,v)$ by CAM(x_t^i, CLS, y_i);
        Select action $a_t^{x_i}$ with ϵ-greedy;
        if a==6 **do**
          status=0;
        Obtain $Mask_{x_i,t}^c$ and $erase_t^{x_i}$;
        Do erase $x_{i,t+1}^c = x_{i,t}^c \otimes Mask_{x_i,t}^c$;
        Get $s_{t+1}^{x_i}$ and $cls_{t+1}^{x_i}$;
        Calculate reward $r_t^{x_i}$ by Eq. 4,5;
        Store $transition(s_t^{x_i}, a_t^{x_i}, r_t^{x_i}, s_{t+1}^{x_i})$;
        Update confidence $cls_t^{x_i} = cls_{t+1}^{x_i}$;
        Do memory replay;
end for
Ensure: Trained DQN $DQN = \{DQN_c\}_{c=1}^{class\_num}$

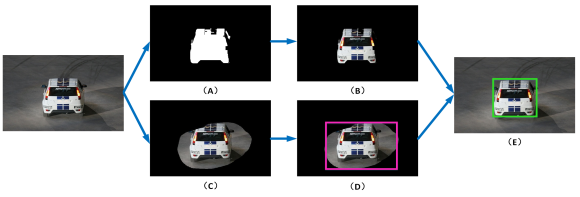

Figure 4: Examples of the post-processing refined bounding box. The first row of each sample represents the binary mask (A) and its segmentation cues (B) generated by saliency technology. The second row of each sample shows the erased image (C) and the selected bounding box proposed by the agents (D). (E) shows the refined results.

Table 1: Quantitative comparison of the detection task on the PASCAL VOC 2007 *test* set. The average precision of our method is computed by detecting one object of the same category per image. The average detection precision is compared for common detection tasks in the same dataset. Hence, we consider a horizontal comparison in this section.

Method	plane	bike	bird	boat	btl	bus	car	cat	chair	cow	table	dog	horse	moto	pers	plant	sheep	sofa	train	tv	mAP
Cinbis *et al.*	35.8	40.6	8.1	7.6	3.1	35.9	41.8	16.8	1.4	23.0	4.9	14.1	31.9	41.9	**19.3**	11.1	27.6	12.1	31.0	40.6	22.4
Wang *et al.*	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
LocNet	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
OICR-VGG	**58.0**	**62.4**	31.1	**19.4**	**13.0**	**65.1**	**62.2**	**28.4**	24.8	**44.7**	30.6	25.3	37.8	**65.5**	15.7	**24.1**	**41.7**	46.9	**64.3**	**62.6**	**41.2**
Ours-H	52.5	34.0	32.0	16.9	5.1	55.3	46.1	54.5	12.6	29.0	49.0	41.0	49.6	54.1	28.4	15.4	26.5	42.8	44.1	10.0	34.9
Ours-M	57.5	35.6	39.9	25.6	6.1	58.0	50.0	56.7	13.6	29.0	49.4	48.0	62.5	58.2	29.3	15.4	27.1	45.4	51.4	10.0	38.4
Ours-M-D	**57.5**	42.0	**39.9**	31.2	10.0	**58.8**	**50.0**	**57.2**	13.6	35.6	**49.4**	49.0	**62.5**	**58.2**	30.4	15.5	33.8	45.4	**57.0**	28.4	**41.2**

Table 2: Quantitative comparison of localization on the PASCAL VOC 2012 *val* set. Note that RCNN and Fast-RCNN are supervised with a bounding box label, while the other methods have only category labels.

Method	plane	bike	bird	boat	btl	bus	car	cat	chair	cow	table	dog	horse	moto	pers	plant	sheep	sofa	train	tv	mAP
RCNN	**92.0**	80.0	80.0	73.0	49.9	86.8	**77.7**	87.6	50.4	72.1	57.6	82.9	79.1	**89.8**	**88.1**	56.1	83.5	50.1	82.0	76.6	74.8
Fast-RCNN	79.2	74.7	74.7	65.8	39.4	82.3	64.8	85.7	54.5	77.2	58.8	85.1	86.1	80.5	76.6	46.7	79.5	68.3	85.0	60.0	71.3
Oquab *et al.*	90.3	77.4	77.4	79.2	41.1	87.8	66.4	91.0	47.3	**83.7**	55.1	88.8	**93.6**	85.2	87.4	43.5	**86.2**	50.8	86.8	66.5	74.5
ProNet-P	91.6	82.0	**85.1**	78.6	45.9	**87.9**	67.1	92.2	51.0	72.9	60.8	89.3	85.1	85.3	86.4	45.6	83.5	55.1	85.6	65.9	74.8
Ben *et al.*	90.0	81.2	81.2	**82.2**	47.5	86.7	64.9	85.7	53.9	75.8	67.9	82.2	84.1	83.4	83.9	**71.7**	83.1	63.7	**89.4**	78.2	77.1
Ours	91.5	**87.0**	77.6	65.7	**54.3**	85.1	68.1	**94.0**	**55.9**	78.3	**87.0**	**90.0**	86.1	89.3	68.0	59.7	68.2	**81.2**	83.0	**79.3**	**77.5**

The training process of the proposed method is described in Algorithm 1.

Experiments

In the following sections, the details of the experiments with the proposed method are discussed.

Experimental Settings

We evaluate the proposed method on the PASCAL VOC 2007 and 2012 datasets. Average precision (AP) and correct localization CorLoc are used to evaluate the performance of our method. For both metrics, the true positive bounding box is determined as correct only if it has an at least 50% intersection-over-union (IoU) ratio with the ground-truth object instance annotation. To compare the state-of-the-art weakly supervised localization methods, we also apply the localization metrics proposed in (Oquab et al. 2015) and (Zhu et al. 2017).

Implementation details. First, for the feature extractor, we use Resnet-50, pretrained on the ImageNet dataset and fine-tuned on the corresponding training sets, for the backbone architecture. We train the network with stochastic gradient descent (SGD) with momentum at a learning rate of 0.01 for 20 epochs. Second, the image descriptor and the history vector are the inputs of the deep Q-network, whose structure is composed of two fully connected layers with 1024 neurons and an action layer with 6 neurons. The architecture of the deep Q-network is shown in Fig. 3. Furthermore, we train each deep Q-network model for 50 epochs, apply an experience replay of 1000 memory capacity, and set the target parameter update step to 100. In post-processing, we adopt saliency detection technology (Jiang et al. 2013) to refine the detected results, as shown in Fig. 4. "Ours-H" indicates the hard threshold post-processing, and "Ours-M" represents the mean value threshold. "Ours-M-D" uses the dilation CNN to obtain the heat maps (Wei et al. 2018), which demonstrates the flexibility of our proposed method.

Detection Results

In Table 1, compare the object detection results on the PASCAL VOC 2007 *test* set with those of (Gokberk Cinbis, Verbeek, and Schmid 2014; Wang et al. 2014; Kantorov et al. 2016; Tang et al. 2017). The proposed method achieves a higher average detection precision in the conveyance categories $i.e.$, plane, bus, motorbike and train, reaching 57.5%, 58.0%, 58.2% and 51.4%, respectively. First, the proposed reward function forces the agent to consider both the classification confidence and the area when focusing on only a specific part of an object. Second, the irregular shape, such as "cat" and "horse", is the difference with respect to the conveyance category. CNN-based weakly supervised methods often propose rectangle bounding boxes, and the proposed method retains more of the original shapes of objects.

In addition, the proposed method achieves a CorLoc of 32.7%, outperforming the compared transfer learning (32.1%) and mining (30.2%) methods in Table 3. We compare two pre-deep learning methods in Table 3, mainly because we do not extract and classify proposals in the training stage so poorer performance occurs in the *trainval* set.

Localization Results

As shown in Table 2, the proposed method achieves comparable performance to that of the CNN methods (Girshick et al. 2014; Girshick 2015; Oquab et al. 2015; Sun et al. 2016; Bency et al. 2016), achieving 77.5%. The proposed method produces the best results on the categories "bike", "bottle", "chair", "table", "sofa" and "tv". Specifically, for "table", the proposed method outperforms (Bency et al. 2016) by

Table 3: Quantitative comparison of Corloc on the PASCAL VOC 2007 *trainval* set. The method Ours-top3* is the percentage of positive images in which an object is correctly located by at least one of the top-3 proposals.

Method	plane	bike	bird	boat	btl	bus	car	cat	chair	cow	table	dog	horse	moto	pers	plant	sheep	sofa	train	tv	mAP
Siva *et al.*	45.8	21.8	30.9	20.4	5.3	37.6	40.8	51.6	7.0	29.8	27.5	41.3	41.8	47.3	24.1	12.2	28.1	32.8	48.7	9.4	30.2
Shi *et al.*	**54.7**	22.7	33.7	**24.5**	4.6	33.9	42.5	**57.0**	7.3	**39.1**	24.1	**43.3**	41.3	**51.5**	25.3	13.3	28.0	29.5	54.6	**11.8**	32.1
Ours-H	41.1	24.2	39.0	15.8	9.4	**42.0**	42.0	52.0	10.1	29.7	37.5	41.4	42.8	49.3	28.1	14.7	31.2	45.2	49.5	9.8	32.7
Ours-M	42.3	**25.5**	**39.2**	19.1	**9.4**	**42.0**	**43.7**	53.3	**10.3**	30.4	**37.5**	41.7	**49.0**	50.6	**28.3**	**16.3**	**31.2**	**45.2**	**55.6**	9.8	**34.7**
Ours-top3*	53.3	30.5	48.5	21.0	9.8	49.5	49.5	63.2	10.8	34.8	46.5	57.5	59.2	60.4	35.9	19.6	35.4	54.1	63.2	10.2	40.6

Table 4: Pointing localization accuracy (%) on the PASCAL VOC 2007 *test* set. **Center** is a baseline method that uses the image center as the estimate of the object center.

Method	localization accuracy(%)
Center	69.5
Deconv(Zeiler and Fergus 2014)	73.1
LRP(Bach et al. 2015)	68.1
MWP(Zhang et al. 2016)	73.7
Ours-H	76.0
Ours-M	**76.7**

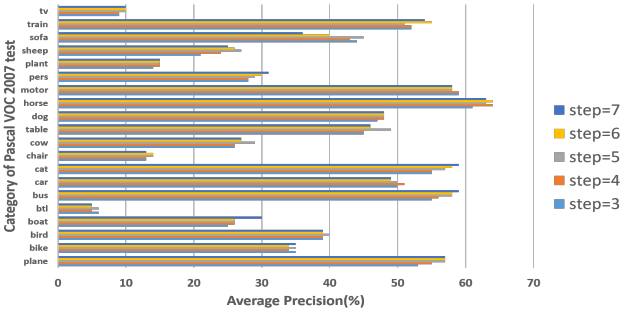

Figure 5: Detection performance with respect to the total number of steps T on the PASCAL VOC2007 test set. The histogram represents the categories (x-axis) and the corresponding average precision (y-axis) for different total number of steps T.

19.1% (87% vs 67.9%), and for "sofa", the proposed method outperforms Fast-RCNN by 12.9% (81.2% vs 68.3%), illustrating the substantial improvement on furniture categories. Furniture class objects are usually located in complicated scene, often overlapping each other, as shown in the second row of Fig. 7. Therefore, furniture class objects are often difficult to identify separately. Our proposed method is able to separate the main objects to produce good results because the overlapping regions are erased correctly. Moreover, the proposed method has comparable localization accuracy results, as shown in Table 4. The proposed method achieves a localization accuracy of 34.7%, outperforming the compared transfer learning (32.1%) and mining (30.2%) methods. When considering the top-3 condition, the localization accuracy of the proposed method increases 8%, reaching 40.6%.

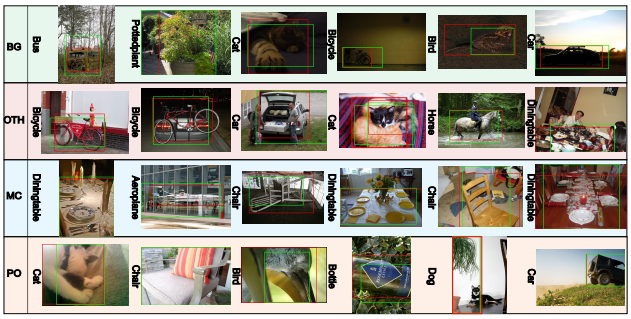

Figure 6: Sample detection results. Red boxes represent the ground-truth annotation. Green boxes indicate correct detection results. "BG" represents samples that are easily confused with the background. "OTH" indicates samples that are easily confused with other categories. "MC" shows samples that are located in messy surroundings. "PO" denotes samples that are shown partially.

Further Analysis

Impact of the total number of steps T. To validate the choice of the total number of steps T, we evaluate various values, as displayed in Fig. 5. This comparison demonstrates that the detection precision may be influenced by changing T from 3 to 7. For most of the compared categories, except "horse", "moto" and "sofa", the greater the number of steps is, the higher the average precision. These three categories always appear with overlap, indicating that the erasing action may damage the structural information.

Impact of hyperparameters. We investigate the effect of the parameters ξ, σ, β and ζ on the detection performance. The results in Fig. 8 show that a lower classification confidence, i.e., $\xi = 0.4$, is beneficial for detection. We further note that a higher classification reward, i.e., $\sigma = 3.2$, yields the best results. Moreover, the erasing degree β also affects the performance and is set to $\beta = 0.5$ to achieves the best results. Another important parameter is the termination reward ζ, which penalizes the agents when the terminate action is triggered. $\zeta = -0.5$ leads to the highest precision.

Analysis of visualization. Fig. 6 shows the sample detection results. Our proposed method can detect the object under conditions where it is difficult to discriminate between the object and the background. Our method can also handle cases of multiple categories or overlap among the same category. Moreover, problems with messy surroundings and the appearance of only parts of objects are solved, as shown in

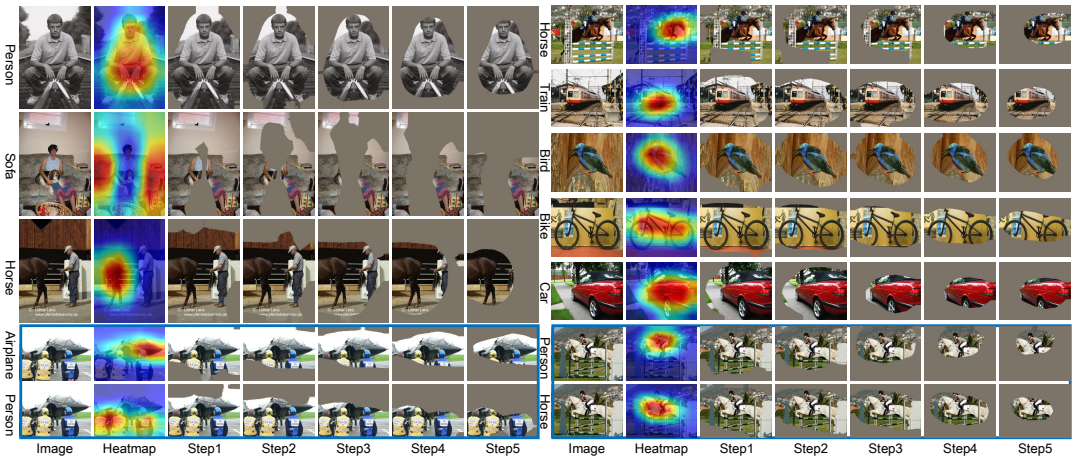

Figure 7: Examples of the erased object regions produced by the proposed method. The third to seventh and the tenth to fourteenth columns show the erasing procedure. The second and ninth columns present the produced heat maps.

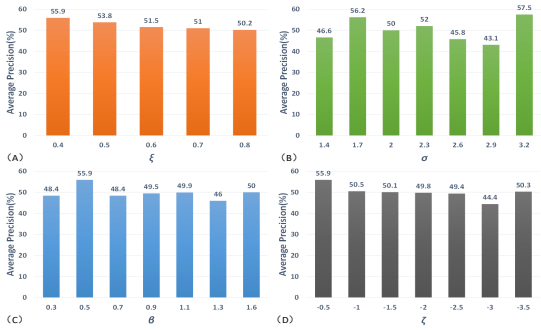

Figure 8: Detection results of category "aeroplane" with different parameters on the PASCAL VOC 2007 test set. (A) Impact of ξ. (B) Impact of σ. (C) Impact of β. (D) Impact of ζ. Each histogram represents the value of the parameter (x-axis) and the corresponding average precision (y-axis).

the third and fourth rows. The heat maps and the corresponding erasing action are visualized to help us to understand the proposed method in Fig. 7. In the last erasing step, key components of people are retained, which demonstrates the meaningful procedure of the proposed method. The ideal results shown in the second row and the appearance of overlapping areas indicates that our proposed method has the ability to distinguish class-specific regions of interest in regions of overlap. Certain relations exist among different objects in the training sets. Therefore, searching for a class-specific object in an image is unreasonable when the areas of correlative objects have responses. However, our proposed method can effectively avoid this situation by fully utilizing top-down information. Clearly, the hard-to-interpret appearance, which often occurs in CNN models, has been overcome to some extent.

Analysis of classification confidence. As is commonly believed, the more focused an object is, the better the classification accuracy of the neural network. However, the opposite phenomenon may occasionally occur, as shown in Fig. 9. CNN always acts as a black box to some extent, but the proposed method provides an explanation of the reason why each action is taken. From another perspective, it is reasonable that the background is related to the object during training on large amounts of data. Nonetheless, the background information would affect the performance in some extent. Hence, a sufficient amount of data is the only way to cope with the adequate condition for CNN models. This problem can be solved by ignoring the background and focusing on only the area where the object has a higher possibility of appearing, as is done with our method.

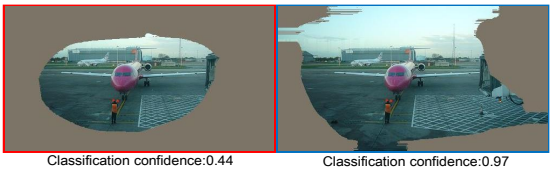

Figure 9: Red frame images, which indicate more focused objects, have lower classification confidence, and the blue frame images have higher confidence under conditions where a limited region is erased.

Conclusions

We present a novel human-like delicate region erasing strategy to solve the weakly supervised object localization problem. The difference between the proposed method and previous works is that a top-down scene analysis is performed by agents to erase pixel-level regions of the background via a human visual mechanism. A deep Q-network, acting as an agent, is applied to learn the localization policy and to optimize the policy to iteratively determine the location of an object. The experimental results demonstrate that the weakly

supervised localization performance of the proposed model is comparable to that of CNN-based methods and that the efficiency is improved simultaneously. We conclude that the proposed method with a human-like mechanism is applicable to weakly supervised localization.

Acknowledgments

This work was supported in part by the National Key R&D Program of China (No. 2018YFB1004600, No. 2017YFC0803705), the National Natural Science Foundation of China (No. 61761146004, No. 61836084 No. 61773375, No. 61572004, No. 61771026), the Beijing Municipal Natural Science Foundation (No. Z181100008918010), the Key Project of Beijing Municipal Education Commission (Research and application on a co-evolutionary model of visual perception and coginition); and the innovation Platform Construction of QingHai Province (2016-ZJ-Y04).

References

Bach, S.; Binder, A.; Montavon, G.; Klauschen, F.; Müller, K.-R.; and Samek, W. 2015. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PloS one* 10(7):e0130140.

Bency, A. J.; Kwon, H.; Lee, H.; Karthikeyan, S.; and Manjunath, B. 2016. Weakly supervised localization using deep feature maps. In *ECCV*.

Caicedo, J. C., and Lazebnik, S. 2015. Active object localization with deep reinforcement learning. In *ICCV*.

Cao, Q.; Lin, L.; Shi, Y.; Liang, X.; and Li, G. 2017. Attention-aware face hallucination via deep reinforcement learning. In *CVPR*, 1656–1664. IEEE.

Durand, T.; Mordan, T.; Thome, N.; and Cord, M. 2017. Wildcat: Weakly supervised learning of deep convnets for image classification, pointwise localization and segmentation. In *CVPR*.

Fu, J.; Zheng, H.; and Mei, T. 2017. Look closer to see better: recurrent attention convolutional neural network for fine-grained image recognition. In *CVPR*.

Girshick, R.; Donahue, J.; Darrell, T.; and Malik, J. 2014. Rich feature hierarchies for accurate object detection and semantic segmentation. In *CVPR*.

Girshick, R. 2015. Fast r-cnn. In *ICCV*.

Gokberk Cinbis, R.; Verbeek, J.; and Schmid, C. 2014. Multi-fold mil training for weakly supervised object localization. In *CVPR*.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *CVPR*.

Jiang, H.; Wang, J.; Yuan, Z.; Wu, Y.; Zheng, N.; and Li, S. 2013. Salient object detection: A discriminative regional feature integration approach. In *CVPR*.

Kantorov, V.; Oquab, M.; Cho, M.; and Laptev, I. 2016. Contextlocnet: Context-aware deep network models for weakly supervised localization. In *ECCV*.

Kong, T.; Sun, F.; Yao, A.; Liu, H.; Lu, M.; and Chen, Y. 2017a. Ron: Reverse connection with objectness prior networks for object detection. In *CVPR*.

Kong, X.; Xin, B.; Wang, Y.; and Hua, G. 2017b. Collaborative deep reinforcement learning for joint object search. In *CVPR*.

Kumar Singh, K., and Jae Lee, Y. 2017. Hide-and-seek: Forcing a network to be meticulous for weakly-supervised object and action localization. In *CVPR*.

Lan, X.; Wang, H.; Gong, S.; and Zhu, X. 2017. Identity alignment by noisy pixel removal. *arXiv preprint arXiv:1707.02785*.

Lu, Y.; Javidi, T.; and Lazebnik, S. 2016. Adaptive object detection using adjacency and zoom prediction. In *CVPR*.

Mnih, V.; Kavukcuoglu, K.; Silver, D.; Graves, A.; Antonoglou, I.; Wierstra, D.; and Riedmiller, M. 2013. Playing atari with deep reinforcement learning. *arXiv preprint arXiv:1312.5602*.

Oquab, M.; Bottou, L.; Laptev, I.; and Sivic, J. 2015. Is object localization for free?-weakly-supervised learning with convolutional neural networks. In *CVPR*.

Selvaraju, R. R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; and Batra, D. 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *ICCV*.

Silver, D.; Huang, A.; Maddison, C. J.; Guez, A.; Sifre, L.; Van Den Driessche, G.; Schrittwieser, J.; Antonoglou, I.; Panneershelvam, V.; Lanctot, M.; et al. 2016. Mastering the game of go with deep neural networks and tree search. *Nature* 529(7587):484–489.

Song, H. O.; Lee, Y. J.; Jegelka, S.; and Darrell, T. 2014. Weakly-supervised discovery of visual pattern configurations. In *NIPS*.

Sun, C.; Paluri, M.; Collobert, R.; Nevatia, R.; and Bourdev, L. 2016. Pronet: Learning to propose object-specific boxes for cascaded neural networks. In *CVPR*.

Tang, P.; Wang, X.; Bai, X.; and Liu, W. 2017. Multiple instance detection network with online instance classifier refinement. In *CVPR*.

Wang, C.; Ren, W.; Huang, K.; and Tan, T. 2014. Weakly supervised object localization with latent category learning. In *ECCV*.

Wei, Y.; Feng, J.; Liang, X.; Cheng, M.-M.; Zhao, Y.; and Yan, S. 2017. Object region mining with adversarial erasing: A simple classification to semantic segmentation approach. In *CVPR*.

Wei, Y.; Xiao, H.; Shi, H.; Jie, Z.; Feng, J.; and Huang, T. S. 2018. Revisiting dilated convolution: A simple approach for weakly-and semi-supervised semantic segmentation. In *CVPR*.

Zeiler, M. D., and Fergus, R. 2014. Visualizing and understanding convolutional networks. In *ECCV*.

Zhang, J.; Lin, Z.; Brandt, J.; Shen, X.; and Sclaroff, S. 2016. Top-down neural attention by excitation backprop. In *ECCV*.

Zhou, B.; Khosla, A.; Lapedriza, A.; Oliva, A.; and Torralba, A. 2016. Learning deep features for discriminative localization. In *CVPR*.

Zhu, Y.; Zhou, Y.; Ye, Q.; Qiu, Q.; and Jiao, J. 2017. Soft proposal networks for weakly supervised object localization. In *ICCV*.