Feedback Convolutional Neural Network for Visual Localization and Segmentation

Chunshui Cao, Yongzhen Huang, *Senior Member, IEEE*, Yi Yang, *Member, IEEE*, Liang Wang, *Senior Member, IEEE*, Zilei Wang, *Member, IEEE*, and Tieniu Tan, *Fellow, IEEE*

Abstract—Feedback is a fundamental mechanism existing in the human visual system, but has not been explored deeply in designing computer vision algorithms. In this paper, we claim that feedback plays a critical role in understanding convolutional neural networks (CNNs), e.g., how a neuron in CNNs describes an object's pattern, and how a collection of neurons form comprehensive perception to an object. To model the feedback in CNNs, we propose a novel model named Feedback CNN and develop two new processing algorithms, i.e., neural pathway pruning and pattern recovering. We mathematically prove that the proposed method can reach local optimum. Note that Feedback CNN belongs to weakly supervised methods and can be trained only using category-level labels. But it possesses a powerful capability to accurately localize and segment category-specific objects. We conduct extensive visualization analysis, and the results reveal the close relationship between neurons and object parts in Feedback CNN. Finally, we evaluate the proposed Feedback CNN over the tasks of weakly supervised object localization and segmentation, and the experimental results on ImageNet and Pascal VOC show that our method remarkably outperforms the state-of-the-art ones.

Index Terms—feedback, convolutional neural networks (CNNs), weakly supervised, object localization, object segmentation

1 Introduction

Visual attention is mainly dominated by *"goals"* from our mind in a top-down manner, especially in the case of object detection. Cognitive science explains this mechanism in the "Biased Competition Theory" [1]: human visual cortex would be enhanced by top-down stimuli, and non-relevant neurons will be suppressed in feedback loops when searching for objects. This process actually contains the selectivity of neuron activations [2], which reduces the chance of recognition to be interfered by either noise or distractive patterns.

Inspired by the above evidence, in this paper we propose a novel *Feedback Convolutional Neural Network* (Feedback CNN) architecture to imitate such selectivity. Specifically, we propose to jointly reason the outputs of class nodes and the activations of hidden layer neurons in the feedback loop. Fig. 1 illustrates the main idea of Feedback CNN. The proposed network does the inference for input images in a bottom-up manner, as in traditional Convolutional Neural Networks [3], [4], [5]. Then high-level semantic labels (e.g., outputs of class nodes) would be produced and they are set as the "goals" in visual search. Finally, we select the target-relevant neurons by pruning the neural pathway in feedback loops. To capture the object of interests, it is in the pixel space to reconstruct the objects by recovering all patterns carried by the selected target relevant neurons. In this work, we show that the Feedback CNN is effective for visualization of classification models, object localization, and semantic segmentation.

Specifically, we propose a simple yet efficient method to analyze image compositions represented by Convolutional Neural Networks, and then assign neuron activations given by goals during visual search. Inspired by Deformable Part-Based Models (DPMs) [6] that model middle level part locations as latent variables and search for them during object detection, we introduce *latent gate-variables* to control the effects of hidden neurons. Then we formulate the feedback computation as an optimization problem and we develop two new algorithms to solve it, i.e., Feedback Selective Pruning (FSP) and Feedback Recovering (FR). The two proposed algorithms both maximize the response of network output to the target high-level semantic concepts in a top-down manner. More specifically, FSP focuses on selecting the target-relevant neurons in the hidden layers, and FR is able to restore visual pattern information in the receptive field of a certain neuron. By combining FSP and FR, Feedback CNN can effectively produce the task-specific gradient maps which allow us to obtain visualization maps and energy maps with high quality. In particular, we visualize several

- C. Cao is with the University of Science and Technology of China, Hefei Shi 230000, China, and also with Center for Research on Intelligent Perception and Computing (CRIPAC), National Laboratory of Pattern Recognition (NLPR), Institute of Automation, Chinese Academy of Sciences (CASIA), Beijing 100864, China. E-mail: ccs@mail.ustc.edu.cn.
- Z. Wang is with the University of Science and Technology of China, Hefei Shi 230000, China. E-mail: zlwang@ustc.edu.cn.
- Y. Huang, L. Wang and T. Tan are with the University of Chinese Academy of Sciences, Huairou 101408, China, and are also with Center for Research on Intelligent Perception and Computing, National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing 100864, China. E-mail: {yzhuang, wangliang, tnt}@nlpr.ia.ac.cn.
- Y. Yang is with Baidu Research, Sunnyvale, CA 94089, USA. E-mail: yangyi05@baidu.com.

Manuscript received 21 Dec. 2016; revised 13 Sept. 2017; accepted 26 Apr. 2018. Date of publication 0 . 0000; date of current version 0 . 0000.
(Corresponding author: ~~Chunshui Cao~~.)
Recommended for acceptance by D. Ramanan.
For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org, and reference the Digital Object Identifier below.
Digital Object Identifier no. 10.1109/TPAMI.2018.2843329

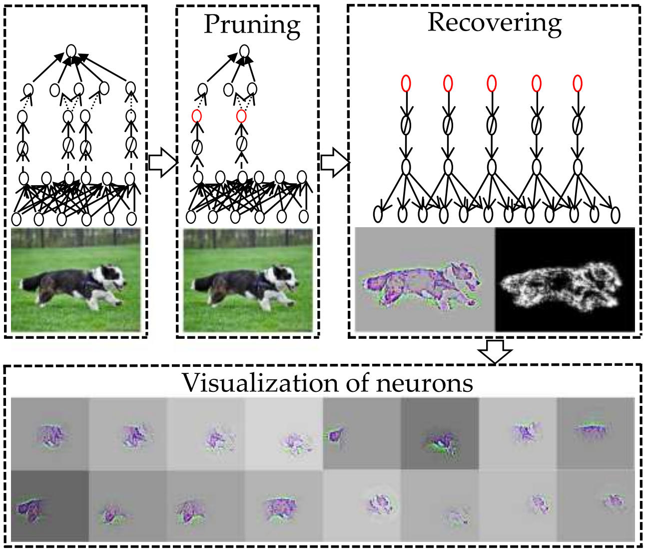


Fig. 1. Feedback CNN. Given an input image, we perform a normal feedforward to predict the class label and set it as the target. Then use the pruning operation to select related neurons, and perform the recovering operation on these selected neurons to obtain target-relevant visualization and energy maps. Each selected neuron is highly related to object parts, which is shown by visualizing the selected neurons respectively.

exemplar neurons that are selected by FSP in Fig. 1. It can be seen that the selected neurons are highly relevant to the target object and correspond to different parts of the object. As a consequence, Feedback CNN can select the target-relevant neurons and suppress the irrelevant ones by the top-down inference, which makes the model mainly respond to the most salient regions of images that are highly related to the target category.

Accordingly, Feedback CNN enables a CNN for object classification to localize and segment the interested objects in natural images, as illustrated in Fig. 2. Specifically, a CNN classifier performs feedforward inference for an input image as usual. The predicted category, e.g., "Train" for the first image, is set as the "goal" for following feedback in Fig. 2c, where only the neurons associated with "Train" would be activated. As a result, in Figs. 2d and 2e, only the salient regions related to "Train" are highlighted in the visualization and energy maps. With the help of these maps, it is easy to localize and segment target objects in images, as shown in Figs. 2f and 2g, and the whole process only needs weakly supervised class annotation for training. As suggested by these results, the feedback networks *provide important flexibility to Convolutional Networks towards integrating object recognition, localization, and segmentation into a unified framework.*

A preliminary version of this work was reported in [7]. Compared with [7], apart from more comprehensive description, analysis, and experiments, this paper develops two new algorithms and gives the mathematical proofs on achieving local optimum. Consequently, the methods in this paper present much stronger capability for task-specific neuron selection and object capturing. Furthermore, we can obtain energy maps with higher signal-to-noise ratio and clearer object boundaries. We believe that this work paves a way for weakly supervised object localization and segmentation. By contrast, the previous model in [7] suffers from more noise and is confined to weakly supervised object localization.

The main contributions of this paper are summarized as follows: 1) We develop two novel algorithms (i.e., FSP and FR) to model the feedback mechanism in CNNs, and provide mathematical proofs on achieving local optimum. 2) We demonstrate that the proposed Feedback CNN has the capability to select the neurons associated with goal objects through extensive visualization. 3) We apply Feedback CNN to weakly supervised object localization and segmentation, and obtain significant performance improvement compared with previous state-of-the-art methods.

2 Related Work

2.1 Deep CNNs

In recent years, it has been witnessed the great success of deep CNNs in various computer vision tasks [3], [4], [5], [8], [9]. Particularly, deep CNNs have basically achieved human-level performance for object recognition [3], [4], [5]. Studies in [10], [11] show that the convolutional units in CNNs that are trained only for the purpose of classification have the potential to learn a part of semantic patterns, e.g., object parts. The discriminative ability of deep CNNs can be further improved by some approaches, such as dropout [12], skip connections [5], and batch normalization [13]. Moreover, many researchers take considerable interests in enhancing deep CNNs to possess greater capacity by making the networks deeper or wider [3], [5], [14].

The great progresses of CNNs provide a solid foundation for constructing a feedback model in CNNs. By introducing the feedback mechanism, it is expected that object localization and semantic segmentation can be conducted more easily, especially under weakly supervised conditions.

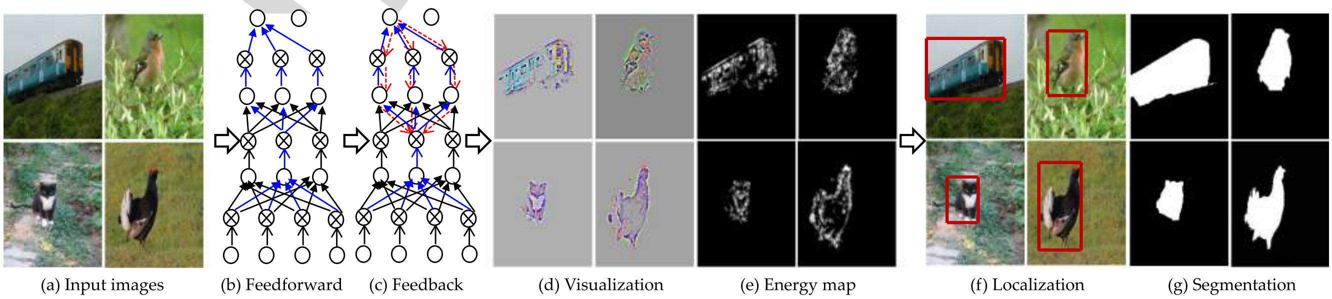


Fig. 2. A simple pipeline for object localization and segmentation via the proposed Feedback CNN model. (a)(b)(c) When given an input image, the proposed Feedback CNN is designed to utilize both bottom-up image inputs and top-down semantic labels to infer the hidden neuron activations. (d)(e)(f)(g) Salient areas captured in the visualization and energy maps by feedback often correspond to related target objects. And based on these maps, objects can be easily localized and segmented from the input image. Best viewed in color.

2.2 Top-Down Feedback

Top-down Feedback is one of the important mechanisms in the human visual system that plays a critical role in many visual tasks, e.g., objects localization and segmentation, feature grouping, perceptual filling, and tuning receptive fields of neurons [15]. Before our work, some efforts have been made to embed the feedback mechanism into deep neural networks. The convolutional latent variable models (CLVMs) in [16] take feedback by treating units as the latent variables of a global energy function. [17], [18] invert the learned convolutional neural networks for understanding deep image representations. The DasNet [19] adds a feedback structure that can dynamically alter the sensitivities of convolutional filters during classification, where the feedback mechanism is learned via reinforcement learning. [20] introduces the top-down module to incorporate fine details into the detection framework. Recently, [21] presents a feedback based learning model. The key idea is to make predictions based on a notion of the thus-far outcome in an iterative manner. An earlier study is presented in Deep Boltzmann Machines (DBM) for feature selection [22]. Meanwhile, Recurrent Neural Networks (RNNs) [23] and Long Short-Term Memory (LSTM) [24] are explored to capture attention drifting in a dynamic environment. Deconvolutional Neural Networks [10] attempt to formulate feedback as a reconstruction problem in the training phase.

In this work, we propose to formulate feedback as an optimization problem for neuron selection. Different from previous works, our proposed feedback is used to selectively modulate the status of hidden neurons during the testing phase. Thus it does not affect the training procedure of CNNs, i.e., many sophisticated models can be directly adopted.

2.3 Weakly-Supervised Object Localization and Segmentation

In recent years, many methods have been developed for weakly supervised object localization based on CNNs [25], [26], [27], [28], [29], [30]. For example, [28] proposes a self-teaching method for object localization. [26], [27] propose to use the global average pooling and max pooling to generate class-specific energy maps for localizing objects. [25] proposes to segment objects in an image using the noisy energy map generated by class-specified gradients. [29] proposes to retrain the recognition model after embedding the average pooling layer. [30] employs a probabilistic "winner-takes-all" process, in which marginal winning probability is computed by taking activation values and positive convolutional weights. The energy maps generated by [29] and [30] mainly highlight the most discriminative parts of objects while losing fine details on object boundaries, which consequently suffer from noise and interference.

Meanwhile, some other approaches are proposed for weakly-supervised semantic segmentation [31], [32], [33], [34], [35]. The approaches presented in [31] and [34] train deep networks using multiple instance learning and adopt different pooling strategies. CCNN [33] and EM-Adapt [35] develop a self-training framework and enforce the consistency between the per-image annotation and the predicted segmentation masks with different constraints.

Different from the previous methods, the proposed Feedback CNN in this paper can simultaneously perform object recognition, localization, and semantic segmentation with the same weakly supervised settings. That is, our method only needs to train a classification model, and then object localization and semantic segmentation can be automatically performed based on the energy maps generated by the proposed feedback selection mechanism. Here, the bounding boxes or segmentation masks are not required at all for the training samples.

3 Feedback CNN

3.1 Re-Interpreting ReLU and Max-Pooling

The recent state-of-the-art deep CNNs consist of many stacked feedforward layers, including convolutional, rectified linear units (ReLU), and max-pooling layers. For each layer, the input $\mathbf{x}$ can be an image or output of the previous layer, which is composed of C input channels with the width M and height N, i.e., $\mathbf{x} \in \mathcal{R}^{M \times N \times C}$. Similarly, the output $\mathbf{y}$ consists of C' output channels with the width M' and height N', i.e., $\mathbf{y} \in \mathcal{R}^{M' \times N' \times C'}$.

Convolutional Layer. The convolution layer is used to extract different features of the input, which is commonly parameterized by C' filters with the kernel $\mathbf{k} \in \mathcal{R}^{K \times K \times C}$.

$$\mathbf{y}_{c'} = \sum_{c=1}^{C} \mathbf{k}_{c'c} * \mathbf{x}_c, \ \forall c', \tag{1}$$

where $\mathbf{k}_{c'c}$ represents the convolutional kernel of the c'th filter over the cth input channel.

ReLU Layer. The ReLU layer is used to increase the nonlinear properties of the decision functions without affecting the receptive fields of convoluional layers. Formally, it is defined as

$$\mathbf{y} = \max(\mathbf{0}, \mathbf{x}). \tag{2}$$

Max-Pooling Layer. The max-pooling layer is used to reduce the dimensionality of the output, and the feature variance of deformable objects for producing the similar image representations. The max-pooling operation is applied to each pixel (i, j) by taking its small neighborhood $\mathcal{N}$, namely,

$$y_{ijc} = \max_{u,v \in \mathcal{N}} x_{i+u, j+v, c}, \ \forall i, j, c. \tag{3}$$

Here y_{ijc} represents the pixel value of (i, j) over the cth output channel.

Selectivity in Feedforward Network. To understand the selectivity mechanism in neural networks and formulate the feedback, we re-interpret the behaviors of ReLU and max-pooling layers by introducing a set of binary activation variables $\mathbf{z} \in \{0, 1\}$ instead of the $\max()$ operations in Equations (2) and (3). In particular, we formulate the behaviors of ReLU and max-pooling as $\mathbf{y} = \mathbf{z} \circ \mathbf{x}$, where $\circ$ denotes the element-wise product (Hadamard product); and $\mathbf{y} = \mathbf{z} * \mathbf{x}$, where $*$ denotes the convolution operator and $\mathbf{z}$ is a set of convolutional filters except that they are location variant.

By interpreting the ReLU and max-pooling layers as "gates" controlled by input $\mathbf{x}$, the network selects the important information in a *bottom-up* manner during the feedforward phase, and then eliminates the signals with minor contributions to predictions. However, for a pre-trained

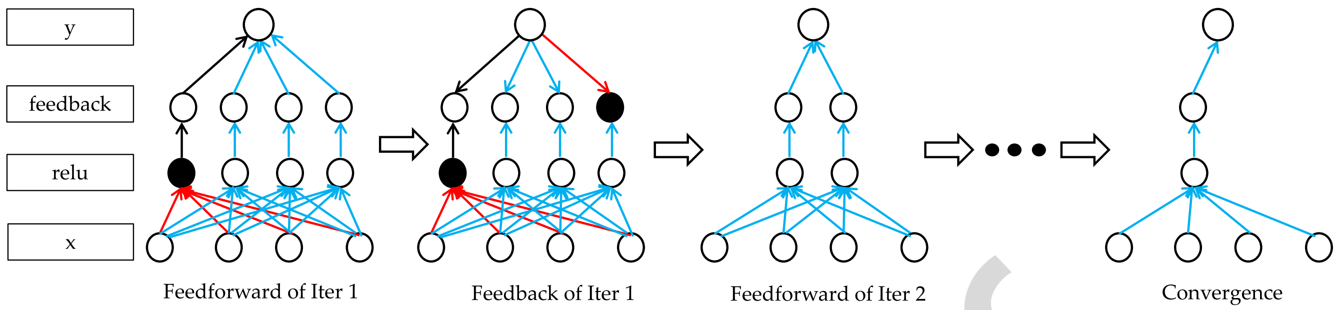


Fig. 3. Illustration of our feedback model and its inference process. At the first iteration, the model performs as a feedforward neural network. Then, the neurons in the feedback hidden layers update their activation status to maximize the confidence output of the target top neuron. This process continues until convergence. Note that the black nodes represent neurons that are not activated or turned off in the feedback loop. (We show only one layer here, but feedback layers can be tacked in the deep CNNs.)

network model, the activations of some neurons could be harmful to image classification since they may involve irrelevant noise, e.g., cluttered backgrounds in complex scenes. This could be one of the reasons that most of CNN classifiers on ImageNet have relatively low top-1 accuracies.

3.2 Introducing Feedback Layer

For a given neural network model, most of the gates acquiesce to be opened so that maximum amount of information can pass through the network for generalization. When targeting at a particular semantic label, however, we can increase the discriminativeness of features by turning off the gates that provide irrelevant information. Such a strategy is explained as the neuron selectivity in the Biased Competition Theory [1], and is critical to implement the top-down attention. Moreover, the evidences from [36] show that the performance of both human recognition and detection can be increased significantly by the goal-directed selectivity after a first time glimpse. A method called ”look and think twice” in [7] mimics this process and consequently the CNN prediction accuracy is effectively boosted.

Technically, to increase the flexibility of models to images and prior knowledge, we introduce an extra layer called *feedback layer* to the existing convolutional neural networks. The feedback layer contains a set of binary variables $\mathbf{Z} \in \{0, 1\}$ to represent activation status of neurons. In practice, these binary variables are determined by top-down messages from outputs rather than inputs. The feedback layer is stacked upon each ReLU layer. Then the feedback and ReLU layers form a hybrid control unit to neuron response, which indeed combines the bottom-up and top-down messages:

- Bottom-Up: Inherent selectivity from *ReLU layers*, and the dominant features would be passed to the upper layers;
- Top-Down: Controlled by *Feedback Layers*, which propagate the high-level semantic information back to image representations. Only the gates associated with the target neurons would be activated.

Fig. 3 illustrates a simple architecture of our feedback model with only one ReLU layer and one feedback layer.

3.3 Problem Formulation

In this paper, we formulate the feedback mechanism as an optimization problem by introducing additional control gate-variables $\mathbf{Z}$. Given an image I and a neural network with learned parameters w, we optimize the output of the target neuron by jointly inferring the binary neuron activation $\mathbf{Z}$ over all the hidden feedback layers. In particular, if the target neuron is a tth class node in the top layer, we maximize the class score $S_t(I)$ by re-adjusting the activation of each neuron, namely,

$$\begin{aligned} \max_{Z} \quad & S_t(I, Z) \\ \text{s.t.} \quad & z_{ijc}^{(l)} \in \{0, 1\}, \ \forall\ l, i, j, c, \end{aligned} \tag{4}$$

where $z_{ijc}^{(l)}$ denotes the gate-variable for the neuron (i, j) of the channel c in the feedback layer l.

The formulation in (4) leads to an integer programming problem, which is NP-hard for the current non-linear deep network architecture. Here we derive locally optimal approximate solutions since S_t is linearly approximated for our considered cases.

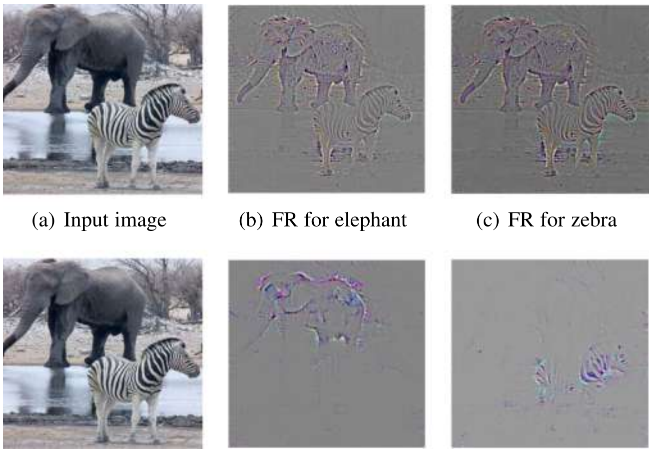
Linear Approximation. It is well known that a CNN presents a nonlinear mapping function owing to the nonlinear layers such as ReLU and max-pooling. Thus $S_t(I)$ is a highly nonlinear function about the input image I. However, given an input image I_0, we can approximate $S_t(I)$ using a linear function in the neighborhood of I_0 [25], [37], [38], [39], e.g., computing the first-order Taylor Expansion as follows:

$$S_t(I) \approx S_t(I_0) + S_t'(I_0)(I - I_0). \tag{5}$$

In this work, we implement such approximations through two layer-wise operations after the neural network finishes the regular feedforward: 1) fixing the “gate” status of the ReLU and max-pooling layers, and 2) approximating other nonlinear layers with the first order Taylor Expansion. In this case, the class score $S_t(I)$ turns to be the output of a linear neural network. After stacking the feedback layer upon each ReLU layer, the objective function in (4) is updated to a linearly nested function $S_t^*(I, Z)$. It can be expanded linearly from any feedback layer l as

$$S_t^*(I, Z) = \sum_{ijc} \alpha_{ijc}^{(l)} z_{ijc}^{(l)} x_{ijc}^{(l)}, \tag{6}$$

where $x_{ijc}^{(l)}$ is the input of the neuron (i, j) of the channel c in the feedback layer l, $z_{ijc}^{(l)}$ is the latent gate-variable, and $\alpha_{ijc}^{(l)}$ is the Contribution Weight (CW) that is determined by the neuron pathways from $z_{ijc}^{(l)}$ to the target neuron S_t.



(a) Input image (b) FR for elephant (c) FR for zebra

(d) Input image (e) FSP for elephant (f) FSP for zebra

Fig. 4. Visualizations by running the FR and FSP. (a)(d) The same input image for FR and FSP. (b)(c) Visualization of gradient maps via running FR for elephant and zebra respectively. (e)(f) Visualization of gradient maps via running FSP for elephant and zebra respectively. Best viewed in color.

From Equation (6), the feedback optimization problem is transformed as

$$\begin{aligned} \max_{Z} \quad & S_t^*(I, Z) \\ \text{s.t.} \quad & z_{ijc}^{(l)} \in \{0, 1\}, \ \forall\, l, i, j, c. \end{aligned} \tag{7}$$

Note that $x_{ijc}^{(l)}$ is the output of a ReLU neuron, namely, the constant values produced by approximating the non-linear layers before the current ReLU layer have been calculated in $x_{ijc}^{(l)}$. Thus Equation (6) does not contain a constant term.

3.4 Solutions

The objective function of the feedback optimization problem in (7) is a linearly nested function. So we can expand the objective function and use a greedy strategy to update the hidden gate-variables. Next, according to the optimization strategies, we propose two different greedy algorithms, i.e., *Feedback Recovering* and *Feedback Selective Pruning*. For convenience, we simplify the target $S_t^*(I, Z)$ as S in the following descriptions.

3.4.1 Feedback Recovering (FR)

In order to maximize S, we propose to optimize the latent gate-variables Z layer-by-layer in a top-down order. For a specific feedback layer l, the input $x_{ijc}^{(l)}$ represents a particular pattern, and the Contribution Weight $\alpha_{ijc}^{(l)}$ tells us how this input pattern contributes to the target neuron S, as demonstrated in Equation (6). Intuitively, we can reserve the $x_{ijc}^{(l)}$'s with positive CWs and remove the ones with negative CWs to maximize S, which can be implemented by updating the latent gate-variable $z_{ijc}^{(l)}$ according to the sign of $\alpha_{ijc}^{(l)}$. And then we expand the remaining $x_{ijc}^{(l)}$ to the next feedback layer $l-1$. This strategy is applied on each feedback layer in a top-down order. We summarize the above processes as Feedback Recovering in Algorithm 1. Note that here we denote the target S after updating the feedback layer l as S_l, and use the subscript k to replace i, j, c for simplicity. A sign function $\delta(x)$ is employed with $\delta(x) = 1$ for $x > 0$ and otherwise $\delta(x) = 0$. A mathematical proof of FR is provided in the appendix, which can be found on the Computer Society Digital Library at http://doi.ieeecomputersociety.org/10.1109/TPAMI.2018.2843329.

Algorithm 1. Feedback Recovering (FR)

INPUT : *image* I_0, *target neuron with score function* S
DO :
Initialize all Z *with* 1
for *iteration* $= 1$ *to max iteration* **do**
  Feedforward
  if *iteration* $== 1$ **then**
    Do Linear approximation operations
  end if
  for $l = N$ *to* 1 **do**
    if $l = N$ **then**
      $\alpha_k^{(N)} = \frac{\partial S}{\partial x_k^{(N)}}$
      $z_k^{(N)} = \delta(\alpha_k^{(N)})$
      update $\alpha_k'^{(N)} = z_k^{(N)} * \alpha_k^{(N)}$
      update $S \to S_N = \sum_k \alpha_k'^{(N)} x_k^{(N)}$
    else
      Fix $z_k^{(N)}, z_k^{(N-1)}, \ldots, z_k^{(l+1)}$
      $S_{l+1} = \sum_k \alpha_k'^{(l+1)} x_k^{(l+1)}$
      $\alpha_k^{(l)} = \frac{\partial S_{l+1}}{\partial x_k^{(l)}}$
      $z_k^{(l)} = \delta(\alpha_k^{(l)})$
      update $\alpha_k'^{(l)} = z_k^{(l)} * \alpha_k^{(l)}$
      update $S_{l+1} \to S_l = \sum_k \alpha_k'^{(l)} x_k^{(l)}$
    end if
    $l--$
  end for
end for

In order to qualitatively analyze the effect of FR, we conduct the proposed FR algorithm over the VggNet [4] that is pre-trained on the ImageNet 2012 dataset. As shown in Fig. 4, given an input image in the first column which contains an elephant and a zebra, we run FR for these two categories separately, and then the concerned objects can be highlighted.

Visualization and Energy Map. After FR achieves convergence, the back-propagation from the target neuron to the image space is performed and a gradient map can be obtained. This gradient map is a three-channel matrix. To visualize it in the RGB space, we normalize it as the visualization map by Min-Max normalization with a scale factor: 255*$\frac{x-\min}{\max-\min}$. To describe the importance of each pixel to the target category, the energy map is constructed by calculating the summation of absolute gradient values of the three channels for each pixel and normalizing the produced one-channel map by ℓ_2 normalization.

For the input image in Fig .4a, the visualization maps for the elephant and zebra are depicted in Figs. 4b and 4c. It can be observed that the FR fails to distinguish particular patterns for different target objects, but can roughly restore visual information in the receptive field of a target neuron. That is the main reason why we name Algorithm 1 as Feedback Recovering. The major cause for these results is that we sequentially update the CWs of hidden neurons in a top-down manner. A more detailed analysis will be provided in the discussion section.

3.4.2 Feedback Selective Pruning (FSP)

FR modulates CWs during the optimization processes, which causes it to lose the discriminativeness of resulting maps. In this section, we propose to update all the gate-variables Z with CWs unchanged. To this end, we compute the latent gate-variable Z in a bottom-up order. Specifically, we modulate the input $x^{(l)}_{ijc}$ to maximize the target score S. We summarize all the operations as Feedback Selective Pruning in Algorithm 2. Note that $w^{(l-1)}_{k'}$ denotes the weight between $x^{(l-1)}_{k'}$ and $x^{(l)}_{k}$ when the convolutional operation is performed from the layer $l-1$ to layer l. Similarly, the mathematical proof of FSP is provided in the appendix, available in the online supplemental material.

Algorithm 2. Feedback Selective Pruning (FSP)

INPUT : *image I_0, target neuron with score function S*
DO :
Initialize all Z with 1
for *iteration* $= 1$ *to max iteration* **do**
  Feedforward
  if *iteration* $== 1$ **then**
    Do Linear approximation operations
  end if
  for $l = 1$ *to* N **do**
    if $l = 1$ **then**
      $\alpha^{(1)}_k = \frac{\partial S}{\partial x^{(1)}_k}$
      $z^{(1)}_k = \delta(\alpha^{(1)}_k)$
      update $x'^{(1)}_k = z^{(1)}_k * x^{(1)}_k$
      update $S \to S_1 = \sum_k \alpha^{(1)}_k x'^{(1)}_k$
    else
      fix $z^{(l)}_k, z^{(2)}_k, \ldots, z^{(l-1)}_k$
      $S_{l-1} = \sum_{k'} \alpha^{(l-1)}_{k'} x'^{(l-1)}_{k'}$
      and also,
      $S_{l-1} = \sum_k \alpha^{(l)}_k x^{(l)}_k$
      $x^{(l)}_k = relu(\sum_{k'} w^{(l-1)}_{k'} z^{(l-1)}_{k'} x'^{(l-1)}_{k'})$
      $\alpha^{(l)}_k = \frac{\partial S_{l-1}}{\partial x^{(l)}_k}$
      $z^{(l)}_k = \delta(\alpha^{(l)}_k)$
      update $x'^{(l)}_k = z^{(l)}_k * x^{(l)}_k$
      update $S_{l-1} \to S_l = \sum_k \alpha^{(l)}_k x'^{(l)}_k$
    end if
    $l++$
  end for
end for

We run the FSP under the same experimental settings as for the FR, and the results are shown in Figs. 4e and 4f. From the results, the salient regions in Figs. 4e and 4f focus on different target objects. That is, the FSP algorithm is able to select target-relevant neurons in deep CNN. This is the reason why we name it as Feedback Selective Pruning. Compared with FR, therefore, FSP possesses more powerful ability to distinguish different target objects. For such ability of FSP, it is mainly because the status of gate-variables is determined by the CWs of hidden neurons and the inputs are modulated instead of the CWs during the optimization. We will provide more discussion in the next section.

3.5 Discussion

To maximize the target score S, the FR algorithm updates CWs layer-by-layer from top to bottom. During optimization, the gate-variable $z^{(l)}_{ijc}$ for $x^{(l)}_{ijc}$ is determined by the modified $\alpha'^{(l)}_{ijc}$ instead of original $\alpha^{(l)}_{ijc}$. As a consequence, it would destroy the ability of the target neuron to judge whether a pattern is beneficial to the semantic information it represents. But the FR algorithm presents a good method to visualize the content in the receptive field of a neuron.

In contrast, the FSP algorithm updates the activations of hidden neurons to maximize the target score S from bottom to top. For a particular neuron, its value $x^{(l)}_{ijc}$ may be changed because of the updating of $x^{(l-1)}_{ijc}$. However, the CW, i.e., $\alpha^{(l)}_{ijc}$ for $x^{(l)}_{ijc}$ will not be changed in a single iteration, since the status of neurons in the ReLU and max-pooling layers have been fixed after the first feedforward and the network is optimized in a bottom-up order. Due to adopting different computational strategies, FSP and FR result in different $z^{(l)}_{ijc}$.

From another point of view, S can be expanded from each feedback layer. Suppose that we have N feedback layers in total, and we do expand S for N times at all the feedback layers. Then S can be reformulated as

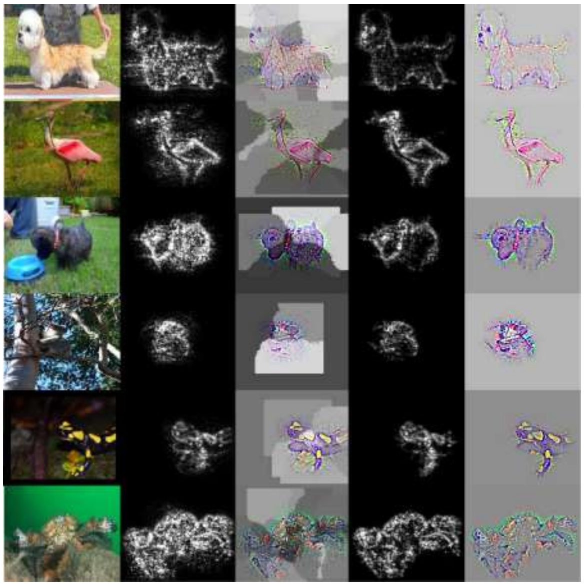
$$S = \frac{1}{N}\sum_{l=1}^{N}\sum_{ijc}\alpha^{(l)}_{ijc} z^{(l)}_{ijc} x^{(l)}_{ijc}. \tag{8}$$

If each $x^{(l)}_{ijc}$, $\forall\, c, l \in 1, 2, \ldots, N$ represents a particular pattern, then S is a linear combination of all those patterns from Equation (8). All the patterns with the negative CW will be removed by FSP. This will change the values of the reserved $x^{(l)}_{ijc}$'s, but not change the relationship between the reserved patterns and the target neuron. Actually, the FSP offers a natural way to seek the patterns closely related to a particular object.

Indeed, neither FR nor FSP provides a global optimum solution, and thus it is difficult to produce perfect visualization and energy maps only using one of them. In this paper, we propose Feedback CNN to combine the advantages of FR and FSP, and consequently impressive results can be produced.

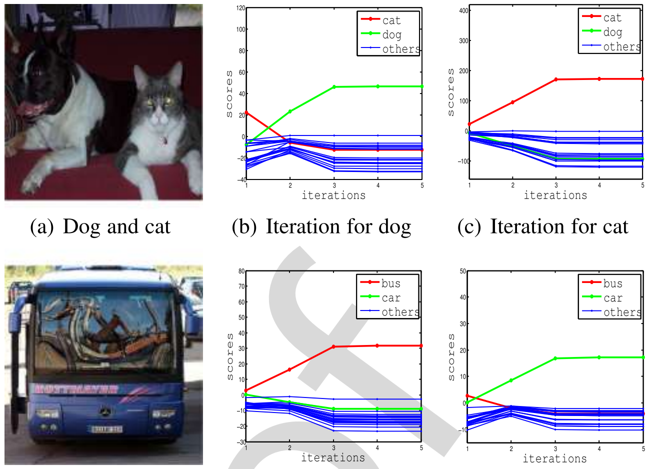
For both FR and FSP, the target neuron S is not limited to be a class node in the top layer. According to the optimization process, the target neuron S can be any hidden neuron in the neural network. In our proposed Feedback CNN, the FSP algorithm is used to select target-relevant neurons in every layer for a particular class node, and the FR algorithm is employed to reconstruct the target object by restoring the visual pattern information carried by the selected neurons. More specifically, the target object would be roughly reconstructed by: (1) running FR over the selected neurons in one of the middle layers simultaneously, and (2) performing a back-propagation from the target neurons (via setting the gradients as 1) to the image space.

Fig. 5 presents some examples to intuitively show the results generated by the proposed Feedback CNN. Specifically, Figs. 5b and 5c give the results of FR on the selected neurons separately, where the energy and the visualization maps are merged via the average operation. Figs. 5d and 5e give the results of FR on the selected neurons simultaneously, which is much more efficient in practice. It can be



(a) (b) (c) (d) (e)

Fig. 5. Results generated by combining both FR and FSP. (a) Input images. (b)(c) Merged energy maps and visualization maps by running FR separately on neurons selected by FSP. (d)(e) Energy and visualization maps by running FR simultaneously on neurons selected by FSP. Best viewed in color.



(a) Dog and cat (b) Iteration for dog (c) Iteration for cat

(d) Bus and cars (e) Iteration for bus (f) Iteration for cars

Fig. 6. The iteration curves of FSP for different objects. (a)(d) Input images. (b)(c) The iteration curves for dog and cat respectively. (e)(f) The iteration curves for bus and car respectively.

seen that the target objects can be effectively captured by Feedback CNN even for the images containing cluttered background. Thus the neurons associated with the target objects can be selected while the irrelevant ones can be turned off. Note that this kind of selectivity occurs in each hidden layer. In particular, for a deep CNN, we determine the status of gate-variables according to the mean value of all CWs in a layer, which would be more robust to noisy patterns.

4 Experimental Results

In this section, we conduct extensive experiments to verify the effectiveness of Feedback CNN. The iteration process of FSP is analyzed in Section 4.1 and the effectiveness of neuron selection is studied in Section 4.2. We evaluate the discriminative ability of FSP in Section 4.3. Besides, we conduct quantitative experiments of weakly supervised object localization in Section 4.4 and weakly supervised semantic segmentation in Section 4.5. It should be noted that since the FR algorithm is like a kind of image reconstruction, we evaluate FR together with FSP in Sections 4.4 and 4.5.

4.1 Analysis on Iteration Process of FSP

In order to verify our theoretical analysis described in Section 3 that the score of the target neuron would keep increasing until convergence when running the FSP algorithm, we specially visualize the iterative process of the FSP algorithm here. For the experimental purpose, the VggNet (16 layers) [4], which is pre-trained with ImageNet 2012 training set, is fine-tuned on the Pascal VOC2012 data set.

First, as shown in Figs. 6a, 6b and 6c, given the input image, the FSP algorithm is applied respectively on two neurons which represent the categories of "dog" and "cat" in the last fully connected layer named as "fc8" in the VggNet. The scores of all the 20 neurons in "fc8", corresponding to the 20 classes of Pascal VOC2012, are recorded during the iteration procedure. The iteration process for category "cat" is plotted with the red curve, "dog" with the green curve, and other 18 classes with the blue curves. As can be seen, all the iterative procedures converge after about 5 iterations. And the scores of the target neuron keep increasing until convergence, while the scores of other classes are suppressed even if the corresponding objects are presented in the image. Similar results are derived from the image which contains a bus and several cars, as shown in Figs. 6d, 6e and 6f. These results prove that FSP will converge to a local optimum efficiently and increase the score of the target neuron effectively. In addition, it should be noted that there are several small cars in the top left and right corners in Fig. 6d. When feedback is applied with respect to category "car", the scores of the target neuron keep increasing while the scores for "bus" decrease heavily though there is a big bus in the center of the image, as demonstrated in Fig. 6f. The reason is that neurons carrying useful information for particular targets can be selected effectively while irrelevant neurons will be turned off in the feedback loops.

Furthermore, we apply FSP on the ImageNet 2012 classification validation set which contains 50,000 images. The ground-truth label of each image is set as the target for the feedback model, and the scores of 5 iterations for all images are recorded. Then we calculate their mean and standard deviation of each iteration, and plot them in Fig. 7. We find that the FSP algorithm is also effective even for a very large image data set.

4.2 Effectiveness of Neuron Selection

In this section, we evaluate the effectiveness of neuron selection of FSP. Given an image with multiple class objects, e.g, images in Fig. 8, we run the FSP algorithm with the same VggNet in Section 4.1 for different targets, and take a middle layer named as "conv5_2" for explanation. This layer has 512 filter kernels, indicating that it may express 512

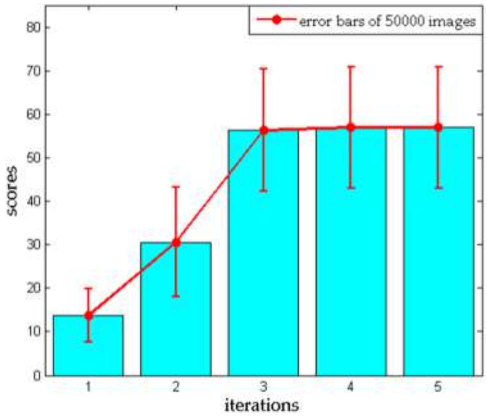


Fig. 7. The mean iteration curve of 50000 images from the ImageNet 2012 classification validation set.

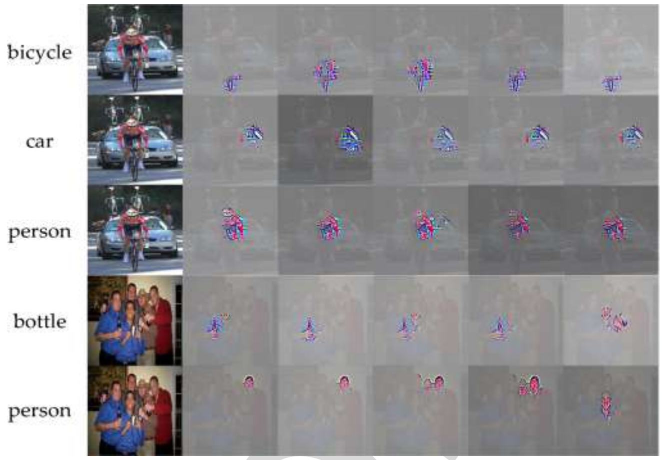
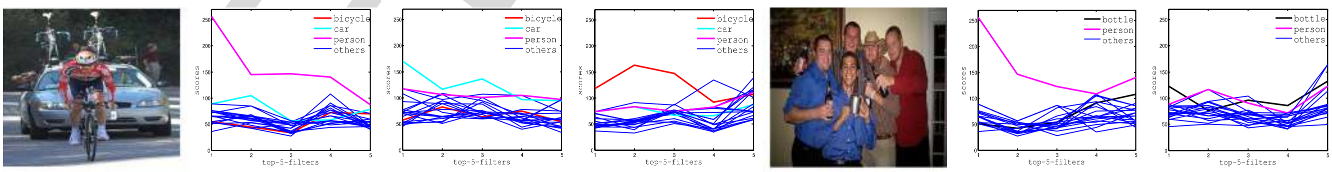


Fig. 8. Visualization of 5 neurons that have the maximum scores in each of the top 5 channels selected by the FSP algorithm. Best viewed in color.

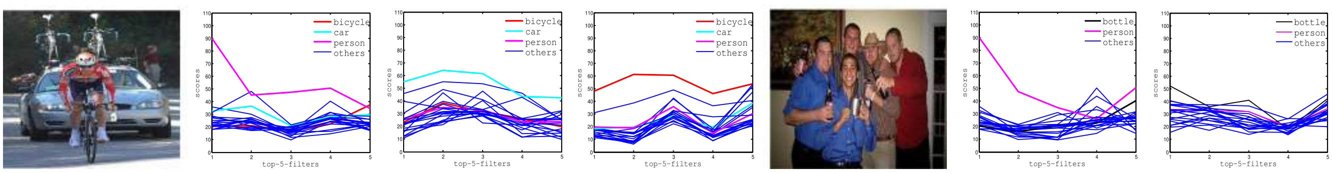
patterns related to different classes. In Fig. 8, the first input image contains 2 people, a bicycle and a car, and the FSP algorithm is run for the three classes. After convergence, for each target, we can select the top 5 channels by ranking the 512 feature maps according to their maximum scores. Further, we select the 5 neurons that have the maximum activation scores in each of the top 5 channels. As illustrated in Fig. 8, the FR algorithm is run to visualize these 5 neurons, and it turns out that they represent the most discriminative parts of the corresponding objects. The similar results appear in another image with some people and bottles.

To be more convincing, we evaluate these selected top 5 filters on the whole Pascal VOC2012 segmentation validation set which contains 1,449 images. We calculate the maximum and mean responses of the selected top 5 filters related to each of the 20 categories by using the images of each category. Since many images have multiple class labels, the calculation process using some of these images will be slightly adjusted. Suppose that we have correctly selected 5 filters for category A and a given input image is labeled as category A and B. Then the responses of these 5 filters will be mainly caused by the objects from category A, thus it is not reasonable to put these responses to category B.To avoid this mutual influence, we ignore these images when evaluating the performance on category B.

In Fig. 9, the maximum and mean responses are presented in the first and second rows respectively. We take the category "person" as an example for detailed analysis. FSP is run for the category "person" on the person-bicycle-car image until convergence, and top 5 filters are acquired. All images that are labeled as "person" are fed to the original CNN model. The responses are drawn with the magenta lines. Then, images of other 19 classes which do not contain any "person" are fed to the same CNN to get the corresponding responses. Specially, the responses for the category "bicycle" and "car" which appear in the image are plotted with the red and cyan lines respectively, and the rest 17 classes are plotted with blue lines. In Fig. 9b, the fact that the magenta lines is higher than other lines indicates that the corresponding filters are highly related to its target category, which means that the FSP algorithm has effectively selected the meaningful filters. The results are similar for another image, as shown in Fig. 9e. We find that this kind of neuron selection happens in all hidden layers. The FSP algorithm has the ability to correctly select the corresponding neurons (filters) to preset targets, as well as suppress irrelevant neurons at the same time.



(a) input image (b) person(max) (c) car(max) (d) bicycle(max) (e) input image (f) person(max) (g) bottle(max)



(h) input image (i) person(mean) (j) car(mean) (k) bicycle(mean) (l) input image (m) person(mean) (n) bottle(mean)

Fig. 9. Filter selection. The FSP algorithm is run for different objects in the input images (a) and (e). After FSP achieves convergence, we select top 5 channels (corresponding to 5 filters) for each target object according to the maximum scores of 512 channels in the "conv5_2" layer. We calculate the maximum and mean responses of these 5 filters to the images of 20 different classes from the Pascal VOC2012 segmentation validation set. The first row reports the maximum scores and the second row reports the mean scores. The filters selected by FSP well respond to the corresponding class images. For example, the selected filters for "person" have much higher responses to the images from the category "person". Best viewed in color.

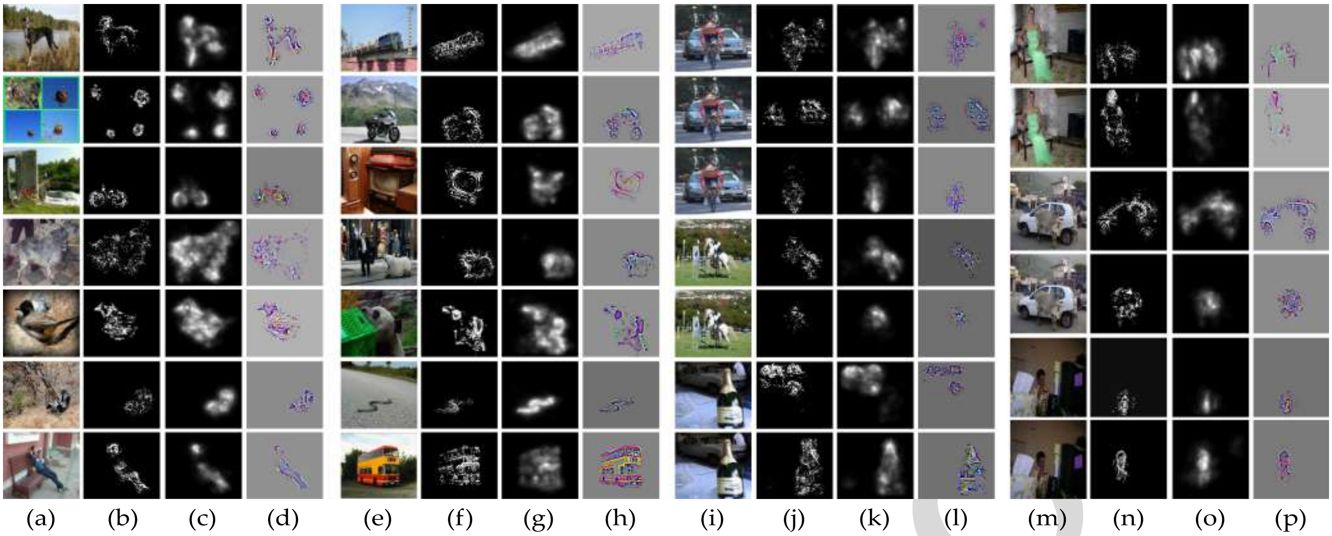


Fig. 10. Visualization and energy maps. (a)(e)(i)(m) Input images. (b)(f)(j)(n) FSP-FR energy maps. (c)(g)(k)(o) Summation Energy Maps. (d)(h)(l)(p) Visualization maps. Note that (i-p) demonstrate some results when input images contain multi-class objects. Best viewed in color.

4.3 Analysis on the Discriminative Ability of FSP

To evaluate this discriminative ability of FSP, we conduct several experiments on the Pascal VOC2012 classification validation set. The same VggNet in Section 4.1 is employed, and Feedback CNN is utilized to generate category-specific energy maps.

As a result, Fig. 10 depicts several examples. The energy maps generated by Feedback CNN are highly relevant to the target objects in the input images, as shown in Figs. 10b, 10f, 10j and 10n. For convenience, these energy maps are named as *FSP-FR energy maps*.

Summation Energy Map. Due to the selection ability of the FSP algorithm, most of the neurons preserved in all hidden layers are highly relevant to the same semantic class. Meanwhile, a whole object can be divided into several parts and be expressed in several different hidden layers. Thus, it is reasonable to combine all the selected neurons to generate a new energy map. We achieve this simply by the summation operation. After applying the FSP algorithm, we resize the gradients of feature maps in all ReLU layers behind convolutional layers with the same size of the input image, and calculate the summation of all the resized gradient maps along the channel direction. The summation map is normalized by ℓ_2 normalization, named as *Summation Energy Map*. Note that the energy value of each pixel indicates how important this pixel is to the target category and the total energy of a Summation Energy Map is 1. Figs. 10c, 10g, 10k and 10o illustrate some results. As can be seen, the Summation Energy Maps have better distributions over target objects.

The Summation Energy Map integrates information of the selected neurons in all hidden layers. So it is more convincing that we evaluate the discriminative ability of FSP using Summation Energy Maps instead of FSP-FR energy maps. We calculate the Summation Energy Maps for each image of the Pascal VOC2012 segmentation validation set. As the data set provides ground-truth masks for all objects of each category, we calculate the sum of energy that falls into the target object regions in each image. We call this value as the *coverage rate*. The mean coverage rate is computed for each class on all validation images and drawn in Fig. 11 with the blue curve. Specially, since the deconvolutional operation in back-propagation causes dilation of the objects in the energy maps, we further report the results of dilating the ground truth masks by 5 pixels and 10 pixels in Fig. 11, with green and red curves respectively. As a contrast, the mean coverage rate of energy maps based on the original gradients of the input image is reported too. As can be seen, all coverage rates of Summation Energy Maps (left) are much higher. That is, the Summation Energy Maps generated by FSP effectively highlight the expected objects and almost focus on the target areas.

To provide a more convincing evaluation of the discriminative ability of Summation Energy Map, we also calculate the coverage rate only for images with multi-class labels in PASCAL VOC2012 segmentation validation set. The corresponding results are shown in Fig. 12, demonstrating the effectiveness of FSP. All these results indicate that the FSP algorithm has a strong discriminative ability. Object-related neurons can be correctly selected and class-specific energy maps can also be effectively produced, which well paves the road for weakly supervised object localization and weakly-supervised semantic segmentation.

4.4 Weakly-Supervised Object Localization

4.4.1 The ImageNet 2012 Localization Task

In this section, we evaluate the object localization power of Feedback CNN on the ImageNet 2012 localization task. The

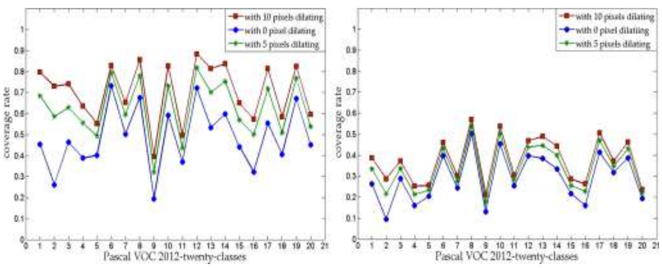


Fig. 11. Coverage rates of Summation Energy Map over all 20 classes images of Pascal VOC2012. (left) Coverage rates of Summation Energy Maps. (right) Coverage rates of the energy maps generated by original gradients. Best viewed in color.

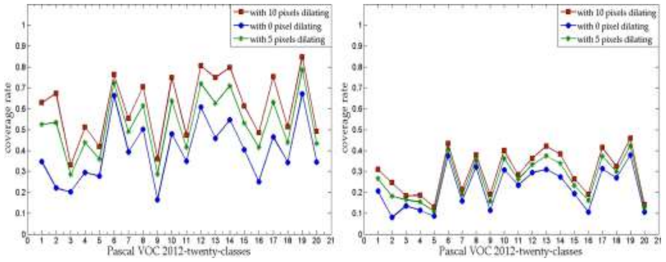


Fig. 12. Coverage rates of Summation Energy Map of multi-class images of Pascal VOC2012. (left) Coverage rates of Summation Energy Maps. (right) Coverage rates of the energy maps generated by original gradients. Best viewed in color.

goal of the task is bounding box localization of dominant objects in an image. The top 5 localization evaluation metric [25] is employed, in which a correct prediction is counted when one of the top 5 guesses meets the requirement that both object category prediction and its associated bounding box are correct. To generate the top 5 category predictions, the VggNet is trained on the ImageNet 2012 classification training set. For weakly supervised localization, several steps are performed to get a bounding box. We summarize the experimental procedure in Procedure 1.

Procedure 1. Weakly Supervised Object Localization

1: Given a test image and a predicted label;
2: Run FSP according to the predicted label and obtain Summation Energy Map;
3: Get a bounding box which preserves 99 percent energy of the Summation Energy Map. Crop the box region from the original image as new input;
4: Apply FSP again on the new input image;
5: Set one of the middle-level layers (e.g., “conv5_2”) as the target layer for FR;
6: Set the preserved neurons in “conv5_2” as the target, and run FR on those neurons simultaneously to get a new energy map;
7: Get a bounding box which preserves 99 percent energy of the new energy map.

In particular, as described in Procedure 1, objects are localized in two stages because of the varied scale of objects. We first localize the objects roughly in Steps 1–4. Then the precise localization is obtained in Steps 5–9 by combining FSP and FR. Note that, in Procedure 1, we intuitively select the “conv5_2” layer since the neurons in this layer represent large-size patterns which are beneficial for reconstructing target objects. And to preserve as much energy as possible but avoid meaningless solutions (e.g., a box with the same size of the input image), we intuitively set the rate of the preserved energy as 99 percent. Better performance could be produced when carefully select those hyper-parameters.

We compare the localization performance of Feedback CNN on the ILSVRC2012 validation set (50,000 images) with several state-of-the-art methods in Table 1. Compared with the VGGnet-GAP [29], our method wins 5.01 percent in terms of the accuracy of weakly supervised object localization. To avoid the influence of different classification performance of the compared models, we employ different image cropping strategies, such as no cropping, 5 cropping, and dense cropping [4], to produce different classification

TABLE 1
Localization Results on ILSVRC2012

methods	classification top 5 error	localization top 5 error
deepinside [25]	-	44.6
VGGnet-GAP [29]	12.2	45.14
Backprop-on-VGGnet [29]	11.4	51.46
GoogLeNet-GAP [29]	13.2	43.00
GoogLeNet [29]	11.3	49.34
Feedback CNN-no crop	**15.68**	**42.82**
Feedback CNN-5 crop	**12.95**	**41.72**
Feedback CNN-dense crop	**9.22**	**40.32**
MWP [30]	with GT	38.70
Feedback CNN	with GT	**36.50**

accuracies and we compare the localization accuracy when classification error rates are close. It is important to note that, the classification accuracies of the compared methods are all produced by using 5 cropping operation. As illustrated in Table 1, When our classification accuracy is 3.48 percent (without cropping operation) and 0.75 percent (with 5 cropping operations) lower than the compared approach VGGnet-GAP [29], we still achieve 2.32 and 3.42 percent higher localization accuracy, respectively. Moreover, when given ground truth labels, 36.50 percent error rate is obtained, which is an accuracy of 2.2 percent higher than the recent best-performing approach MWP [30] under the same experimental set-up.

Due to the powerful selection ability of FSP and the better object boundaries in energy maps, the proposed Feedback CNN outperforms the state-of-the-art approaches. The energy maps generated by our Feedback CNN are more precise and contain more complete objects. Accordingly, the bounding boxes are more close to the ground truth bounding boxes. Fig. 13 displays some examples.

4.4.2 Object Localization on Pascal VOC

We now turn to a different evaluation setting. We follow the evaluation protocol of weakly supervised object localization in [27], [30], [40]. Experiments are conducted on the test set of Pascal VOC2007 with 4,952 images and Pascal VOC2012 classification validation set with 5,823 images. Here, we compare Feedback CNN with the following methods: Excitation Backprop (EB) [30]; Exemplar-Driven(ED) [40]; Deep inside CNN (DICNN) [25]; Deconvolutional neural networks (Deconv) [10], and Weakly-supervised learning CNN (WeakSup) [27].

For Feedback CNN, the Summation Energy Maps are used as localization score maps. We extract the maximum

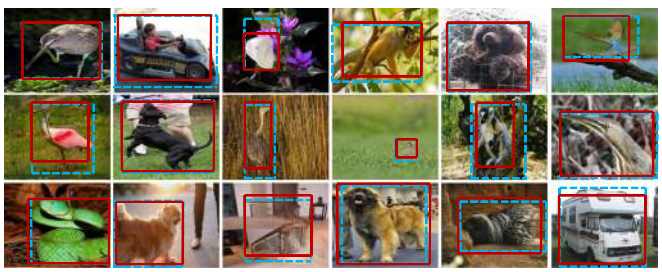


Fig. 13. Examples of weakly supervised object localization of our approach. The predicted bounding boxes are plotted in red color.

TABLE 2
Mean Accuracy (%) of Object Localization on the Test Set of VOC2007 and Classification Validation Set of VOC2012

	DICNN	WeakSup	Deconv	ED	EB	Ours
VOC07	76.0	-	75.5	-	80.0	**86.5**
VOC12	-	65.3	64.7	73.4	78.6	**82.7**

point on a Summation Energy Map as a location prediction. A hit is counted if the maximum point falls into the ground truth bounding box of the target category, otherwise a miss is counted. Unlike [27], which sets a 18-pixel tolerance to the predicted location, we restrict the correct predicted location to be within the ground truth bounding box for a more accurate evaluation. The localization accuracy is measured by $Acc = \frac{Hits}{Hits+Misses}$ for each category. The mean accuracy across all categories is reported. Table 2 presents the experimental results.

Note that we use the ground truth category labels as targets for Feedback CNN on Pascal VOC2007 for fair comparing with DICNN [25] and EB [30], and use the predicted ones on Pascal VOC2012 for fair comparing with ED [40] and WeakSup [27]. The results demonstrate that Feedback CNN significantly outperforms the compared methods with a large performance gap.

A visual comparison between Deconv [10], WeakSup [27], EB [30] and Feedback CNN is shown on the left side of Fig. 14. All the three input images contain a motorbike, and we present the localization maps for the motorbike class produced by the above four methods. More examples are presented on the right side of Fig. 14, in which all the input images contain objects from two categories of PASCAL VOC. As can be seen, our Feedback CNN generates more accurate localization maps with less noise. Both the qualitative and quantitative experiments support that Feedback CNN performs very well for the weakly-supervised object localization.

4.5 Weakly-Supervised Semantic Segmentation

In this section, we focus on the weakly-supervised semantic segmentation task with experimental analysis on the Pascal VOC2012 semantic segmentation challenge. We employ the standard Pascal VOC2012 segmentation metric: mean intersection-over-union (mIoU). Note that we only make use of class-level labels to fine-tune VggNet for classification on the Pascal VOC2012 segmentation training set, and evaluate our method on the Pascal VOC2012 semantic segmentation validation set (containing 1,449 images). In the training phase, the input images are randomly cropped, mirrored, scaled and rotated to obtain a better model. As for the multi-label classification task, the loss function we adopt is the sigmoid cross entropy instead of soft-max. To segment objects from an input image based on the energy map, the saliency cut proposed in [41] is utilized.

Procedure 2 demonstrates the experimental procedure. In particular, distinct parts of an object may be expressed in different layers, and their information can be all integrated into the Summation Energy Maps, which makes the Summation Energy Maps suitable for the segmentation task. On the other hand, the FSP-FR energy maps have the property to highlight object boundaries. Thus, we acquire

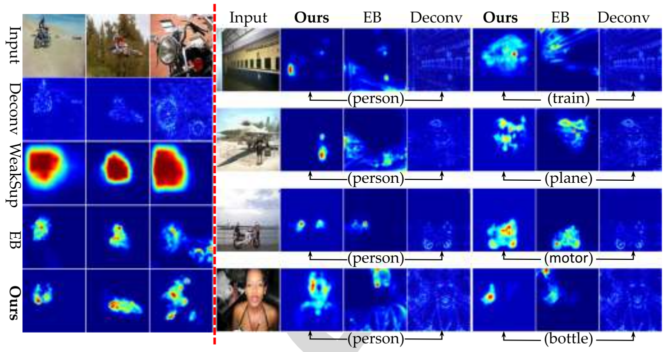


Fig. 14. Visual comparison. The left side is the comparison of localization score maps of the *motorbike* class between Deconv [10], WeakSup [27], EB [30] and Feedback CNN. The right side shows more examples.

both these energy maps for the target objects in an input image and simply add them together as the final energy map, which is called as the Summation-FSP-FR energy map. Specially, for the overlapped objects, the pixels of the overlapped regions are simply determined by their energy values in the corresponding Summation-FSP-FR energy maps. It should be noted that the deconvolutional operation in a CNN model in the back-propagation process will cause the offset in the energy map, which leads to the dilation around object edges. Thus, we regularize the Summation-FSP-FR energy into the super-pixels generated by the method proposed in [42].

Procedure 2. Weakly Supervised Semantic Segmentation

1: Given a test image and a predicted label;
2: Run FSP and obtain Summation Energy Map;
3: Select one of the middle-level layers (e.g., “conv5_2”) as the target layer for FR, and get the FSP-FR energy map;
4: Add Summation Energy Map and FSP-FR energy map to obtain Summation-FSP-FR energy map;
5: use super-pixels [42] to refine the Summation-FSP-FR map.
6: Run the saliency cut to get the segmentation results.

The quantitative results on the over-all validation set are listed in Table 3. We compare the performance of our weakly supervised approach with several state-of-the-art approaches with the same experimental setup, i.e., using only images from Pascal VOC2012 and only image-level labels. The results reveal that our approach largely outperforms previous techniques. Particularly, we achieve a 10.76 percent higher mIOU score than the state-of-the-art approaches and update the best records of 16 classes of Pascal VOC2012. Fig. 15 illustrates some successful examples, where we find that even for very complex scenes, the proposed approach still works well. We also show some failure cases and their corresponding objects' energy maps in Fig. 16. We observe that the energy maps are quite meaningful but the segmentation results are not satisfactory. The reason derives from the saliency cut [41], which implies that our approach has the potential to be further improved.

4.6 Discussion

The proposed Feedback CNN achieves good performance on both weakly supervised object localization and semantic segmentation. It is intuitive that, if all the target-relevant neurons in a classification neural network can be ideally

TABLE 3
Results of Weakly Supervised Semantic Segmentation on the Pascal VOC2012 Validation Dataset

	bkg	plane	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	motor	person	plant	sheep	sofa	train	tv	mAP
img+obj [31]																						32.2
stage1 [32]	71.7	30.7	**30.5**	26.3	20.0	24.2	39.2	33.7	50.2	17.1	29.7	22.5	41.3	35.7	43.0	36.0	29.0	34.9	23.1	33.2	33.2	33.6
EM-Adapt [35]	67.2	29.2	17.6	28.6	22.2	29.6	47.0	44.0	44.2	14.6	35.1	**24.9**	41.0	34.8	41.6	32.1	24.8	37.4	24.0	38.1	31.6	33.8
CCNN [33]	68.5	25.5	18.0	25.4	20.2	36.3	46.8	47.1	48.0	15.8	37.9	21.0	44.5	34.5	46.2	**40.7**	**30.4**	36.3	22.2	38.8	36.9	35.3
MIL+ILP+SP [34]	77.2	37.3	18.4	25.4	28.2	31.9	41.6	48.1	50.7	12.7	45.7	14.6	50.9	44.1	39.2	37.9	28.3	44.0	19.6	37.6	35.0	36.6
ours	**81.1**	**62.1**	25.9	**51.5**	**32.5**	**47.7**	**57.7**	**51.0**	**65.1**	**20.6**	**55.6**	23.7	**54.5**	**54.6**	**57.3**	38.5	27.2	**65.9**	31.2	**50.7**	**40.3**	**47.4**

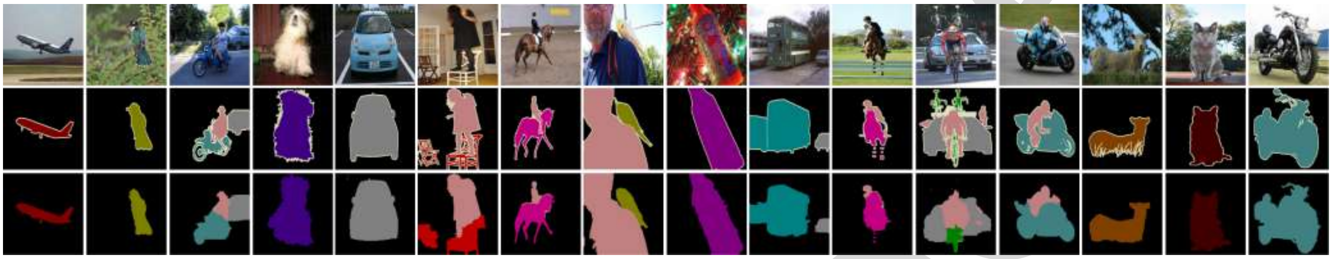
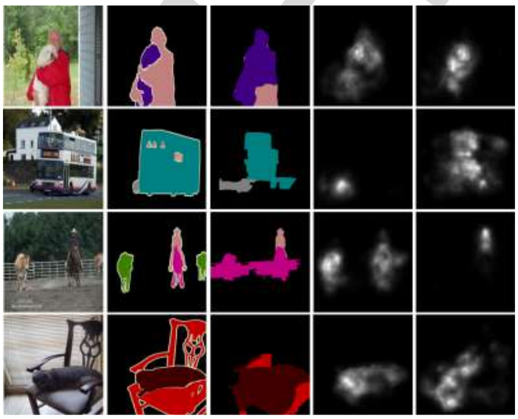


Fig. 15. Examples of weakly supervised semantic segmentation on the Pascal VOC2012 validation set. The first row is input images, the second row is ground truth segmentations, and the last row is our results.

selected when given an input image, the target objects can be accurately localized and even segmented from the input image based on the spatial and pattern information carried by all the target-relevant neurons.

In Feedback CNN, most semantic patterns can be well learned and expressed by the neurons in the basic classification CNN, which is the most fundamental premise. And the FSP algorithm can effectively select target-relevant neurons when given an input image, which is the essential part of the Feedback CNN. Due to the effectiveness of neuron selection, the spatial information carried by the selected neurons is able to be integrated into an energy map, i.e., the Summation Energy Map. And the pattern information can be restored and visualized by using the FR algorithm, which enables us to reconstruct target objects. Based on these advantages, we can obtain highly discriminative target-relevant energy maps with good quality (e.g., complete objects with clear boundaries). These are main reasons that the Feedback CNN works well on object localization and segmentation.



(a) (b) (c) (d) (e)

Fig. 16. Failed examples. (a) Input images. (b) Ground truth segmentations. (c) Our segmentation results. (d)(e) Energy maps for different objects generated by our approach.

Another issue to be discussed is what happens if an irrelevant class neuron is chosen as the target for an input image with objects of specific classes. In fact, given an input image, almost all the class nodes in common CNN models produce non-zero responses, which means that most class nodes can always find visual patterns from the input image that contribute to themselves. Therefore, when we set up an irrelevant target for the input image, the feedback model will always find the corresponding neurons that contribute to the target. More precisely, the proposed feedback mechanism is able to infer what makes the CNN model produce a specific prediction, no matter the prediction is right or wrong, weak or strong.

5 Conclusion

In this paper, we proposed a novel Feedback CNN consisting of the pruning and recovering operations. Feedback CNN gives an effective approach to implement the selectivity mechanism of neuron activation by jointly inferring the outputs of class nodes and activations of neurons in hidden layers. Feedback CNN is able to capture high-level semantic concepts and transform it into the image space to generate the energy maps. By embedding the feedback mechanism, a CNN that is only used for general object classification can be enhanced to accurately localize and segment the interested objects in images. A large number of qualitative and quantitative experiments have verified the effectiveness of our Feedback CNN. The feedback mechanism is significantly important in both the human visual system and machine vision systems, and thus deserves more attention. In the future, we plan to further explore it, e.g., how neurons represent multiple object instances of the same category, which is critical for instance segmentation.

Acknowledgments

This work is jointly supported by National Key Research and Development Program of China (2016YFB1001000), and

National Natural Science Foundation of China (61721004, 61420106015). We also thank Xianming Liu for participating in our initial experiments.

REFERENCES

[1] R. Desimone and J. Duncan, "Neural mechanisms of selective visual attention," *Annu. Rev. Neuroscience*, vol. 18, no. 1, pp. 193–222, 1995.

[2] N. Kruger, P. Janssen, S. Kalkan, M. Lappe, A. Leonardis, J. Piater, A. J. Rodriguez-Sanchez , and L. Wiskott, "Deep hierarchies in the primate visual cortex: What can we learn for computer vision?" *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 8, pp. 1847–1871, Aug. 2013.

[3] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 1–9.

[4] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *Int. Conf. Learn. Representations*, 2015.

[5] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.

[6] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan, "Object detection with discriminatively trained part-based models," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 9, pp. 1627–1645, Sep. 2010.

[7] C. Cao, X. Liu, Y. Yang, Y. Yu, J. Wang, Z. Wang, Y. Huang, L. Wang, C. Huang, W. Xu et al., "Look and think twice: Capturing top-down visual attention with feedback convolutional neural networks," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 2956–2964.

[8] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Region-based convolutional networks for accurate object detection and segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 1, pp. 142–158, Jan. 2016.

[9] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 3431–3440.

[10] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 818–833.

[11] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Object detectors emerge in deep scene CNNs," *Int. Conf. Learn. Representations*, 2015.

[12] N. Srivastava, G. E. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *J. Mach. Learning Res.*, vol. 15, no. 1, pp. 1929–1958, 2014.

[13] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," *Int. Conf. Mach. Learn.*, pp. 448–456, 2015.

[14] S. Zagoruyko and N. Komodakis, "Wide residual networks," *British Mach. Vis. Conf.*, pp. 87.1–87-12, 2016.

[15] C. D. Gilbert and W. Li, "Top-down influences on visual processing," *Nature Rev. Neuroscience*, vol. 14, no. 5, pp. 350–363, 2013.

[16] P. Hu and D. Ramanan, "Bottom-up and top-down reasoning with hierarchical rectified Gaussians," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 5600–5609.

[17] A. Dosovitskiy and T. Brox, "Inverting convolutional networks with convolutional networks," *Comput. Res. Repository*, vol. abs/1506.02753, 2015.

[18] A. C. Gilbert, Y. Zhang, K. Lee, Y. Zhang, and H. Lee, "Towards understanding the invertibility of convolutional neural networks," in *Proc. 26th Int. Joint Conf. Artif. Intell.*, 2017, pp. 1703–1710.

[19] M. F. Stollenga, J. Masci, F. Gomez, and J. Schmidhuber, "Deep networks with internal selective attention through feedback connections," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2014, pp. 3545–3553.

[20] A. Shrivastava, R. Sukthankar, J. Malik, and A. Gupta, "Beyond skip connections: Top-down modulation for object detection," *Comput. Res. Repository*, vol. abs/1612.06851, 2016.

[21] A. R. Zamir, T.-L. Wu, L. Sun, W. Shen, J. Malik, and S. Savarese, "Feedback networks," *2017 IEEE Conf. Comput. Vis. Pattern Recognit.*, pp. 1808–1817, 2017.

[22] Q. Wang, J. Zhang, S. Song, and Z. Zhang, "Attentional neural network: Feature selection using cognitive feedback," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2014, pp. 2033–2041.

[23] V. Mnih, N. Heess, A. Graves et al., "Recurrent models of visual attention," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2014, pp. 2204–2212.

[24] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.

[25] K. Simonyan, A. Vedaldi, and A. Zisserman, "Deep inside convolutional networks: Visualising image classification models and saliency maps," *Workshop Int. Conf. Learn. Representations*, 2014.

[26] K. Sohn, G. Zhou, C. Lee, and H. Lee, "Learning and selecting features jointly with point-wise gated boltzmann machines," in *Proc. Int. Conf. Mach. Learning*, 2013, pp. 217–225.

[27] M. Oquab, L. Bottou, I. Laptev, and J. Sivic, "Is object localization for free?-weakly-supervised learning with convolutional neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 685–694.

[28] A. Bergamo, L. Bazzani, D. Anguelov, and L. Torresani, "Self-taught object localization with deep networks," *2016 IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, pp. 1–9, 2016.

[29] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," *2016 IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 2921–2929, 2016.

[30] J. Zhang, Z. Lin, J. Brandt, X. Shen, and S. Sclaroff, "Top-down neural attention by excitation backprop," in *Proc. Eur. Conf. Comput. Vision*, 2016, pp. 543–559.

[31] A. Bearman, O. Russakovsky, V. Ferrari, and L. Fei-Fei , "What's the point: Semantic segmentation with point supervision," *Eur. Conf. Comput. Vis.*, pp. 549–565, 2016.

[32] H.-E. Kim and S. Hwang, "Scale-invariant feature learning using deconvolutional neural networks for weakly-supervised semantic segmentation," *Comput. Res. Repository*, vol. abs/1602.04984, 2016.

[33] D. Pathak, P. Krahenbuhl, and T. Darrell, "Constrained convolutional neural networks for weakly supervised segmentation," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 1796–1804.

[34] P. O. Pinheiro and R. Collobert, "From image-level to pixel-level labeling with convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 1713–1721.

[35] G. Papandreou, L.-C. Chen, K. P. Murphy, and A. L. Yuille, "Weakly-and semi-supervised learning of a deep convolutional network for semantic image segmentation," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 1742–1750.

[36] R. M. Cichy, D. Pantazis, and A. Oliva, "Resolving human object recognition in space and time," *Nature Neuroscience*, vol. 17, no. 3, pp. 455–462, 2014.

[37] M. Figurnov, A. Ibraimova, D. P. Vetrov, and P. Kohli, "Perforatedcnns: Acceleration through elimination of redundant convolutions," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2016, pp. 947–955.

[38] P. Molchanov, S. Tyree, T. Karras, T. Aila, and J. Kautz, "Pruning convolutional neural networks for resource efficient transfer learning," *Int. Conf. Learn. Representations*, 2017.

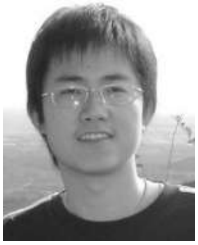
[39] G. Montavon, S. Lapuschkin, A. Binder, W. Samek, and K.-R. Müller, "Explaining nonlinear classification decisions with deep taylor decomposition," *Pattern Recognit.*, vol. 65, pp. 211–222, 2017.

[40] S. He, R. W. Lau, and Q. Yang, "Exemplar-driven top-down saliency detection via deep association," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 5723–5732.

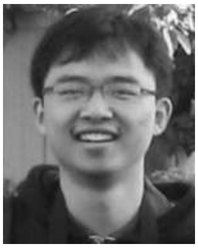
[41] M.-M. Cheng, N. J. Mitra, X. Huang, P. H. Torr, and S.-M. Hu, "Global contrast based salient region detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 3, pp. 569–582, Mar. 2015.

[42] P. Dollár and C. L. Zitnick, "Fast edge detection using structured forests," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 8, pp. 1558–1570, Aug. 2015.

Chunshui Cao received the BS degree from the University of Science and Technology of China, in 2013. He is currently working toward the PhD degree at the University of Science and Technology of China and studies in the Center for Research on Intelligent Perception and Computing (CRIPAC), National Laboratory of Pattern Recognition (NLPR), Institute of Automation, Chinese Academy of Sciences (CASIA). His research interests are in artificial intelligence, machine learning and computer vision.



Yongzhen Huang received the BE degree from the Huazhong University of Science and Technology, in 2006 and the PhD degree from the Institute of Automation, Chinese Academy of Sciences (CASIA), in 2011. In July 2011, he joined the National Laboratory of Pattern Recognition (NLPR), CASIA, where he is currently an associate professor. He has published more than 70 papers in the areas of computer vision and pattern recognition at international journals such as the *IEEE Transactions on Pattern Analysis and Machine Intelligence*, the *International Journal of Computer Vision*, the *IEEE Transactions on Image Processing*, the *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, the *IEEE Transactions on Circuits and Systems for Video Technology*, the *IEEE Transactions on Multimedia*, and conferences such as CVPR, ICCV, NIPS, AAAI. His current research interests include pattern recognition, computer vision and machine learning. He is a senior member of the IEEE.



Yi Yang received the BS degree with honors from Tsinghua University, in 2006 and master of philosophy degree in industrial engineering from the Hong Kong University of Science and Technology, in 2008, and the PhD degree in computer science from the UC Irvine, in 2013. He is currently a research scientist with the Institute of Deep Learning, Baidu Research. His research interests are in artificial intelligence, machine learning and computer vision. He is a member of the IEEE.

Liang Wang received the BEng and MEng degrees from Anhui University, in 1997 and 2000, respectively, and the PhD degree from the Institute of Automation, Chinese Academy of Sciences (CASIA), in 2004. From 2004 to 2010, he worked as a research assistant with Imperial College London, United Kingdom and Monash University, Australia, a research fellow with the University of Melbourne, Australia, and a lecturer with the University of Bath, United Kingdom, respectively. Currently, he is a full professor of hundred talents program with the National Lab of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, P. R. China. His major research interests include machine learning, pattern recognition and computer vision. He has widely published at highly-ranked international journals such as the *IEEE Transactions on Pattern Analysis and Machine Intelligence* and the *IEEE Transactions on Image Processing*, and leading international conferences such as CVPR, ICCV and ICDM. He has obtained several honors and awards such as the Special Prize of the Presidential Scholarship of Chinese Academy of Sciences. He is currently a senior member of the IEEE and a fellow of IAPR, as well as a member of BMVA. He is an associate editor of the *IEEE Transactions on Cybernetics* and the *IEEE Transactions on Information Forensics and Security*.

Zilei Wang received the BS and PhD degrees from the University of Science and Technology of China (USTC), in 2002 and 2007, respectively. He is currently an associate professor with the Department of Automation, USTC, and the founding lead of the Vision and Multimedia Research Group (http://vim.ustc.edu.cn). His research interests include computer vision, multimedia, and deep learning. He is a member of Youth Innovation Promotion Association, Chinese Academy of Sciences. He is a member of the IEEE.

Tieniu Tan received the BSc degree in electronic engineering from Xi'an Jiaotong University, China, in 1984, and the MSc and PhD degrees in electronic engineering from Imperial College London, United Kingdom, in 1986 and 1989, respectively. In October 1989, he joined the Computational Vision Group with the Department of Computer Science, The University of Reading, Reading, United Kingdom, where he worked as a research fellow, senior research fellow and lecturer. In January 1998, he returned to China to join the National Laboratory of Pattern Recognition (NLPR), Institute of Automation of the Chinese Academy of Sciences (CASIA), Beijing, China, where he is currently a professor and the director of Center for Research on Intelligent Perception and Computing (CRIPAC), and was former director (1998–2013) of the NLPR and director general of the Institute (2000–2007). He is currently also deputy director of Liaison Office of the Central Peoples Government in the Hong Kong S.A.R. He has published 14 edited books or monographs and more than 600 research papers in refereed international journals and conferences in the areas of image processing, computer vision and pattern recognition. His current research interests include biometrics, image and video understanding, and information content security. He is a fellow of the CASIA, TWAS (The World Academy of Sciences for the advancement of science in developing countries) and IAPR (the International Association of Pattern Recognition), and an international fellow of the UK Royal Academy of Engineering. He has served as chair or program committee member for many major national and international conferences. He is or has served as associate editor or member of editorial boards of many leading international journals including the *IEEE Transactions on Pattern Analysis and Machine Intelligence* (PAMI), the *IEEE Transactions on Automation Science and Engineering*, the *IEEE Transactions on Information Forensics and Security*, the *IEEE Transactions on Circuits and Systems for Video Technology*, the *Pattern Recognition*, the *Pattern Recognition Letters*, the *Image and Vision Computing*, etc. He is editor-in-chief of the *International Journal of Automation and Computing*. He was founding chair of the IAPR Technical Committee on Biometrics, the IAPR-IEEE International Conference on Biometrics, the IEEE International Workshop on Visual Surveillance and Asian Conference on Pattern Recognition (ACPR). He has served as the president of the IEEE Biometrics Council. He is currently the deputy president of Chinese Artificial Intelligence Association and president of China Society of Image and Graphics. He has given invited talks and keynotes at many universities and international conferences, and has received many national and international awards and recognitions. He is a fellow of the IEEE.