The Thirty-Second AAAI Conference
on Artificial Intelligence (AAAI-18)

Lateral Inhibition-Inspired Convolutional Neural Network for Visual Attention and Saliency Detection

Chunshui Cao,[1,2] **Yongzhen Huang,**[2,3,4] **Zilei Wang,**[1] **Liang Wang**[2,3,4]
Ninglong Xu,[3,4,5] **Tieniu Tan**[2,3,4]
[1]University of Science and Technology of China
[2]Center for Research on Intelligent Perception and Computing (CRIPAC),
Institute of Automation, Chinese Academy of Sciences (CASIA)
[3]Center for Excellence in Brain Science and Intelligence Technology (CEBSIT)
[4]University of Chinese Academy of Sciences (UCAS)
[5]Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences

Abstract

Lateral inhibition in top-down feedback is widely existing in visual neurobiology, but such an important mechanism has not be well explored yet in computer vision. In our recent research, we find that modeling lateral inhibition in convolutional neural network (LICNN) is very useful for visual attention and saliency detection. In this paper, we propose to formulate lateral inhibition inspired by the related studies from neurobiology, and embed it into the top-down gradient computation of a general CNN for classification, i.e. only category-level information is used. After this operation (only conducted once), the network has the ability to generate accurate category-specific attention maps. Further, we apply LICNN for weakly-supervised salient object detection. Extensive experimental studies on a set of databases, e.g., ECSSD, HKU-IS, PASCAL-S and DUT-OMRON, demonstrate the great advantage of LICNN which achieves the state-of-the-art performance. It is especially impressive that LICNN with only category-level supervised information even outperforms some recent methods with segmentation-level supervised learning.

Introduction

Visual attention is an important mechanism of visual information processing in human brain. It has been extensively studied in neurobiology and preliminarily borrowed in computer vision, including the top-down models for generating category-specific attention maps (Zhang et al. 2016; Cao et al. 2015; Zhou et al. 2015; Simonyan, Vedaldi, and Zisserman 2013) and bottom-up models for salient object detection (Li and Yu 2016; Zhao et al. 2015; Wang et al. 2015b). In most of existing methods, however, the attention maps usually suffer from heavy noise or fail to well preserve a whole object. More importantly, the state-of-the-art models usually require strong supervision information for training, e.g., manually labeled segmentation masks of salient objects, which is pretty time- and labor-consuming. However, human performance of visual attention is robust. In neurobiology, classic theories like object binding (Wyatte, Herd, and Mingus 2012) and selective attention (Desimone and Duncan 1995), show that top-down feedback conveys attentional signals to the sensory area of visual cortex and supplies the criteria to select neurons that most relevant to a specific task (Desimone and Duncan 1995), and lateral inhibition (Nielsen 2001) can further modulate the feedback signals by creating competition among neurons, which leads to enhanced visual contrast and better perception of interested objects (Wyatte, Herd, and Mingus 2012). This attention mechanism is inspiring to model both stimulus-driven and goal-oriented visual attention, and it is noted that no strong supervised information, e.g., object segmentation mask, is needed.

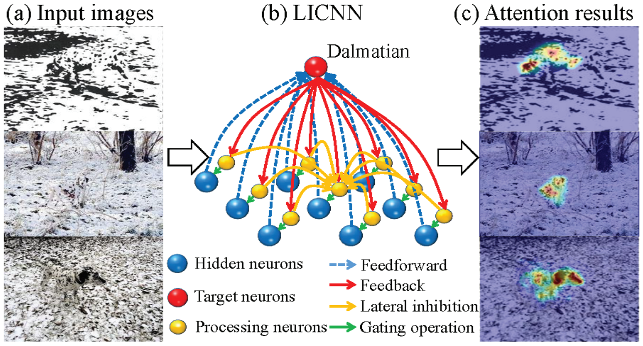
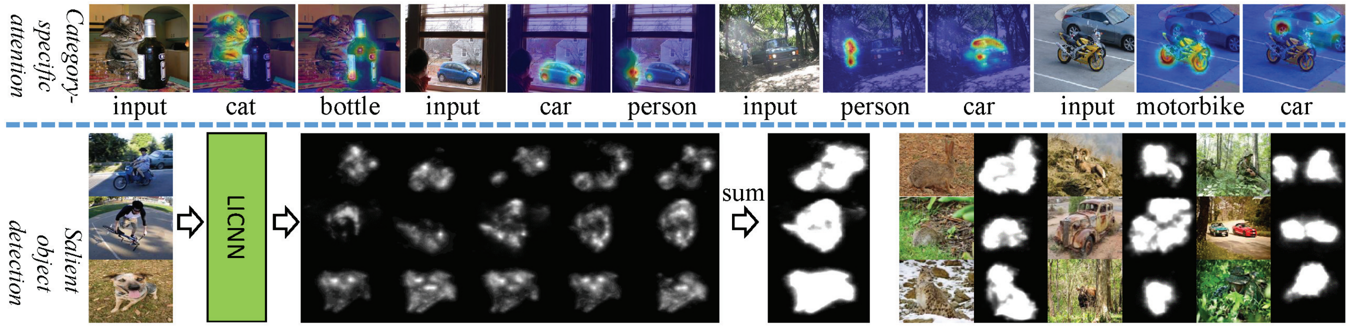


Figure 1: LICNN for visual attention. (a) Input images. Here, three very challenging "dalmatian-on-snow" images are particularly shown. (b) Modeling lateral inhibition in each hidden layer of CNN. (c) Category-specific maps generated by LICNN. Best viewed in color.

Motivated by the aforementioned findings, we propose a method to formulate lateral inhibition in convolutional neural network (LICNN). As we all know that, various patterns can be expressed in a CNN classifier. Given an input, they compete with each other during the feed-forward phase and eventually contribute to one or more classes, leading to the distinct scores over different categories. And the category-specific gradients can roughly estimate the importance of each pattern (Simonyan, Vedaldi, and Zisserman 2013). Thus, we employ the category-specific gradients as the feedback signals for attention in LICNN. And to bind



(a) Inputs (b) LICNN (c) Top-5 category-specific attention maps (d) Salient objects (e) More salient object detection results

Figure 2: Category-specific attention (top row) and salient object detection (bottom row) based on LICNN.

target-relevant patterns together and capture interested objets, a new lateral inhibition model is developed and embedded into the top-down gradient feedback process. The main principal is illustrated in Fig. 1. A pre-trained CNN is employed to process the input image with normal feedforward. Then the lateral inhibition is introduced among the hidden neurons in the same layer during category-specific feedback (back-propagation for the class of dalmatian). As shown in Fig. 1(b), *the status of a neuron is determined by the* ***top-down signals of its own and neighbors via a new lateral inhibition model.*** Afterwards, we can obtain a gradient-based map for each input image, as depicted in Fig. 1(c). We show more selective attention results in Fig. 2 (upper part), in which different kinds of objects are highlighted in color. As can be seen, the energy of attention is mainly around interesting objects and the shapes of objects are well preserved. With LICNN, discriminative category-specific attention maps can be derived with much less noise.

Moreover, LICNN can also detect salient objects from complicated images with only weak supervision cues. Since semantic patterns from the salient objects in a given image contribute to the decision making of classification, a saliency map can be produced with LICNN by choosing categories with the highest scores in the top level as the category-specific feedback signals. As shown in Fig. 2 (lower part), the five class nodes with the five highest outputs of a CNN classifier are regarded as a **bottom-up salient pattern detector**. The lateral inhibition is applied over hidden neurons in category-specific feedback for these five class respectively. As a result, we can obtain five attention maps, as demonstrated in Fig. 2(c). We produce the saliency map for each input image by integrating these five attention maps together, as depicted in Fig. 2(d). More results of challenging images are shown in Fig. 2(e).

It is interesting that LICNN can effectively locate salient objects even when the input image does not contain predefined objects in the CNN classifier. This is because a powerful CNN for classification has learned many visual local patterns shared by different objects, and lateral inhibition would make interesting objects more obvious. Although these objects do not belong to any training category, their parts are shared by other categories of objects.

We conduct extensive experiments on the PASCAL VOC dataset (Everingham et al. 2010) to evaluate category-specific attention. The results demonstrate that LICNN achieves much better results than the state-of-the-art approaches. We further apply LICNN for salient object detection on several benchmark datasets, including ECSSD, HKU-IS, PASCAL-S, and DUT-OMRON. It is noteworthy that LICNN with only category-level labeling can achieve comparable results to recent strongly supervised approaches with segmentation labeling.

In summary, the main contributions of this paper are two-fold: 1) To the best of our knowledge, this work for the first time implements lateral inhibition with an effective computational model to produce accurate attention maps, which is pretty useful for visual attention. 2) Based on LICNN, we develop an effective weakly supervised approach for salient object detection, largely enhancing the state-of-the-art and even achieving comparable performance to strongly supervised methods.

Related Work

Lateral Inhibition

The concept of lateral inhibition means a type of process in which active neurons suppress the activity of neighboring neurons through inhibitory connections. It has already been applied in several neural models. The analysis of the recurrent neural network with lateral inhibition was presented by Mao and Massaquoi (Mao and Massaquoi 2007). They showed that lateral suppression by neighboring neurons in the same layer makes the network more stable and efficient. Fernandes et al. (Fernandes, Cavalcanti, and Ren 2013) proposed a pyramidal neural network with lateral inhibition for image classification. Lateral inhibition has also been incorporated into some other vision tasks, such as structured sparse coding (Szlam, Gregor, and Cun 2011), color video segmentation (Fernández-Caballero et al. 2014), and image contour enhancement (Chen and Fu 2010). These approaches mainly adopt lateral inhibition mechanism during the bottom-up procedure. Different from these works, we incorporate the lateral inhibition mechanism into an algorithmic model with the top-down feedback signals of a deep CNN to combine both bottom-up and top-down information, and achieve powerful category-specific attention capability.

Top-down Attention

Top-down attention plays an important role in human visual system, and various top-down attention models have been proposed. Since deep CNNs greatly improve the performance of object recognition (Simonyan and Zisserman 2014; Szegedy et al. 2015), a number of CNN based category-specific attention models have been proposed (Zhang et al. 2016; Cao et al. 2015; Simonyan, Vedaldi, and Zisserman 2013; Zhou et al. 2015). In the work by Simonyan et al. (Simonyan, Vedaldi, and Zisserman 2013), target-relevant regions for a predicted category are visualized by error back-propagation. The work by Zhou et al. (Zhou et al. 2015) replaces the fully-connected layer with the average pooling layer to generate coarse class-activation maps that highlight target-relevant regions. The work in (Cao et al. 2015) embeds a top-down feedback model which introduces latent gate variables into CNN classifiers to capture task related regions in input images. Recently, an attention model based on the winner-take-all principle is proposed in (Zhang et al. 2016) to generate highly discriminative attention maps for object localization. Unlike these previous methods, our proposed LICNN is based on the lateral inhibition mechanism. Compared with the models in (Cao et al. 2015; Zhang et al. 2016), our model produces shape-preserved category-specific maps with less noise and exhibits more discriminative capability.

Salient Object Detection

CNN based approaches for salient object detection have been developed rapidly in recent years. The work by Wang et al. (Wang et al. 2015b) integrates both local estimation and global search using deep CNNs to detect salient objects. In (Zhao et al. 2015), both global and local contexts are adopted and integrated into a unified deep learning framework. The approach proposed in (Li and Yu 2016) derives a saliency map by integrating multi-scale features extracted from trained deep CNNs. Most of these CNN based methods work very well. However, these methods regard salient object detection as a problem independent from other vision tasks, which is thus different from the human visual perception. Meanwhile, these methods require strong supervised annotations to train models. In contrast, the approach proposed in this paper only requires weak supervision cues for training a CNN classifier. Despite of the use of weak supervision cues, our approach outperforms traditional methods and achieves competitive performance on popular benchmarks compared with the strongly supervised approaches.

Our Method

Top-down Feedback Signals

Recently, the evidence from (Zeiler and Fergus 2014; Cheng et al. 2017) indicates that various semantic patterns can be learned by convolutional neurons of CNN classifiers. *That is, specific neurons will be activated if there are positively correlated patterns lying in their receptive field. These activated patterns give rise to distinct scores of different classes.* Therefore, it is of great significance for selective attention to quantitatively estimate ***how much a pattern contributes to a category***. Top-down feedback is a brain-inspired way to achieve this. It provides an extra criteria to judge which patterns are relevant to a given category, and supplies an important foundation for modeling lateral inhibition. We implement this by utilizing category-specific gradients, which serve as the top-down feedback signals in LICNN. Given a CNN classifier, the output of a neuron is denoted as x. And S represents the output of a class node. A mapping function between x and S can be abstracted from the network, denoted as $f(x)$. Given an input image, x will take a certain value x_0. For this situation, S can be approximated by the first order Taylor expansion of $f(x)$ at x_0, i.e.,

$$\begin{aligned} S &= f(x) \\ &= f(x_0) + f'(x_0)(x - x_0) + o(x - x_0) \\ &\approx f'(x_0)x + f(x_0) - x_0 f'(x_0) \end{aligned} \tag{1}$$

It linearly approximates the relationship between S and x. Hence, after a CNN completes feed-forward, it is intuitive to quantitatively estimate the contribution of a neuron to a specific category by $f'(x_0)$, named as ***contribution weight (CW)***. Note that $f'(x_0) = \frac{\partial S}{\partial x_0}$ can be acquired by a simple back-propagation in CNNs. Since the activated patterns can be derived from the target objects, background or disturbed objects, it is rather rough to directly adopt the above approximation as ideal attention maps. *Fortunately, when this estimation is applied on all hidden neurons, a distribution of neurons' contributions emerges for all layers*. And the mechanism of lateral inhibition paves an effective way to refine and enhance the contrast of the estimated distribution.

Lateral Inhibition Model

In neurobiology, lateral inhibition in vision was inferred by Ernst Mach in 1865 as depicted in his Mach Band (Nielsen 2001), which reveals that lateral inhibition disables the spreading of action potentials from excited neurons to neighboring neurons in the lateral direction. This creates a contrast in stimulation that allows increased sensory perception. We mimic this mechanism with a new computational model and incorporate it into the top-down procedure of a CNN classifier. An interesting finding is that lateral inhibition combined with feedback signals(CWs) can *ensure that only the most relevant neurons are grouped together.*

Generally, the outputs of the convolutional layer are modulated by the subsequent ReLU layer to decide whether a pattern is activated. Thus, we introduce lateral inhibition to activated neurons in the ReLU layers. Note that the CWs of all ReLU neurons can be calculated via back-propagation. Assume that the layer l produces a cub of CWs with the dimension of (W, H, C), where W, H, C denote width, height and channels, respectively. Normally the C neurons at the same location represent different patterns though they share information in the same receptive field.

Firstly, a simple inhibition along the channel axis is performed by selecting the maximum CW at each location, thus we can obtain a CW map normalized by L2 norm with the dimension of (W, H), named as Max-C Map. Next, we construct lateral connections between different points in the Max-C Map to compute inhibition strength for each location. Formally, the lateral inhibition value is computed as

$$x_{ij}^{LI} = a * \underbrace{e^{-\overline{x_{ij}}}}_{average} + b * \underbrace{\sum_{uv}(d_{uv}e^{-d_{uv}}\delta(x_{uv} - x_{ij}))}_{differential} \tag{2}$$

$$y_{ij} = \begin{cases} y_{ij} & \text{if } x_{ij} - x_{ij}^{LI} > 0 \\ 0 & \text{if } x_{ij} - x_{ij}^{LI} \leq 0 \end{cases} \tag{3}$$

where x_{ij} denotes a point in the Max-C Map at the location (i, j) and x_{ij}^{LI} is its inhibition value. We calculate x_{ij}^{LI} within a square zone, named as lateral inhibition zone (LIZ), which LIZ is formed by the k neighboring points of x_{ij}, $\overline{x_{ij}}$ is the mean value of x_{ij} within the LIZ. The x_{uv} is a neighbor of x_{ij} in its LIZ. The Euclidean distance between x_{ij} and x_{uv} is denoted by d_{uv}, which is normalized by the length of the LIZ. And δ is a function defined as

$$\delta(x) = \begin{cases} x & \text{if } x > 0 \\ 0 & \text{if } x \leq 0 \end{cases} \tag{4}$$

In Equation (2), x_{ij}^{LI} consists of two parts, namely ***average term*** and ***differential term***. *The average term protects the neurons within a high response zone. And the differential term sharpens the objects' boundaries and increases the contrast between objects and background in the protected zone created by the average term.* The differential term includes two components, $d_{uv}e^{-d_{uv}}$ and $\delta(x_{uv} - x_{ij})$. Note that $d_{uv}e^{-d_{uv}}$, which has a function surface of an inverted "Mexican hat", is a computational model of lateral inhibition in neurobiology (Müller et al. 2005; Casanova, Buxhoeveden, and Gomez 2003; Ringach 1998). It denotes that the nearest and farthest neighbors have lowest inhibit effects on the central neuron. $\delta(x_{uv} - x_{ij})$ indicates that the inhibition is caused by the difference between the central neuron and neighboring neurons. And the central neuron would not receive inhibition from the neighboring neurons with values lower than itself. a and b are the balance coefficients. In this paper, a is set to 0.1, b is set to 0.9, and the length of LIZ is set to 7. Both Edge-enhancement and Mach band effects can be derived from Equation (2). Finally, the new output y_{ij} is determined by gating operation described in Equation (3). Note that all the hyper-parameters mentioned above are selected intuitively. It is possible that better performance can be produced by carefully selecting these parameters with more experiments empirically.

An example is illustrated in Fig. 4. A VggNet (VGG16) (Simonyan and Zisserman 2014) pre-trained on ImageNet2012 is adopted from Caffe (Jia et al. 2014) Model Zoo. Our lateral inhibition model is applied to each ReLU layer in this VggNet. Without loss of generality, we present the inhibition procedure of a middle layer called "relu4_3". The input image and the final attention map are shown in Fig. 4(a). To visualize the original gradients, we calculate the summation of the gradient cub along the channel axis, as shown in Fig. 4(b). The resulting map is very noisy. Fig. 4(c) illustrates the Max-C Map generated by inhibition along channels. It reveals a trace to infer about the related object. But it also contains lots of noise, and these noise will ***spread rapidly*** during the hierarchical top-down back-propagation. Fig. 4(d) and (e) demonstrate the inhibition effects produced by the average term and differential term. The average term creates a protected zone for the target objects and suppresses small noise or irrelevant values. The differential term imposes a heavy penalty on the background around the boundaries of target objects. Consequently, the edges of the objects are enhanced. Fig. 4(f) shows the final result of "relu4_3" layer by combining both the average term and differential term. *The example provides the intuitive evidence that, the proposed lateral inhibition model can effectively sharpen the object edges, suppress noise and increase the contrasts between target objects and background.*

Category-specific Attention Maps

Actually, CW is a kind of weak and vague top-down signal for measuring how much the neurons in hidden layers contribute to a class node. It will suffer from interference and noise, and decay layer by layer. However, benefiting from the hierarchical lateral inhibition mechanism, ***LICNN eliminates interference and noise layer by layer*** and acquires the precise location information of expected objects. That is, it can be used to produce the accurate category-specific attention maps. For clarity, we summarize the involved operations in Algorithm 1.

Specifically, after the first time of applying lateral inhibition, we block the suppressed neurons and perform another feed-foward and category-specific back-propagation to generate attention maps. For comparative analysis, Fig. 3 shows the responses and gradients in high, medium and low level layers of the original VggNet and LICNN, respectively. To visualize a multi-channel layer, we generate a map by summing the outputs along channels ***(SUM-C map)***. From the visual results, it can be seen that, *owing to the lateral inhibition, different objects are highlighted in each hidden layer with less noise and better contrast*, even though the "dog" is very close to the "person" in the image. Specially, Fig. 3(e) and Fig. 3(i) are obtained by setting the positive values in the gradient cub of the input image as one, others as zero. The results indicate that objects can be precisely located, while irrelevant information and noise in each hidden layer can be effectively suppressed.

We resize the response and gradient SUM-C maps to the same size as the input image and then combine all the resized SUM-C maps together simply by adding operation. The category-specific ***response-based or gradient-based*** attention maps can be finally derived by normalizing the combined SUM-C maps with L2 norm. Note that, *the gradient-based attention maps are essentially generated by measuring the contribution of activated patterns, which account for selective attention, while the response-based attention maps depend on the strength of input patterns, which are more related to saliency.* Although, they both capture the location information, the distribution of intensity is slightly different. We utilize the gradient-based attention maps for the top-down attention task, and the response-based attention maps for salient object detection.

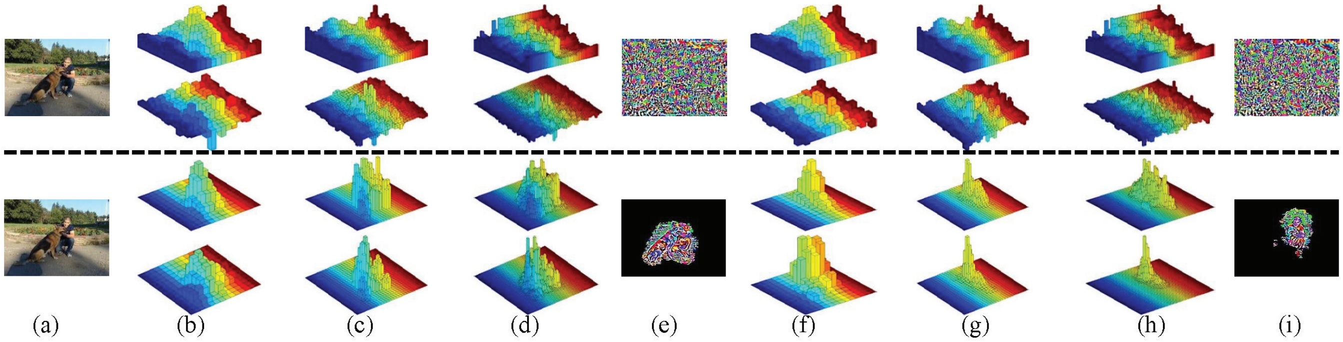


Figure 3: Comparison of responses and gradients. The first two rows are response and gradient SUM-C maps of the original VggNet. The third and forth rows are the results of LICNN. (a) Input image. (b)(c)(d) Response and gradient SUM-C maps of layer "relu5_2","relu4_1"and"relu1_1" for the dog category. (f)(g)(h) Response and gradient SUM-C maps of layer "relu5_2","relu4_1"and"relu1_1" for the person category. (e)(i) Non-zero gradients of the input image for dog and person, respectively.

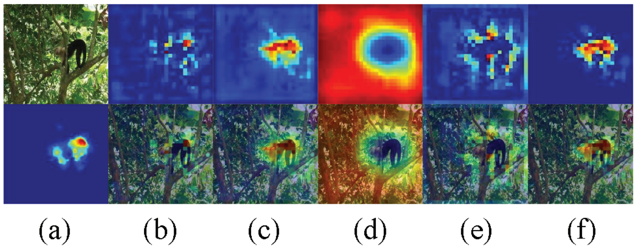


Figure 4: Visualization of the lateral inhibition model. (a) The original input image with the corresponding final category-specific attention map below it. (b) Original gradient sum of the layer "relu4_3". (c) Max-C Map obtained by the first step of our approach. (d) Inhibition from the average term. (e) Inhibition from the differential term. (f) Result of "relu4_3" layer after lateral inhibition.

Algorithm 1 Implementation details of LICNN

```
1: Given an image and a pre-trained CNN classifier;
2: Perform feed-forward and obtain a predicted category, then carry out gradient back-propagation of the predicted category to produce CWs for all neurons;
3: In each ReLU layer, compute the Max-C Map and apply the lateral inhibition model on the obtained Max-C Map;
4: Block the neurons across all channels on the inhibited locations by fixing them to zeros, then perform feed-forward and category-specific gradient back-propagation again;
5: Calculate and normalize the SUM-C maps of both response and gradients of each layer;
6: Resize all the gradient (or response) SUM-C maps to the same size of the input image and combine them together by adding operation;
7: Obtain gradient-based (or response-based) attention maps by normalizing the combined gradient (or response) SUM-C maps with L2 norm;
```

Experiments

The Pointing Game

In this section, we evaluate the discriminative power of the top-down attention maps generated by LICNN. Here Pointing Game (Zhang et al. 2016) is adopted, and the test set of PASCAL VOC 07 is used with 4952 images. We follow the protocol of Excitation Backprop (Zhang et al. 2016) for fair comparison. The same VGG16 (Simonyan and Zisserman 2014) model as mentioned above is employed. The model is fine-tuned on PASCAL VOC07 training set with the multi-label cross-entropy loss. We compare LICNN with the following methods: Excitation Backprop (c-MWP) (Zhang et al. 2016) which is the recent best-performing approach for category-specific attention, error back-propagation (Grad) (Simonyan, Vedaldi, and Zisserman 2013), and deconvolutional neural network for neuron visualization (Deconv) (Zeiler and Fergus 2014). The top-down attention maps are produced according to the ground truth object labels. Bounding boxes of the PASCAL VOC07 test set are only utilized to calculate the accuracy. We extract the maximum point on a category-specific attention map as the final prediction. A hit is counted if the maximum point falls into one of the annotated instances of the corresponding object category, and otherwise a miss is counted. The localization accuracy is measured by $Acc = \frac{Hits}{Hits+Misses}$ for each category. The mean accuracy across all categories is reported. To obtain more rigorous results, we imitate the test protocol in Excitation Backprop (Zhang et al. 2016) to select a difficult subset from the whole test set. The images in the difficult subset meet two criteria: 1) The total area of bounding boxes of the testing category is smaller than $1/4$ area of the image; and 2) there is at least one other distracter category in the image. The experimental results are reported in Table 1.

To be more convincing, we further embed our LI model into the Googlenet (Szegedy et al. 2015) to conduct the pointing game experiment. The results are reported in Table 2. Note that we also compare LICNN with the method CAM proposed in (Zhou et al. 2015). As suggested by all

	Center	Grad	Deconv	c-MWP	LICNN
ALL	69.5	76.0	75.5	80.0	**85.3**
Difficult	42.6	56.8	52.8	66.8	**70.0**

Table 1: Mean accuracy (%) of the Pointing Game on the test set of VOC07 by using VggNet. We report the results for the whole test set and a difficult subset respectively. The Center method is the baseline in which the central points of images are directly used.

	Center	Grad	CAM	c-MWP	LICNN
ALL	69.5	79.3	80.8	85.0	**87.9**
Difficult	42.6	61.4	61.9	72.3	**75.7**

Table 2: Mean accuracy (%) of the Pointing Game on the test set of VOC07 by using GoogleNet.

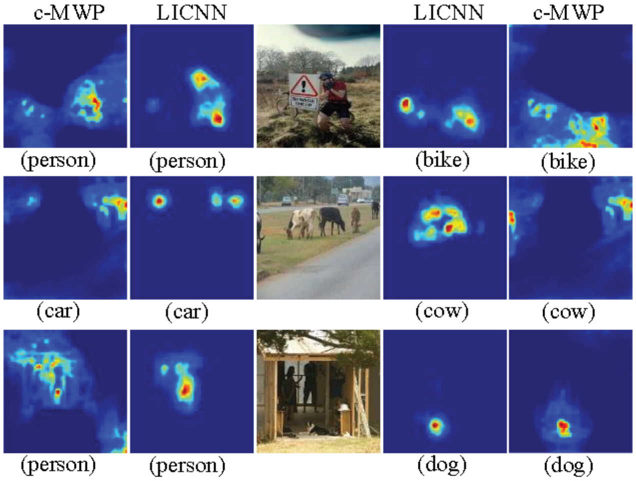


Figure 5: Visual comparison between LICNN and Excitation Backprop (c-MWP). Note that the attention maps of c-MWP and LICNN of different objects are shown on the left and right sides of input images, respectively.

the above results, LICNN significantly outperforms the compared methods with a large performance gap.

A visual comparison between LICNN and c-MWP is shown in Fig. 5. All the input images in Fig. 5 contain objects from two categories of PASCAL VOC. And the attention maps of c-MWP are produced by using the same Vgg model and the code provided by the authors. From the results, our method generates more accurate attention maps with less noise in various challenging cases. Both the qualitative and quantitative experiments support that LICNN performs very well for the category-specific attention.

Salient Object Detection

In this section, we explore the capacity of LICNN to capture salient objects from natural images. As mentioned before, *a CNN classifier can be regarded as a **salient pattern detector***, where many patterns will be activated when given an input. They compete with each other during the feed-forward phase and eventually contribute to one or more classes, leading to the distinct scores over different categories. Salient objects can be acquired by ***assembling different salient patterns*** according to the top-k categories via LICNN. In this paper, k is set to 5. Moreover, due to *the generalization ability of learned patterns*, LICNN can be applied to ***any natural image***, no matter it contains the predefined objects, i.e., the class labels in training the CNN, or not.

Datasets and evaluation criteria. We evaluate our method on four popular datasets which are widely used to evaluate salient object detection methods based on deep learning, including HKU-IS (Li and Yu 2016), PASCAL-S (Li et al. 2014), ECSSD (Yan et al. 2013), and DUTOMRON (Yang et al. 2013). HKU-IS is a large dataset containing 4447 challenging images, most of which have either low contrast or multiple salient objects. PASCAL-S contains 850 images and is built using the validation set of the PASCAL VOC 2010 segmentation challenge. ECSSD contains 1,000 structurally complex images collected from the Internet. And Dut-OMRON is composed of 5,168 challenging images, each of which has one or more salient objects with complex background. We evaluate all methods using maximum F-measure(maxF) and mean absolute error (MAE) as in (Li and Yu 2016).

Implementation. We also obtain the VGG16 (Simonyan and Zisserman 2014) model (pre-trained on Imagenet 2012) from the Caffe Model Zoo website as the classification model of LICNN. By LICNN, we produce 5 category-specific response-based attention maps according to the predicted top-5 categories. As a result, the generated maps highlight the different components of the objects. By combining the 5 maps, we obtain a rough saliency map. Fig. 2 illustrates several examples. Furthermore, to get better results, a simple optimization technique in (Zhu et al. 2014) is adopted. Specifically, we simply set the obtained saliency map SM as the foreground weight matrix, and acquire the background weight matrix by the pixels in $1 - SM$ with the value lower than 50th percentile. The refined saliency maps are used as the final output.

Comparison with the state-of-the-art. We compare our method against several recent state-of-the-art methods, including the traditional methods DRFI (Jiang et al. 2013), wCtr* (Zhu et al. 2014), RC (Cheng et al. 2015), BSCA (Qin et al. 2015), PISA (Wang et al. 2015a), and the strongly supervised CNN based approaches **LEGS** (Wang et al. 2015b), **MC** (Zhao et al. 2015) and **MDF** (Li and Yu 2016). To analyze the importance of LI, we also report three baseline results: In Baseline 1 (B1), LI is turned off (only using CNN with gradients information); In Baseline 2 (B2), we apply average denoising algorithm on the Max-C map and then handle the denoised Max-C map by a thresholding with its mean value; And in Baseline 3 (B3), we turn off the optimization technique (Zhu et al. 2014). We report the quantitative comparison w.r.t. maximum F-measure and MAE in Table 3.

From the results, our method significantly outperforms the traditional approaches. Compared with supervised methods, LICNN achieves much better performance than LEGS and is comparable with MC and MDF. Note that our method

Data Set	Metric	B1	B2	B3	DRFI	wCtr*	RC	BSCA	PISA	**LEGS**	**MC**	**MDF**	**Ours**
HKU-IS	maxF	0.510	0.759	0.798	0.776	0.726	0.726	0.723	0.753	0.770	0.798	0.861	0.841
	MAE	0.377	0.289	0.161	0.167	0.140	0.165	0.174	0.127	0.118	0.102	0.076	0.101
PASCAL-S	maxF	0.46	0.681	0.710	0.690	0.655	0.644	0.666	0.660	0.752	0.740	0.764	0.755
	MAE	0.395	0.322	0.212	0.210	0.201	0.227	0.224	0.196	0.170	0.145	0.146	0.162
ECSSD	maxF	0.477	0.748	0.805	0.782	0.716	0.738	0.758	0.764	0.827	0.837	0.847	0.831
	MAE	0.396	0.302	0.180	0.170	0.171	0.186	0.183	0.150	0.137	0.100	0.106	0.129
DUT-OMRON	maxF	0.360	0.534	0.613	0.664	0.630	0.599	0.617	0.630	0.669	0.703	0.694	0.677
	MAE	0.347	0.378	0.154	0.150	0.144	0.189	0.191	0.141	0.133	0.088	0.092	0.138

Table 3: Comparison of quantitative results including maximum F-measure (the larger is the better) and MAE (the smaller is the better). The best three results are shown in red, blue, and green color, respectively. Note that **LEGS**, **MC**, and **MDF** are strongly supervised CNN based approaches, and our method is based on weak supervision cues.

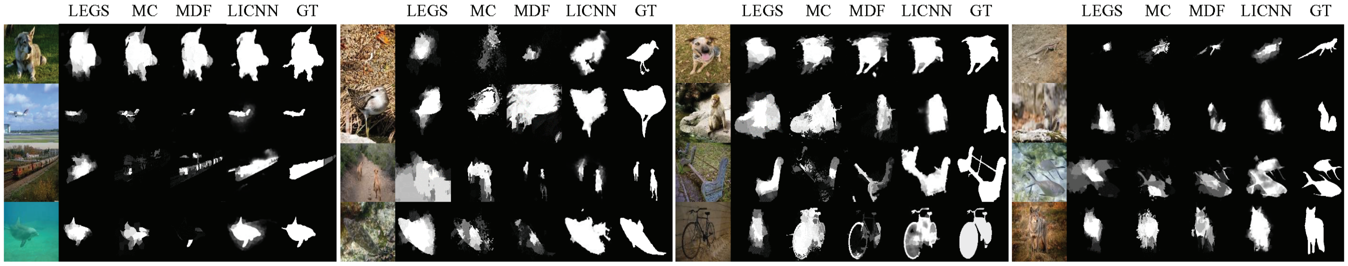


Figure 6: Visual comparison of saliency maps produced by the state-of-the-art CNN based methods, including LEGS, MC, MDF and our LICNN model. The ground truth (GT) is shown in the last column.

is merely based on a pre-trained VGG classifier which only requires the image-level class labels. In contrast, LEGS, MC and MDF strongly rely on the manually labeled segmentation masks of salient objects for model learning. And the three baseline experiments demonstrate the effectiveness of the proposed LI model. Moreover, a visual comparison between LEGS, MC, MDF and LICNN is presented in Fig. 6. LICNN performs much better in difficult cases, e.g., low contrast between objects and background.

ALL these results indicate that the bottom-up salient object detection can be carried out by grouping salient patterns via lateral inhibition. With the help of feedback signals, LICNN can organize different activated patterns effectively and merge them to form different objects. Thus, salient objects can be well perceived by a CNN for classification with only category-level labels.

Conclusion

In this paper, we have proposed a lateral inhibition based attention model, namely LICNN. In contrast to other methods, the LICNN can simultaneously perform object recognition, top-down selective attention, and salient object detection. We reveal that combining lateral inhibition and top-down feedback can construct a competitive environment to ensure that only the most discriminative and salient features are selected. With LICNN, highly discriminative category-specific attention maps are produced, and salient objects are effectively obtained without learning an independent model based on strong segmentation supervision. Both qualitative and quantitative experimental results strongly support the effectiveness of LICNN.

LICNN is an important attempt of modeling visual attention with feedback and lateral inhibition. It provides a new insight to implement brain-inspired concepts, which we believe represents a more promising route in future studies on designing vision algorithms.

Acknowledgments

This work is jointly supported by National Key Research and Development Program of China (2016YFB1001000), National Natural Science Foundation of China (61525306, 61633021, 61572504, 61420106015, 61673362),Youth Innovation Promotion Association CAS (2006121), Beijing Nova Program and Beijing Natural Science Foundation (4162058). This work is also supported by grants from NVIDIA and the NVIDIA DGX-1 AI Supercomputer.

References

Cao, C.; Liu, X.; Yang, Y.; Yu, Y.; Wang, J.; Wang, Z.; Huang, Y.; Wang, L.; Huang, C.; Xu, W.; et al. 2015. Look and think twice: Capturing top-down visual attention with feedback convolutional neural networks. In *Proceedings of the IEEE International Conference on Computer Vision*, 2956–2964.

Casanova, M. F.; Buxhoeveden, D.; and Gomez, J. 2003. Disruption in the inhibitory architecture of the cell minicolumn: implications for autisim. *The Neuroscientist* 9(6):496–507.

Chen, Y., and Fu, H. 2010. Study and application of lateral inhibition models in image's contour enhancement. In *2010*

International Conference on Computer Application and System Modeling (ICCASM 2010), volume 13, V13–23. IEEE.

Cheng, M.-M.; Mitra, N. J.; Huang, X.; Torr, P. H.; and Hu, S.-M. 2015. Global contrast based salient region detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 37(3):569–582.

Cheng, Y.; Cai, R.; Li, Z.; Zhao, X.; and Huang, K. 2017. Locality-sensitive deconvolution networks with gated fusion for rgb-d indoor semantic segmentation. In *The IEEE Conference on Computer Vision and Pattern Recognition*.

Desimone, R., and Duncan, J. 1995. Neural mechanisms of selective visual attention. *Annual Review of Neuroscience* 18(1):193–222.

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

Fernandes, B. J. T.; Cavalcanti, G. D.; and Ren, T. I. 2013. Lateral inhibition pyramidal neural network for image classification. *IEEE transactions on cybernetics* 43(6):2082–2092.

Fernández-Caballero, A.; López, M. T.; Serrano-Cuerda, J.; and Castillo, J. C. 2014. Color video segmentation by lateral inhibition in accumulative computation. *Signal, Image and Video Processing* 8(6):1179–1188.

Jia, Y.; Shelhamer, E.; Donahue, J.; Karayev, S.; Long, J.; Girshick, R.; Guadarrama, S.; and Darrell, T. 2014. Caffe: Convolutional architecture for fast feature embedding. *arXiv preprint arXiv:1408.5093*.

Jiang, H.; Wang, J.; Yuan, Z.; Wu, Y.; Zheng, N.; and Li, S. 2013. Salient object detection: A discriminative regional feature integration approach. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2083–2090.

Li, G., and Yu, Y. 2016. Visual saliency detection based on multiscale deep cnn features. *IEEE Transactions on Image Processing* 25(11):5012–5024.

Li, Y.; Hou, X.; Koch, C.; Rehg, J. M.; and Yuille, A. L. 2014. The secrets of salient object segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 280–287.

Mao, Z.-H., and Massaquoi, S. G. 2007. Dynamics of winner-take-all competition in recurrent neural networks with lateral inhibition. *IEEE transactions on neural networks* 18(1):55–69.

Müller, N. G.; Mollenhauer, M.; Rösler, A.; and Kleinschmidt, A. 2005. The attentional field has a mexican hat distribution. *Vision Research* 45(9):1129–1137.

Nielsen, C. J. 2001. Effect of scenario and experience on interpretation of mach bands. *Journal of Endodontics* 27(11):687–691.

Qin, Y.; Lu, H.; Xu, Y.; and Wang, H. 2015. Saliency detection via cellular automata. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 110–119.

Ringach, D. L. 1998. Tuning of orientation detectors in human vision. *Vision Research* 38(7):963–972.

Simonyan, K., and Zisserman, A. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Simonyan, K.; Vedaldi, A.; and Zisserman, A. 2013. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*.

Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S.; Anguelov, D.; Erhan, D.; Vanhoucke, V.; and Rabinovich, A. 2015. Going deeper with convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1–9.

Szlam, A. D.; Gregor, K.; and Cun, Y. L. 2011. Structured sparse coding via lateral inhibition. In *Advances in Neural Information Processing Systems*, 1116–1124.

Wang, K.; Lin, L.; Lu, J.; Li, C.; and Shi, K. 2015a. Pisa: Pixelwise image saliency by aggregating complementary appearance contrast measures with edge-preserving coherence. *IEEE Transactions on Image Processing* 24(10):3019–3033.

Wang, L.; Lu, H.; Ruan, X.; and Yang, M.-H. 2015b. Deep networks for saliency detection via local estimation and global search. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 3183–3192.

Wyatte, D.; Herd, S.; and Mingus, B. 2012. The role of competitive inhibition and top-down feedback in binding during object recognition. *Inhibition in the Process of Feature Binding* 7.

Yan, Q.; Xu, L.; Shi, J.; and Jia, J. 2013. Hierarchical saliency detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1155–1162.

Yang, C.; Zhang, L.; Lu, H.; Ruan, X.; and Yang, M.-H. 2013. Saliency detection via graph-based manifold ranking. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 3166–3173.

Zeiler, M. D., and Fergus, R. 2014. Visualizing and understanding convolutional networks. In *European Conference on Computer Vision*, 818–833.

Zhang, J.; Lin, Z.; Brandt, J.; Shen, X.; and Sclaroff, S. 2016. Top-down neural attention by excitation backprop. In *European Conference on Computer Vision*, 543–559.

Zhao, R.; Ouyang, W.; Li, H.; and Wang, X. 2015. Saliency detection by multi-context deep learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1265–1274.

Zhou, B.; Khosla, A.; Lapedriza, A.; Oliva, A.; and Torralba, A. 2015. Learning deep features for discriminative localization. *arXiv preprint arXiv:1512.04150*.

Zhu, W.; Liang, S.; Wei, Y.; and Sun, J. 2014. Saliency optimization from robust background detection. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2814–2821.