Look and Think Twice: Capturing Top-Down Visual Attention with Feedback Convolutional Neural Networks*

Chunshui Cao[†] Xianming Liu[‡] Yi Yang[§] Yinan Yu[§] Jiang Wang[§] Zilei Wang[†]
Yongzhen Huang[♮] Liang Wang [♮] Chang Huang[§] Wei Xu[§] Deva Ramanan[♯] Thomas S. Huang[‡]
[†]University of Science and Technology of China [‡]University of Illinois at Urbana-Champaign
[§]Baidu Research [♮]Institution of Automation, Chinese Academy of Sciences [♯]Carnegie Mellon University

Abstract

*While feedforward deep convolutional neural networks (CNNs) have been a great success in computer vision, it is important to note that the human visual cortex generally contains more feedback than feedforward connections. In this paper, we will briefly introduce the background of feedbacks in the human visual cortex, which motivates us to develop a computational feedback mechanism in deep neural networks. In addition to the feedforward inference in traditional neural networks, a feedback loop is introduced to infer the activation status of hidden layer neurons according to the "goal" of the network, e.g., high-level semantic labels. We analogize this mechanism as "**Look and Think Twice**." The feedback networks help better visualize and understand how deep neural networks work, and capture visual attention on expected objects, even in images with cluttered background and multiple objects. Experiments on ImageNet dataset demonstrate its effectiveness in solving tasks such as image classification and object localization.*

1. Introduction

> *"What did you see in this image?"*
> *"Panda, Tiger, Elephant, Lions."*
> *"Have you seen the Gorilla?"*
> *"Oh! I even didn't notice there is a Gorilla !"*

Visual attention typically is dominated by *"goals"* from our mind easily in a top-down manner, especially in the case of object detection. Cognitive science explains this in the "Biased Competition Theory" [1, 5, 6], that human visual cortex is enhanced by top-down stimuli and non-relevant neurons will be suppressed in feedback loops when searching for objects. By "*looking and thinking twice*", both human recognition and detection performances increase significantly, especially in images with cluttered background [3]. This leads to the selectivity in neuron activations [16], which reduces the chance of recognition being interfered with either noises or distractive patterns.

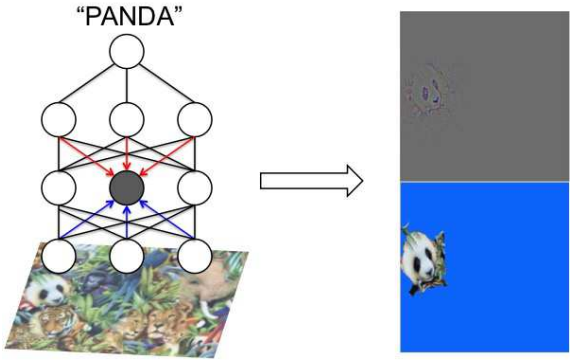
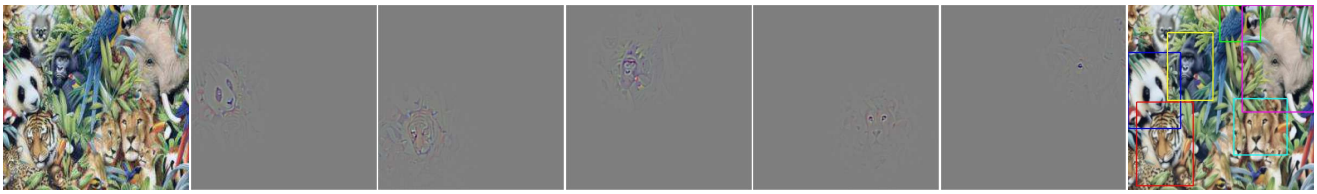


Figure 1. Feedback Convolutional Net model for capturing visual attention by inferring the status of hidden neuron activations. It is designed to utilize both bottom-up image inputs and top-down semantic labels to infer the hidden neuron activations. Salient areas captured by feedback often correspond to related "target" objects, even in images with cluttered background and multiple objects.

Inspired by above evidences, we present a novel *Feedback Convolutional Neural Network* architecture in this paper. It achieves this selectivity by jointly reasoning outputs of class nodes and activations of hidden layer neurons during the feedback loop. As shown in Figure 1, during the feedforward stage, the proposed networks perform inference from input images in a bottom-up manner as traditional Convolutional Networks; while in feedback loops, it sets up high-level semantic labels, (*e.g.*, outputs of class nodes) as the "goal" in visual search to infer the activation status of hidden layer neurons. We show that the network is powerful enough to apply for class model visualization [24, 33], object localization [24] and image classification [15].

Optimization in a Feedback Loop: From a machine learning perspective, the proposed feedback networks *add extra flexibility to Convolutional Networks, to help in capturing visual attention and improving feature detection.*

*This work is conducted during Chunshui Cao's and Xianming Liu's internship at Institute of Deep Learning, Baidu Research



(a) Input Image (b) Panda (c) Tiger (d) Gorilla (e) Lion (f) Elephant (g) Localization

Figure 2. We illustrate the localization power of the feedback net on a multi-object image with cluttered background. (a) shows the original input image which both VggNet [25] and GoogleNet [29] recognize as "comic book". (b) - (f) illustrate our feedback model on understanding the image given different class labels as a prior. We visualize the gradient of each class node with respect to image after the feedback net finish its inference. (g) shows the final localizations for different objects based on the gradients. Better viewed in color.

Convolutional Neural Networks [18, 15, 25] have achieved great success in both machine learning and computer vision in recent years. Benefiting from large-scale training data, (*e.g.,* ImageNet [4]), CNNs are capable of learning filters and image compositions at the same time. Various approaches have been adopted to further increase generalization ability of CNNs, by either adding regularization in training [11, 13], or going deeper [25, 29]. Inspired by Deformable Part-Based Models (DPMs) [8] that characterize middle level part locations as latent variables and search for them during object detection, we utilize a simple yet efficient method to optimize image compositions and assign neuron activations given "goals" in visual search. The algorithm maximizes the posterior response of network given target high-level semantic concepts, in a top-down manner. Compared with traditional bottom-up strategies [11, 13], which aim to regularize the network training, the proposed feedback framework adds flexibilities to the model inference from high-level concepts down to the receptive field.

Figure 2 shows an example on how this flexibility is reflected in detection and localization. Instead of recognizing the input image as a "comic book", the proposed feedback network is capable to localize each component of the "comic book" via salience maps. The example shown in Figure 1 illustrates its working mechanism: given a high-level semantic stimulus "PANDA", only the neurons in hidden layers related with the concept "PANDA" will be activated by iterative optimization in a feedback loop. As a result, only salient regions related with the concept "PANDA" are captured in visualizations. As suggested by those results, the feedback networks achieve certain level of selectivity and provide non-relevant suppression during the top-down inference, allowing the model to focus on the most salient image regions that improve the class confidence.

Weakly Supervised Object Localization: Given gradient visualizations in Figure 2, we further develop an algorithm for weakly supervised object localization. Instead of using large amount of supervision (*e.g.*, bounding box positions) in traditional methods such as R-CNN [9] or using regression model [7, 25], we don't require any localization information in the training stage. Instead, we adopt *a unified network performing both recognition and localization tasks*, to answer questions of "what" and "where" simultaneously, which are the two most important tasks in computer vision. Experimental results suggest that our weakly supervised algorithm using feedback network could achieve competative performance on ImageNet object localization task as GoogLeNet [29] and VGG [25].

Image Classification Revisited: We mimic the human visual recognition process that human may focus to recognize objects in a complicated image after a first time glimpse as the procedure "Look and Think Twice" for image classification. We utilize the weakly supervised object localization during the "first glimpse" to make guesses of ROIs, then make the network refocused on those ROIs and give final classifications list. Experimental results on ImageNet 2014 classification validation dataset suggest that this approach is efficient to eliminate irrelevant clutters and improve classification accuracy especially on small objects.

2. Related Work

2.1. Feedforward and Feedback Mechanism

Deep Neural Network takes a *feedforward-Back Error Propagation* strategy to learn features and classifiers simultaneously, from large scale of training samples [15, 25, 20, 23, 2]. Various approaches have been proposed to further improve the discriminative ability of deep neural network, either by 1) adding regularization to improve the robustness of learnt model and get rid of overfitting, such as Dropout [27], PReLU [11], Batch Normalization [13]; or 2) making the network deeper [29, 25].

Despite great successes achieved by applying Feedforward Networks to image recognition and detection, evidences accumulate from cognitive studies and point to the feedback mechanism that may dominant human perception processes [3, 22, 16, 19]. Recently, tentative efforts have been made to involve feedback strategy into Deep Neural Networks. Deep Boltzmman Machines (DBM) [23, 26] and Deconvolutional Nerual Networks [34] try to formulate

the feedback as a reconstruction process within the training stage. Meanwhile, Recurrent Neural Networks (RNN) and Long Short Term Memory (LSTM) [12] are utilized to capture the attention drifting in a dynamic environment and learn the feedback mechanisms via reinforcement learning [28, 21]. DRAW from Google DeepMind [10] combine above two into a generative model, to synthesize the image generation process.

As in *Biased Competition Theory* [1, 6], feedback, which passes the high-level semantic information down to the low-level perception, controls the selectivity of neuron activations in an extra loop in addition to the feedforward process. This results in the "Top-Down" attention in human cognition. Hierarchical probabilistic computational models [19] are proposed to characterize feedback stimuli in a top-down manner, which are further incorporated into deep neural networks, for example, modeling feedback as latent variables in DBM [31], or using selectivity to resolve fine-grained classification [21], *et al.*. However, generative models used in [21, 31] are limited by low computational efficiency and capacity, and are hardly used in large scale datasets.

2.2. Visualization, Detection, and Localization

Feedback is often related with visualization of CNN and object localization since both of them aim to project the high-level semantic information back to image representations. To visualize neuron responses and class models, various approaches are proposed either using deconvolution [33] or optimization based on gradients [24, 17]. As demonstrated in [24], visualization of Convolutional Neural Network shows semantically meaningful salient object regions and helps understand working mechanism of CNNs.

Object detection and localization are more about feedback, by treating detection / localization as a searching process with clear "goals." To localize and detect objects in images, typical approaches use supervised training, which relies on large amount of supervision, *e.g.*, ground-truth bounding boxes, or manually labeled segmentation in training samples [7]. To behave "searching", sliding window is used [7], or instead region proposals detected by image segmentations [30] in R-CNN [9]. However, both of these approaches are computational intensive and naturally bottom-up: selecting candidate regions, performing feedforward classification and making decisions.

Inspired by visualizations of CNNs [33, 24], a more feasible and cognitive manner for detection / localization could be derived by utilizing the saliency maps generated in feedback visualizations. Moreover, an ideal approach should unify the recognition and detection in a single feedforward-feedback network architecture. However, if possible, the challenge lies on how to obtain semantically meaningful salience maps with high quality for each concept. That's the ultimate goal of our work presented in this paper.

3. Model

3.1. Re-interpreting ReLU and Max-Pooling

The most recent state-of-the-art deep CNNs [25] 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 the output of a previous layer, consisting of C input channels of width M and height N: $\mathbf{x} \in \mathcal{R}^{M \times N \times C}$. The output $\mathbf{y}$ consists of a set of C' output channels of width M' and height N': $\mathbf{y} \in \mathcal{R}^{M' \times N' \times C'}$.

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

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

ReLU Layer: The ReLU layer is used to increase the nonlinear properties of the decision function and of the overall network without affecting the receptive fields of the convoluional layer.

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

Max-Pooling Layer: The max-pooling layer is used to reduce the dimensionality of the output and variance in deformable objects to ensure that the same result will be obtained even when image features have small translations. The max-pooling operation is applied for every pixel (i, j) around its small neighborhood $\mathcal{N}$.

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

Selectivity in Feedward Network: To better understand how selectivity works in neural networks and how to formulate the feedback, we re-interpret behaviors of ReLU and Max-Pooling layers as a set of binary activation variables $\mathbf{z} \in \{0, 1\}$ instead of the $\max()$ operation in Equation 2 and 3. Thus, behaviors of ReLU and Max-Pooling could be formulated as $\mathbf{y} = \mathbf{z} \circ \mathbf{x}$, where $\circ$ is the element wise product (Hadamard product); and $\mathbf{y} = \mathbf{z} * \mathbf{x}$, where $*$ is the convolution operator and $\mathbf{z}$ is a set of convolutional filters except that they are location variant.

Be interpreting ReLU and Max-Pooling layers as "gates" controlled by input x, the network selects information during feedforward phases in a *bottom-up* manner, and eliminates signals with minor contributions in making decisions. However, the activated neurons could be either helpful or harmful for classification, and involve too many noises, for instance, cluttered backgrounds in complex scenes.

3.2. Introducing the Feedback Layer

Since the model opens all gates and allow maximal information getting through to ensure the generalization, to

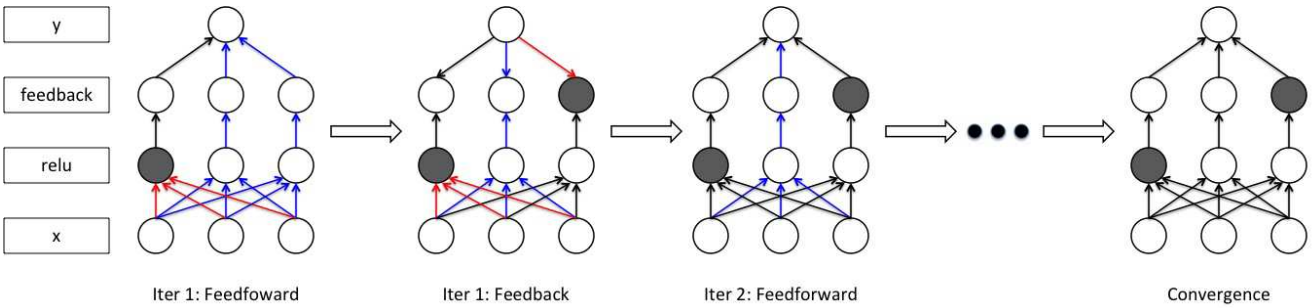


Figure 3. Illustration of our feedback model and its inference process. At the first iteration, the model performs as a feedforward neural net. Then, the neurons in the feedback hidden layers update their activation status to maximize the confidence output of the target top neuron. This process continues until convergence. (We show only one layer here, but feedback layers can be tacked in the deep ConvNet.)

increase the discriminability within feature level, it is feasible to turn off those gates that provide irrelevant information when targeting at particular semantic labels. This strategy is explained as selectivity in biased competition theory [6] and is critical to realize the top-down attention.

More technically, to increase the model flexibility to images and prior knowledges, we introduce an extra layer to the existing convolutional neural network. We call it the *feedback layer*. The feedback layer contains another set of binary neuron activation variables $\mathbf{z} \in \{0, 1\}$, in addition to ReLU. However, these binary variables are activated by top-down messages from outputs, instead of bottom inputs. The feedback layer is stacked upon each ReLU layer, and they compose a hybrid control unit to active neuron response in both bottom-up and top-down manners:

Bottom-Up Inherent the selectivity from *ReLU layers*, and the dominant features will be passed to upper layers;

Top-Down Controlled by *Feedback Layers*, which propagate the high-level semantics and global information back to image representations. Only those gates related with particular target neurons are activated.

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

3.3. Updating Hidden Neurons in Feedback Loops

We formulate the feedback mechanism as an optimization problem, by introducing an addition control gate-variable $\mathbf{z}$. Given an image I and a neural network with learned parameters w, we optimize the target neuron output by jointly inference on binary neuron activations $\mathbf{z}$ over all the hidden feedback layers. In particular, if the target neuron is a k-th class node in the top layer, we optimize the class score s_k by re-adjusting the neuron activations at every neuron (i, j) of channel c, on feedback layer l.

$$\begin{aligned} \max_{\mathbf{z}} \quad & s_k(I, \mathbf{z}) - \lambda||\mathbf{z}|| \\ s.t. \quad & z^l_{i,j,c} \in \{0, 1\},\ \forall\, l, i, j, c \end{aligned} \tag{4}$$

Since the goal of this optimization aims at activating minimal number of neurons yet maximizing the target score, we use $L1$ norm in above target function, as $\|\mathbf{z}\|_1$

This leads to an integer programming problem, which is NP-hard given the current deep net architecture. An approximation could be derived by applying a linear relaxation:

$$\begin{aligned} \max_{\mathbf{z}} \quad & s_k(I, \mathbf{z}) - \lambda||\mathbf{z}||_1 \\ s.t. \quad & 0 \leq z^l_{i,j,c} \leq 1,\ \forall\, l, i, j, c \end{aligned} \tag{5}$$

We use the gradient ascent algorithm to update the hidden variables through all layers simultaneously.

$$\mathbf{z}_{t+1} = \mathbf{z}_t + \alpha \cdot (\frac{\partial s_k}{\partial \mathbf{z}}|_{\mathbf{z}_t} - \lambda) \tag{6}$$

where $\frac{\partial \lambda \|\mathbf{z}\|_1}{\partial z_i} = \lambda$ since we require $0 \leq z^l_{i,j,c} \leq 1$.

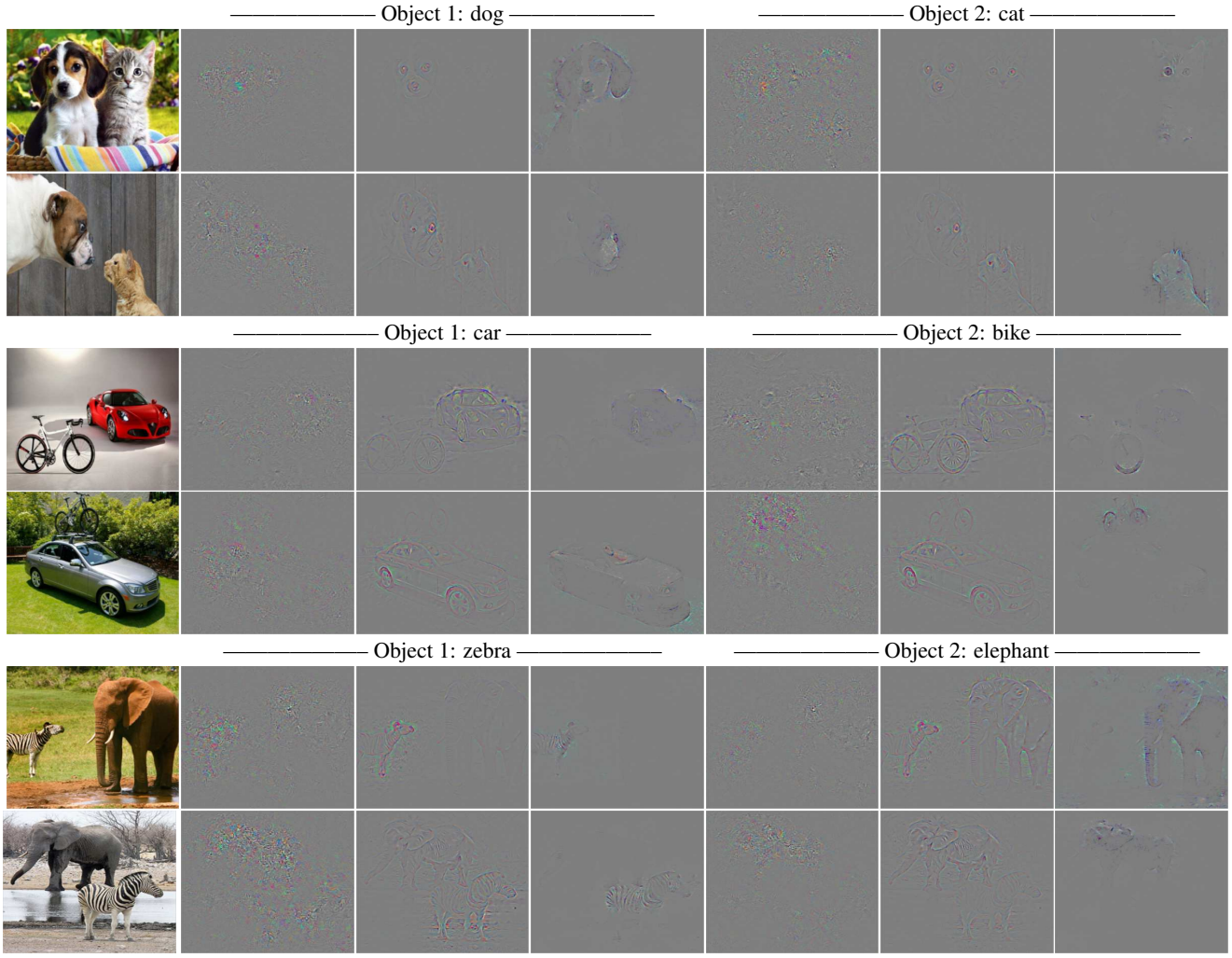
The initialization of feedback layer status z is set to be the corresponding ReLU activation after the first feedforward pass and truncate z when the updated values are either larger than 1 or smaller than 0 during inference.

3.4. Implementation Details

As for the implementation details, we set the feedback layer on top of each ReLU layer. We intialize all the hidden activations from $\mathbf{z} = \mathbf{1}$, making all feedback "gates" opening during the first time feedforward pass. It is suspected that fully connected layers learn embedding spaces rather than particular parts compared to convolutional layers. We update high level feedback layers according to the sign of gradient of each neuron. We set learning rate of hidden activations to 0.1 and update the neurons of all the feedback layers simultaneously. Each iteration performs a feedforward step of the neural net and a backpropagation step to send back gradients. We stop this process in 10 to 50 iterations.

4. Experimental Results

The *Feedback Network* could be used to improve various computer vision problems. In this paper, we demonstrate



(a) Image (b) Gradient (c) Deconv (d) Feedback (e) Gradient (f) Deconv (g) Feedback

Figure 4. We demonstrate the effectiveness of feedback neural networks for class-specific feature extraction, by comparing the class model visualization results against original gradient [24] and Deconv [33] on selected images with multiple objects. All methods compute visualizations using a pre-trained GoogleNet trained on ImageNet 2012 classification dataset. Column (a) shows the input images (*i.e.* dog v.s. cat, car v.s. bike, and zebra v.s. elephant). Column (b) and (e) show the original image gradients given the provided class labels. Column (c) and (f) show the Deconv results. Column (d) and (g) show the image gradients after feedback. Comparing against original gradient and Deconv, the feedback visualization captures more accurate salient area of the target object. For example, in the 4th row, both original template and Deconv see the dog and cat, even provided with the target label. In the last row, when zebra is specified, Deconv finds it hard to suppress the elephant area. Our feedback method suppress the irrelevant object much better. Better viewed in color and zoom in.

its potential, conduct qualitative experiments on class neuron visualizations, and quantitative experiments on weakly supervised object localization task. Furthermore, we show that the image recognition could also benefit from the *Feedback* mechanism, by taking the strategy "Looking and Thinking Twice", which eliminate noisy or cluttered background and makes the network focused on salient regions. We use three most popular pre-trained ConvNet models, AlexNet [15], VggNet [24] and GoogleNet [29] for experiments. All three models are pre-trained with ImageNet 2012 classification training dataset [4], obtained from Caffe [14] model zoo[1].

4.1. Image Specific Class Model Visualization

Given an image I, a class label k and the hidden neuron activation states $\mathbf{z}$, we approximate the neural net class score s_k with the first-order taylor expansion in the neighborhood of I:

$$s_k(I, \mathbf{z}) \approx \mathbf{T}_k(\mathbf{z})^T I + b \tag{7}$$

[1]https://github.com/BVLC/caffe/wiki/Model-Zoo

where $\mathbf{T}_k(\mathbf{z})$ is the derivative of s_k with respect to the image at the point of I and $\mathbf{z}$. $\mathbf{T}_k(\mathbf{z})$ can be viewed as the linear template applied on image I for measuring how likely the image belongs to class k, and could be visualized in the same spatial space since it is of the same size as input image I. We use this technique to visualize our feedback model throughout the paper.

More specifically, for a Convolutional Network composed with a stack of piecewise linear layers (*i.e.* Conv, ReLU and max-pooling) to compute the class scores, once the hidden states $\mathbf{z}$ are determined, the final score is a linear function of the image, which is equivalent to the inner product between the template and the image.

Comparison of Visualization Methods: We compare the image gradient (template $\mathbf{T}$) after the feedback process against the original one in feedforward pass, and Deconvolutional Neural Net [33] on a set of complex images containing multiple objects from different classes, with all using the same pre-trained GoogleNet and being given ground truth class labels as a prior. Qualitative results are shown in Figure 4. Without involving the feedback, where all hidden neurons' status are determined by the bottom-up computation only, the visualization is the same as original image gradient. However, compared with Deconv-like approaches, our feedback model is more efficient in capturing salient regions for each specific class while suppress those irrelevant object areas at the same time after feedback.

Comparison of ConvNet Models: We also qualitatively compare major convolutional network models, *i.e.*, AlexNet, VggNet and GoogleNet, by visualizing their feedback templates in Figure 5. All models are given ground truth labels a prior. From visualizations, we find that VggNet and GoogleNet produce more accurate visual attention than AlexNet, suggesting that using smaller convolution filters and deeper architectures could further distinguish similar and nearby objects. Moreover, although both VggNet and GoogleNet produce very similar image classification accuracies, GoogleNet better captures the salient object areas than VggNet. We hypothesize that the two $4,096$ dimensional fully connected layers (*i.e.*, fc6, fc7) in VggNet (which GoogleNet does not contain) could ruin the spatial distinctiveness of image features, as pointed out in [20].

4.2. Weakly Supervised Object Localization

To quantitatively demonstrate the effectiveness of the feedback model. we experiment on the ImageNet 2014 localization task. As pointed in [24], the magnitude of the elements in the model template $\mathbf{T}_k$ defines the class specific salience map on image I. Pixels with larger magnitudes indicate that they are more important to the class. We adopt the same saliency extraction strategy as [24] that a single class saliency value M_k for class k at pixel (i, j) is computed across all color channels: $M_k(i, j) = \max_{c \in rgb} |T_k(i, j, c)|$.

Method	Localization Error (%)
Oxford [24]	44.6
Feedback	**38.8**

Table 1. Comparison of our weakly supervised localization results on ImageNet 2014 validation set with the simplified testing protocol: the bounding box is predicted from a single central crop of images and the ground truth labels are provided. We show that our feedback method significant outperforms the baseline method (error rate 44.6%) that uses the original image gradient to localize in [24], both on GoogLeNet architecture.

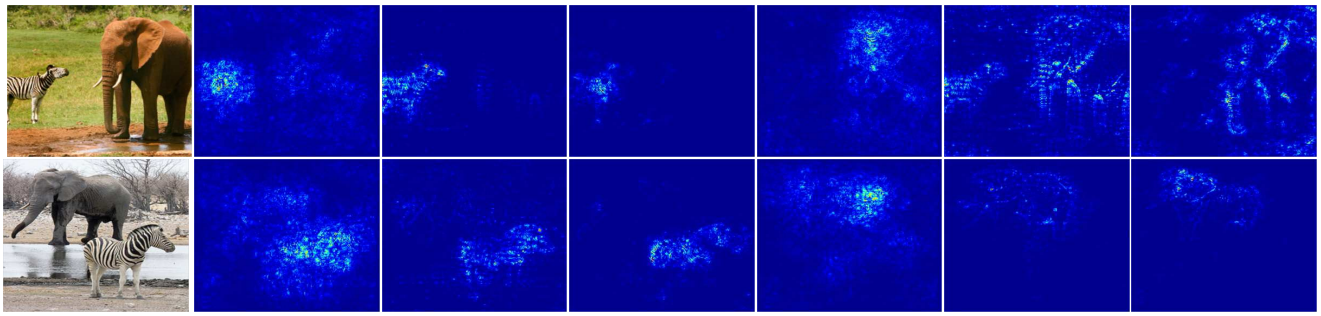
	Weakly Supervised	Supervised
Model	Localization Error (%)	Localization Error (%)
AlexNet [15]	49.6	-
VggNet [25]	40.2	34.3[25]
GooglNet [29]	**38.8**	-

Table 2. Column 2 compares localization errors using feedback on different ConvNet models. VGG and GoogleNet significant outperform AlexNet suggesting they are learning better features. GoogleNet outperforms VGG even further, which matches the observations in Figure 5. We also compare the weakly-supervised feedback mechanism with totally supervised localization model in [25] on VGG, in the third column. It shows that we are competitive to a carefully trained localization model (34.3%) using pixel-wise supervised training data.

We show that the proposed Feedback CNN has the potential to unify recognition and detection into a single network architecture in this experiment, instead of using separate ones to perform different tasks respectively. Although the three ConvNets are pre-trained for image classification, we could use the feedbacked salience map for weakly supervised object localization. Given an image and the corresponding class salience map, we compute the object segmentation mask by simply thresholding so that the foreground area covers 95% energy out of the whole salience map, and calculate a tightest bounding box as the localization result. Different from [24], which uses GraphCut [32], this requires saliency maps of higher quality, but only takes less computation.

We test our localization results on ImageNet 2014 validation set, which contains $\sim 50,000$ images with each image associated with labels and corresponding bounding boxes. A prediction is considered as correct if and only if its overlap with the ground truth bounding box is over 50%. Image are resized to 224x224 to meet the model requirement on resolutions, and ground-truth class labels are provided to predict localizations. Neither further preprocessing nor multi-scale strategy are involved.

Comparison of Localization Methods: Table 1 shows the comparison of our weakly supervised localization accuracy against the baseline method [24]. For fair comparison, we reimplemented the method in [24] following the details



(a) Image (b) AlexNet (c) VggNet (d) GoogleNet (e) AlexNet (f) VggNet (g) GoogleNet

Figure 5. We qualitatively compare the feedback ability of three most popular pre-trained ConvNets: AlexNet, VggNet and GoogleNet, by visualizing final image gradients and salience maps after feedback. We show the input images in column (a); results of these three models feedbacked by "zebra" are shown in column (b), (c), (d), and by "elephant" in column (e), (f), (g) respectively. We find that VggNet performs quite better than AlexNet, especially in capturing salient object details, suggesting the benefit of usage of small convolutional filters and deeper architecture. Although both VggNet and GogoleNet produce similar classification accuracy, GoogleNet provides the better class specific feature separations according to these results. We suspect the two 4096 fully connected layers in VggNet (which GoogleNet does not have) may harm the spatial distinctiveness of image features.

Method	Top 1 (%)	Top 5 (%)
GoogleNet [29]	32.28	11.75
GoogleNet Feedback	**30.49**	**10.46**

Table 3. Classification errors on ImageNet 2014 validation set with the simplified testing protocol: the first row is the performance of GoogleNet given a single central crop of images, the second row shows classification results of the same GoogleNet given the attention cropped images, using the feedback mechanism in 4.3.

in the original paper strictly, and name as "Oxford." For our method, we use GoogLeNet and apply the same segmentation strategy in our model. Our method obtains 38.8% localization error, and significantly outperforms Oxford (44.6%), suggesting that in terms of capturing attention and localizing salient objects, our feedback net is better. Note that our weakly supervised localization error is even closer to a carefully trained supervised localization model (34.3%).

Comparsion of ConvNet Models: We also analyze weakly supervised localization accuracies of above mentioned three ConvNets in Table 2, provided with the same testing protocol. Even provided with ground truth class, VggNet and GoogleNet significantly outperforms AlexNet. This suggests that better feature representations are sharable between the two highly correlated visual tasks: recognition and localization. GoogLeNet outperforms VggNet even further, which matches the observations in Figure 5.

4.3. Image Re-Classification with Attention

Given the weakly supervised attention boxes, the image labels are *re-classified* using zoomed-in image patches cropped around the bounding boxes. We call such method *"Look and Think Twice"*, which mimics the human visual recognition process that human may focus to recognize objects in a complicated image after a first time glimpse. We apply this strategy to the image classification. By looking at the full image first in a coarse scale, our model obtains an initial guess of a set of most probable object classes, we then identify the salient object regions from the predicted top-ranked labels using the feedback neural nets, and reclassify those regions. Below are the implementation details on ImageNet 2014 classification:

- Resize image to size $224 * 224$, run CNN model and predict top 5 class labels.
- *For each* of the top 5 class labels, compute object localization box with feedback model.
- Crop image patch for each of 5 bounding boxes from original image and resize to $224 * 224$. Predict top 5 labels again.
- Given the total 25 labels and the corresponding confidences, rank them and pick the top 5 as final solution.

Note that when cropping image patches, we use the original image with full resolution (*i.e.* $375 * 500$). This gives the cropped patch enough resolution for recognizing objects in details. Figure 7 shows two exemplar images from ImageNet 2014 validation dataset. Both images are incorrectly classified as *"Italian greyhound"* and *"sunglass"* respectively. After feedback, the network determines potential salient ROIs, "reconsider" these salient regions as input and make correct classifications, e.g., *"Ibizan hound"* and *"mask"* with high confidence scores. These two examples thoroughly illustrate how "looking and thinking twice" improves classification tasks.

Classification Accuracy: We test our classification results on ImageNet 2014 validation set, which contains $\sim$

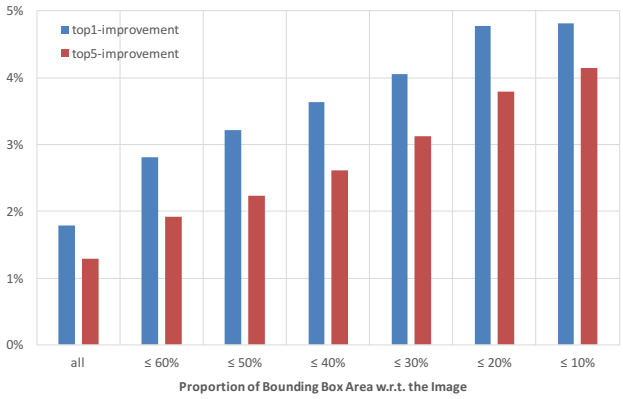
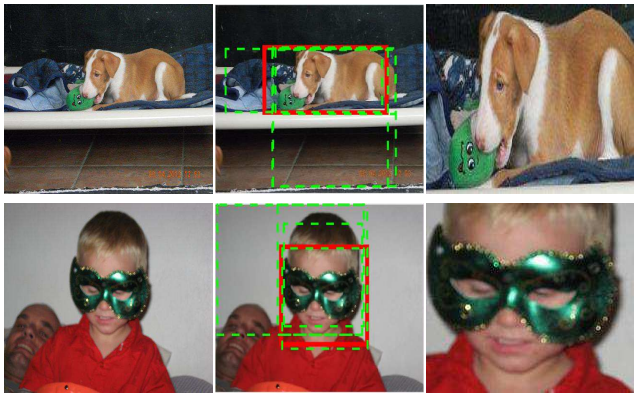


Figure 6. We divide the ImageNet 2014 validation set based on the proportion of the object size in the image. Classification accuracy using feedback crop for images increases with smaller objects. E.g., for those images with object area smaller than 20%, the top-1 classification accuracy increases significantly by almost 5%.



(a) Original Image (b) Attention Boxes (c) Cropped Image

Figure 7. We select two examples in the ImageNet 2014 validation set to demonstrate the re-classification mechanism. column (a) shows the original images which ConvNets predict incorrect labels (*"Italian greyhound"* and *"sunglass"*), column (b) shows the 5 calculated bounding box areas using the top-5 saliency maps, column (c) shows the cropped images obtained from the red box which ConvNets predict the correct labels (*"Ibizan hound"* and *"mask"*) with high confidences.

$50,000$ images with each image associated with one label. Table 3 shows the classification results using a pre-trained GoogleNet on the original full image and on the image patch based on feedback crop [2]. After the re-classification, the top 5 classification errors drops by 1.29%, and, moreover, top 1 error improves even more 1.79%. These results suggest that correct estimations of bounding boxes by a glance can provide more accurate classifications.

Ablative Study: To further understand how "Look and Think Twice" improves the classification task, we divide the ImageNet 2014 validation set based on the proportion of the object size in the image. Figure 6 shows the ablative study. We find that with feedbacks, classification errors drop significantly for images with smaller objects. For example, for objects of less than 20% area of images, the top-1 classification error drops with almost 5%. This suggests traditional ConvNet is powerless in recognzing small objects with cluttered backgrounds, especially when the image is resized to a fixed resolution (*i.e.* $224 * 224$). Our algorithm, on contrary, could focus the network's attention onto the salient areas, extracting image patches with enough resolutions on the potential objects.

5. Conclusion & Discussion

We propose a Feedback Convolutional Neural Network architecture in this paper, which achieves the *top-down* selectivity of neuron activations by jointly reasoning the outputs of class nodes and the activations of hidden layer neurons during the feedback loop. The proposed Feedback CNN is capable of capturing high level semantic concepts and project the information down to image representation as salience maps. Benefiting from the feedback mechanism of our model, we utilize the salience map to build a unified deep neural network for both recognition and object localization tasks, to answer questions of both *"What"* and *"Where"* simultaneously. Experimental results on ImageNet 2014 object localization Challenge show that our model could achieve competitive performance compared with state-of-the-arts, using only weakly supervised information. We also show the feedback could improve image classification task by re-focusing the network onto those salient regions.

We foresee the potential of Feedback CNN to further improve various computer vision and machine learning tasks, such as fine-grained recognition, object detection, and multi-tasks learning. However, instead of simulating the human vision system, we attribute the improvement of Feedback CNN to the efficiency in utilizing computation resources.

[2]Model from Caffe Model Zoo

Acknowledgement

We gratefully acknowledge the support of NVIDIA Corporation with the donation of the Tesla K40 GPUs used in the prototyping stage of this research. This research is supported by National Science Foundation under Grant No. 1318971. This work is jointly supported by National Natural Science Foundation of China (61420106015, 61175003) and National Basic Research Program of China (2012CB316300).

References

[1] D. M. Beck and S. Kastner. Top-down and bottom-up mechanisms in biasing competition in the human brain. *Vision research*, 49(10):1154–1165, 2009.

[2] Y. Bengio, A. Courville, and P. Vincent. Representation learning: A review and new perspectives. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 35(8):1798–1828, 2013.

[3] R. M. Cichy, D. Pantazis, and A. Oliva. Resolving human object recognition in space and time. *Nature Publishing Group*, 17(3):455–462, Jan. 2014.

[4] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 248–255. IEEE, 2009.

[5] R. Desimone. Visual attention mediated by biased competition in extrastriate visual cortex. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, 353(1373):1245–1255, 1998.

[6] R. Desimone and J. Duncan. Neural mechanisms of selective visual attention. *Annual review of neuroscience*, 18(1):193–222, 1995.

[7] D. Erhan, C. Szegedy, A. Toshev, and D. Anguelov. Scalable object detection using deep neural networks. In *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*, pages 2155–2162. IEEE, 2014.

[8] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. Object Detection with Discriminatively Trained Part-Based Models. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 32(9):1627–1645, 2010.

[9] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*, pages 580–587. IEEE, 2014.

[10] K. Gregor, I. Danihelka, A. Graves, and D. Wierstra. Draw: A recurrent neural network for image generation. *arXiv preprint arXiv:1502.04623*, 2015.

[11] K. He, X. Zhang, S. Ren, and J. Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. *arXiv preprint arXiv:1502.01852*, 2015.

[12] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[13] S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*, 2015.

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

[15] A. Krizhevsky, I. Sutskever, and G. E. Hinton. ImageNet classification with deep convolutional neural networks, 2012.

[16] 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? 35(8):1847–1871.

[17] Q. V. Le. Building high-level features using large scale unsupervised learning. In *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, pages 8595–8598. IEEE, 2013.

[18] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[19] T. S. Lee and D. Mumford. Hierarchical bayesian inference in the visual cortex. *JOSA A*, 20(7):1434–1448, 2003.

[20] M. Lin, Q. Chen, and S. Yan. Network in network. *arXiv preprint arXiv:1312.4400*, 2013.

[21] V. Mnih, N. Heess, A. Graves, and K. Kavukcuoglu. Recurrent Models of Visual Attention. *NIPS*, June 2014.

[22] N. C. Rust and J. J. DiCarlo. Selectivity and Tolerance ("Invariance") Both Increase as Visual Information Propagates from Cortical Area V4 to IT. *Journal of Neuroscience*, 30(39):12978–12995, Sept. 2010.

[23] R. Salakhutdinov and G. E. Hinton. Deep boltzmann machines. In *International Conference on Artificial Intelligence and Statistics*, pages 448–455, 2009.

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

[25] K. Simonyan and A. Zisserman. Very Deep Convolutional Networks for Large-Scale Image Recognition. *arXiv.org*, Sept. 2014.

[26] K. Sohn, G. Zhou, C. Lee, and H. Lee. Learning and selecting features jointly with point-wise gated {B} oltzmann machines. In *Proceedings of The 30th International Conference on Machine Learning*, pages 217–225, 2013.

[27] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, 15(1):1929–1958, 2014.

[28] M. Stollenga, J. Masci, F. Gomez, and J. Schmidhuber. Deep Networks with Internal Selective Attention through Feedback Connections. *arXiv.org*, July 2014.

[29] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going Deeper with Convolutions. *arXiv.org*, Sept. 2014.

[30] J. R. Uijlings, K. E. van de Sande, T. Gevers, and A. W. Smeulders. Selective search for object recognition. *International journal of computer vision*, 104(2):154–171, 2013.

[31] Q. Wang, J. Zhang, S. Song, and Z. Zhang. Attentional neural network: Feature selection using cognitive feedback. In *Advances in Neural Information Processing Systems*, pages 2033–2041, 2014.

[32] B. Yuri and J. Marie-Pierre. Interactive graph cuts for optimal boundary and region segmentation of objects in nd images. In *IEEEInternational Conference on Computer Vision. USA: IEEE*, volume 112, 2001.

[33] M. D. Zeiler and R. Fergus. Visualizing and understanding convolutional networks. In *Computer Vision–ECCV 2014*, pages 818–833. Springer, 2014.

[34] M. D. Zeiler, G. W. Taylor, and R. Fergus. Adaptive deconvolutional networks for mid and high level feature learning. *Computer Vision (ICCV . . .*, pages 2018–2025, 2011.