ICCV #1344

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

Anonymous ICCV submission

Paper ID 1344

Abstract

While feedforward deep convolutional neural networks (ConvNets) have been a great success in computer vision, it is important to remember that the human visual contex contains generally more feedback connections than foward 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 the deep neural networks. The proposed networks perform inference from image features in a bottom-up manner as traditional convolutional networks; while during feedback loops it sets up high-level semantic labels as the goal to infer the activation status of hidden layer neurons. The feedback networks help us better visualize and understand on how deep neural networks work as well as capture visual attention on expected objects, even in the images with cluttered background and multiple objects.

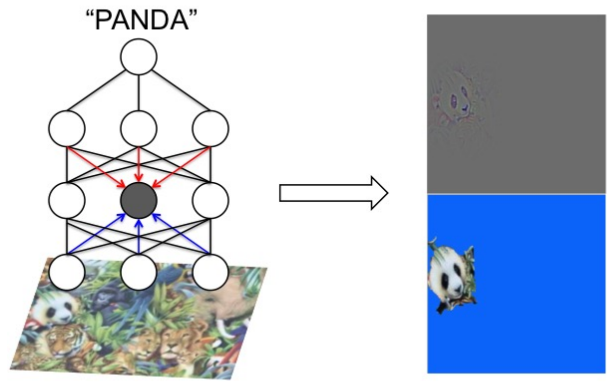


Figure 1. We propose a novel Feedback Convolutional Net model for captuing visual attention by infering the status of hidden neuron activations. The feedback net is designed to utilize both bottom-up image features and top-down semantic labels to infer the hidden neuron activations. The salient area captured by feedback often matches the correponding "target" object, even in the images with cluttered background and multiple objects.

1. Introduction

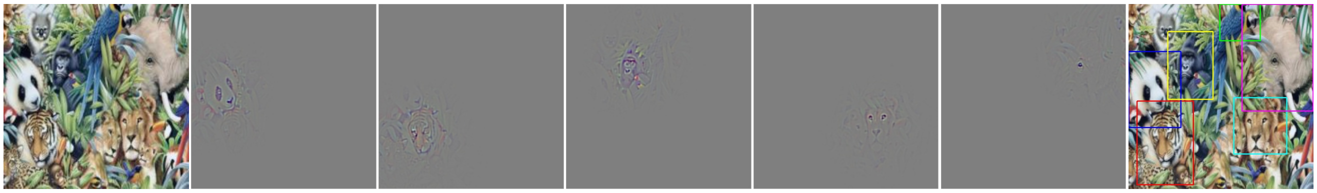
"What did you see in this image?"
"Panda, Tiger, Elephant, Lions."
"Have you seen the Gorilla?"
"Oh! I even didn't notice there is a Gorilla !"

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 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 [15], which reduces the chance of recognition being interfered with either noises or distractive patterns.

Inspired by the above evidences, we present a novel *Feedback Convolutional Neural Network* architecture in this paper. It achieves this selectivity by jointly reasoning the outputs of class nodes and the activations of hidden layer neurons during the feedback loop. As shown in Figure 1, during the feedforward stage, the proposed networks perform inference from image features 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. The networks are powerful enough to apply for class model visualization [22, 29] and object localization even in cluttered scenes with multiple objects.

1.0.1 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*. Convolutional Neural Networks [16, 14, 23] have achieved great success in both machine learning



(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 and GoogleNet recongize as ”comic book”. (b) - (f) illustrate our feedback modelon 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.

and computer vision recent years. Benefit from large scale of 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 ability of CNN, by either adding regularization in training [11, 13], or going deeper [23, 26]. Inspired by the Deformable Part-Based Models (DPM) [8] that model 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 searching. 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 receptive fields.

As the example shown in Figure 1, 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. Figure 2 also shows the visualizations of saliencies given different semantic concepts for the same input image. 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 importatn image regions that improve the class confidence.

Weakly Supervised Object Localization

Given the salience maps shown in Figure 2 and Figure 4, we further develop a weakly supervised object localization algorithm. 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], we don’t require any localization information in the training stage. In this case, we utilize *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 show that our weakly supervised algorithm using feedback network could achieve similar performance on ImageNet object localization task as GoogLeNet [26] and VGG [23].

The remainder of this paper is organized as follows: Section 2 introduces the related work, while we formulate our algorithm in Section 3. Experiments of visualization and object localization are demonstrated in Section 4. We conclude this work and future directions in Section 5

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 [14, 23, 18, 21, 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 [24], PReLU [11], Batch Normalization [13]; or 2) making the network deeper [26, 23].

Despite great successes have been 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, 20, 15, 17]. Recently, tentative efforts have been made to involve feedback strategy into Deep Neural Networks. Deep Boltzmman Machine (DBM) [21] and Deconvlutional Nerual Network [30] try to formulate the feedback as a reconstruction process within the training stage. In the meanwhile, recurrent neural network (RNN) and LSTM [12] are utilized to capture the attention drifting in a dynamic environment and learn the feedback mechanism via reinforcement learning [25, 19]. DRAW from Google DeepMind [10] combined above two into a generative model, to synthesis the image generation procedure.

As formulated 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 the "Top-Down" attention in human cognition. Hierarchical probabilistic computational model [17] is proposed to characterize feedback stimuli in a top-down manner. This is further incorporated into deep neural networks, for example, modeling feedback loops as latent variables in DBM [27], or using selectivity to resolve fine-grained classification [19], *et al.*. Due to the computational efficiency and capacity limitations of generative models used in [19, 27], they are hardly used in large scale datasets.

Vision is inherently ambiguous and could benefit from any prior knowledge from the high-level layers.

Also lateral connec- tions play an important role. This hints at the importance of processes like attention, expectation, top-down reasoning, imagination, and filling in.

Many computer vision systems try to work in a purely feed-forward fashion. However, vision is inherently ambiguous and benefits from any prior knowledge available.

2.2. Visualization, Detection, and Localization

Feedback is always related with visualization of CNN and object localization since both of these aim to project the high-level semantic information back to image representations. To visualize neuron responses and class models, deconvolution, various approaches are proposed either using deconvolution [29] or optimization based on gradients [22]. As demonstrated in [22], visualization of Convolutional Neural Network is showing semantically meaningful salient 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]. R-CNN [9] solves the detection problem by using region proposals instead of sliding windows. 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 [29, 22], a more feasible and cognitive manner for detection / localization could be derived by utilizing the saliency map generated in feedback visualizations. Moreover, an ideal approach should unify the recognition and detection in a single feedforward-feedback network architecture. However, the challenge lies on how to obtain semantically meaningful salience maps of high quality for each concept. That's the ultimate goal of our work presented in this paper.

3. Model

We first review the current state-of-the-art feedforward Deep Convolutional Neural Networks (CNNs) architecture and then propose our feedback model on top of that.

3.1. Review of Convolutional Neural Networks

The most recent state-of-the-art deep CNNs [23] 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 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 provide the nonlinearity for the network.

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

Max-Pooling Layer: The max-pooling layer is used to reduce the dimensionality of the output and model the invariance in deformable objects. The max-pooling operation is applied for every pixel location i, j around its small neighborhood $\mathcal{N}$.

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

3.2. Re-interpretion of ReLU and Max-Pooling

Both ReLU and Max-Pooling layer can be re-interpreted when we introduce the binary latent activation variables $\mathbf{z}$: $\mathbf{z} \in \{0, 1\}$. For the ReLU layer, $\mathbf{z}$ is the same size as the input $\mathbf{x}$, while for the max-pooling, $\mathbf{z}$ is similar as a set of location variant filters. Given that Equations 2 and 3 can be re-interpreted.

For the ReLU layer,

$$\mathbf{y} = \mathbf{z} \circ \mathbf{x} \tag{4}$$

where $\circ$ is the element wise product (Hadamard product). The element value in $\mathbf{z}$ is determine by $\mathbf{x}$ as the property of ReLU.

For the max-pooling layer,

$$\mathbf{y} = \mathbf{z} * \mathbf{x} \tag{5}$$

where $*$ is the convolution operation. The element value in $\mathbf{z}$ is determined by $\mathbf{x}$ as the property of max-pooling.

ICCV #1344

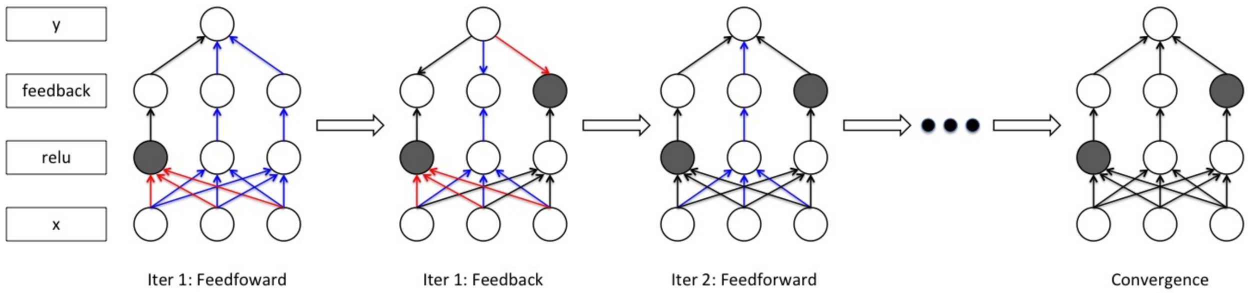


Figure 3. Illustration of our feedback model and its inference process. At the first iteration, the model performs as a feedforward neural net. After then, the neurons in the feedback hidden layers update their activation status to maximize the confidence output of the target top neuron. This process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

For both layers, the $\max()$ functions are replaced with linear operations between the inputs and binary hidden variables. The binary latent variables performs as feature selection. However, the values of the latent variables $\mathbf{z}$ are completely determined by the bottom-up input $\mathbf{x}$, meaning that the feature selection based on bottom-up.

During the feedforward process, since the middle layers have no idea about what is going on on the top layers, they have to keep the information passed from bottom layers as much as possible. Hence the final layer features computed for classifcation are sharable among all classes. It can be imagined that when ConvNets compute the middle layer features, it tries to make the largest possible information pass through the network. However, These features are stationary and even when there provide with prior information about top level neurons, the features won't change themselves.

3.3. Updating Hidden Neuron in Feedback

However, in order to recognize the image, the middle level layers are designed to provide all possible information for the final layer to classify. This works well when there is only one salient object in the image. However when people care about different aspects in the images, the same feature may not be appropriate all the time.

Since the model open all the gates for the gates for all information to pass, when we are targeted on a paticular semantic labels, we want to turn off those gates that provide irrelavant information for seeing that object. This top-down message will be utilized to turn off those relu.

We model the top down as another type of activation variable, similar as ReLu. However, this unit activates based on the the overall information of bottom-up responses and top-down messages.

To understand the model, we allow each ReLU neuron to either turn on, turn off, or pass a proportion of the information computed from the convolutional layer, in order to maximially interpret the image as the target class.

We model a feedback layer on top of the relu layers to improve the activation felxibility, that the feedback neuron activation variables depend on the weights from high level targets. When combined with ReLU layer, the two layers not only capture on the bottom-up features but also top-down weights passed from the target neuron. Figure. 3 illustrates the architecture of our model and the inference process.

Given an image I and a neural network with learned parameters w, we optimize the target neuron output by jointly reasoning neuron activiations h over all the hidden layers. Particularly, if the target neuron is the class node in the top layer, we optimize the class score S_c by looking for the reorganization of hidden neuron activations on the ReLU layer:

$$\begin{aligned} &\max_h S_c(I_0, h) \\ &s.t.\ h_i \in \{0, 1\} \end{aligned} \tag{6}$$

This leads to an integer programming problem, which is a NP-hard problem given the current deep convolutional neural net structure. To obtain a good solution, we apply a linear relaxation.

In the linear relaxation, we rephrase the problem as:

$$\begin{aligned} &\max_h S_c(I_0, h) \\ &s.t.\ 0 \leq h_i \leq 1 \end{aligned} \tag{7}$$

3.4. Relationship to the Deconvolutional Neural Networks

To understand how our inference algorithm work exactly, we compare it to the Deconvolutional neural networks model [].

we show below, DeconvNet-based reconstruction of the n-th layer input Xn is either equivalent or similar to computing the gradient of the visualised neuron ac-tivity f with respect to Xn, so DeconvNet effectively corresponds to the gradient back-propagation through a ConvNet.

We can conclude that apart from the RELU layer, computing the approximate feature map recon- struction Rn using a DeconvNet is equivalent to computing the derivative f/Xn using back- propagation, which is a part of our visualisation algorithms. Thus, gradient-based visualisation can be seen as the generalisation of that of [13], since the gradient-based techniques can be applied to the visualisation of activities in any layer, not just a convolutional one. In particular, in this paper we visualised the class score neurons in the final fully-connected layer.

Following this, deconvnet can be viewed as a one iteration of our hard optimization

3.5. Implementation Details

4. Experiment Results

4.1. Image Specific Class Model Visualization

Given an image I, a class label k and the hidden neuron activation states h, we approximate the neural net class score S_k with the first-order taylor expansion in the neighborhood of I:

$$S_k(I,h) \approx T_k(h)^T I + b \tag{8}$$

where $T_k(h)$ is the derivative of S_k with respect to the image at the point of I and hidden neuron status h.

$T_k(h)$ can be also viewed as the linear template applied on image I for the understanding of how like the image belongs to class k. We can visualize T since it's the same size as the image I. We use this technique to visualize the model through the paper.

Specifically, for a VggNet [23] which uses only a stack of piece-wise linear layers (i.e. convolution, relu, max-pooling) to compute the class scores, once the hidden states h encodes the selection of linear functions for each piece-wise linear operator, the final score is an linear operation on the image, equivalent to the inner product between the template and the image.

Comparison of Methods: We compare the visualization and saliency extraction of the feedbacked results against [22] (Oxford) and [29] (Deconv) on a set of complex images containing multiple objects from different classes. We show the qualititative results in Figure 4. All techniques use the same pre-trained VggNet[23] and ground truth class labels for each image are given. The visualization results before feedback is the same as Oxford, where all the hidden neurons states are determined by bottom-up computation. The visualization results after the feedback process are similar to Deconv, except that our model captures the most salient visual regions for each specific class.

Comparison of ConvNet Models: We compare the three most popluar ConvNets: AlexNet [14], VggNet [23] and Googlenet [26] by visualizing their feedback templates in Figure 5. All models are pre-trained on Caffe with the same classificatino accuracy reported. Ground truth class labels are given for each image before feedback. From the visualization results, we find that VggNet and GoogleNet produce more accurate visual attention than AlexNet, suggesting the benefit of using smaller convolution filters and deeper architectures to further distinguish similar and closeby objects. We also observe that, although both VggNet and GoogleNet produce very smilar image classification accuracies, GoogleNet better captures the salient object areas for each class than VggNet. We hypothesize that the two 4096 dimensional fully connected layers (i.e. fc6, fc7) in VggNet (which GoogleNet does not contain) could harm the spatial distinctiveness of the final image features,

Feedback Attention for Fine-Grained Classification: We show some interesting feedback visualization results for understanding the fine-grained classifcation ability of VggNet [23]. The VggNet is trained on ImageNet dataset [4] with 1000 object classes which contains ∼100 fine-graiend dog categories and ∼500 animal categories. We visualize the feedback templates of a few dog-cat images with respect to some dog classes and animal classes in Figure 6. We observe that each object class has its own special salient image features for distinction. Given the fixed images, some classes tend to emphasize on local region features such as parts (i.e. nose and ears), while others focus on global attributes (i.e. furry and tabby).

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 [22], the magnitude of the elements in the model template T_k in section 4.1 defines the class specific salience map for image I. A pixel with larger magnitude indicates that it is more important to the class k. We adopt the same saliency extraction strategy as [22] that a single class saliency value M_k for class k at pixel (i,j) is computed across all color channels: $M_k(i,j) = \max_{c\in rgb} |T_k(i,j,c)|$.

Although the ConvNet is pre-trained for image classification, we could use the feedbacked salience map for weakly supervised object localization. Following [22], given an image and the corresponding class saliency map, we compute the object segmentation mask using the GraphCut color segmentation [28]. During the inialization of graph cut, We set the pixels with the saliency higher than 95% quantile of the saliency distribution in the image as foreground and those with the saliency lower than 30% quantile as background. Once foreground and background segmentations are computed, the object segmentation mask is set to the largest connected component of the foreground pixels and the tighest bounding box is extracted as the lo-

Localization Error With Top 5 Predictions

Method	AlexNet	VggNet	GoogleNet
Oxford [22]	53.4	51.6	47.8
Deconv [29]	55.2	52.2	49.6
Feedback	52.3	49.0	**46.1**

Table 1. We compare the localization results on ImageNet 2014 validation set against Oxford and Deconv, using three different ConvNet models. Our feedback method clearly outperforms the baseline approaches for weakly supervised object localization. Notably, although VggNet and GoogleNet have ery similar image classfication accuray on ImageNet, our comparison suggests GoogleNet learns better middle level features than VggNet.

Localization Error With Ground Truth Labels

Model	Localization Error (%)
AlexNet [14]	47.1
VggNet [23]	42.3
GooglNet [26]	**40.1**

Table 2. In order to more fairly compare AlexNet agaisnt VggNet ana GoogleNet, we use the ground truth labels as the target labels for evaluating localization.

calization result.

We test our object localization method on the ImageNet 2014 localization validation set. We resize every image to 224x224 as the model required resolution and use the ground-truth class labels for the localization prediction. No further preprocessing or multi-scale strategy is involved. The predicted bounding box is considered as correct if its intersection over union with the ground truth bounding box is over 50%.

Comparison of Methods: Table 3 shows the comparsion of weakly supervised localization accuracy against Oxford and Deconv. We use the same VggNet and apply the same graph cut strategy on all the 3 models. Our method obtains 57% accuracy, outperforming both Oxford 50% and Deconv 53%, suggesting that in terms of capturing attention and localizing salient objects, our feedback net is better. Note that currently this is only top-down and weakly supervised and the object localization task is not taken into account during training.

Comparsion of ConvNet Models: We also compare the weakly supervised localization ability of the three most popular ConvNet models: AlexNet, VggNet and GoogleNet in Table 2, given the ground-truth class labels. AlexNet ... **I need the number to complete the paragraph.**

However, Most of the images in ImageNet 2014 dataset contain only one salient object. We further show some localization results on images with multiple object classes in Figure **??** ... **I need the figures to complete the paragraph**. Obviously our feedback net cannot distinguish multiple object instances from the same class, but could capture the salient areas, which could be utilized by other sophisticated object detection algorithms.

Localization Error With Ground Truth Labels

Method	Localization Error (%)
Oxford [22]	44.7
Deconv [29]	46.9
Feedback	**42.3**
VggNet-Supervised [23]	33.1

Table 3. We show that even given ground truth labels, our feedback method still outperforms baseline methods on ImageNet 2014 validation dataset.

5. Conclusion & Discussion

We propose a Feedback Convolutional Neural Network architecture in this paper, which achieves the selectivity of neuron activations by jointly reasoning outputs of class nodes and activations of hidden layer neurons during the feedback loop. The proposed Feedback CNN is capable to capture high level semantic concepts and project down to image representations as salience maps. Benefit from the feedback mechanism implemented in our model, we utilize the salience map to build a unified deep neural network for both recognition and object detection tasks, to answer the questions both “What” and “Where” simultaneously. Experimental results on ImageNet 2014 object localization Challenge show that our model could achieve competitive or even better performance compared with state-of-the-arts, using only weakly supervised information.

The Feedback CNN has the potential to improve various computer vision and machine learning tasks, such as fine-grained recognition, multi-task learning, and object detection. Moreover, we are seeing the possibilities to implement semi-supervised CNN using the proposed feedback architecture as future work.

References

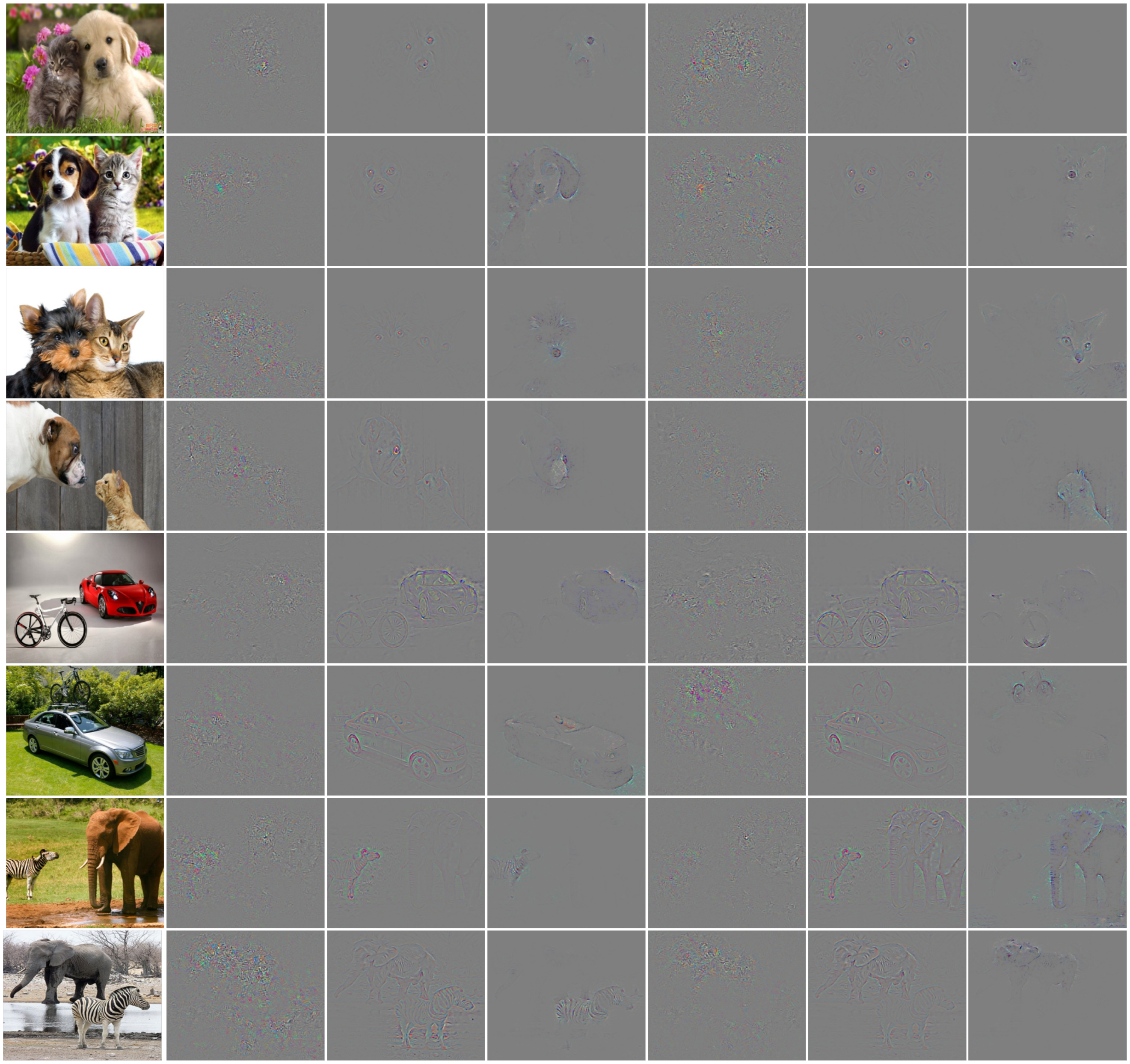
[1] D. M. Beck and S. Kastner. Top-down and bottom-up mechanisms in biasing competition in the human brain. *Vision research*, 49(10):1154–1165, 2009. 1, 2

[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. 2

[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. 1, 2

[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. 2, 5

[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. 1



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

Object 1: dog, car or zebra Object 2: cat, bike or elephant

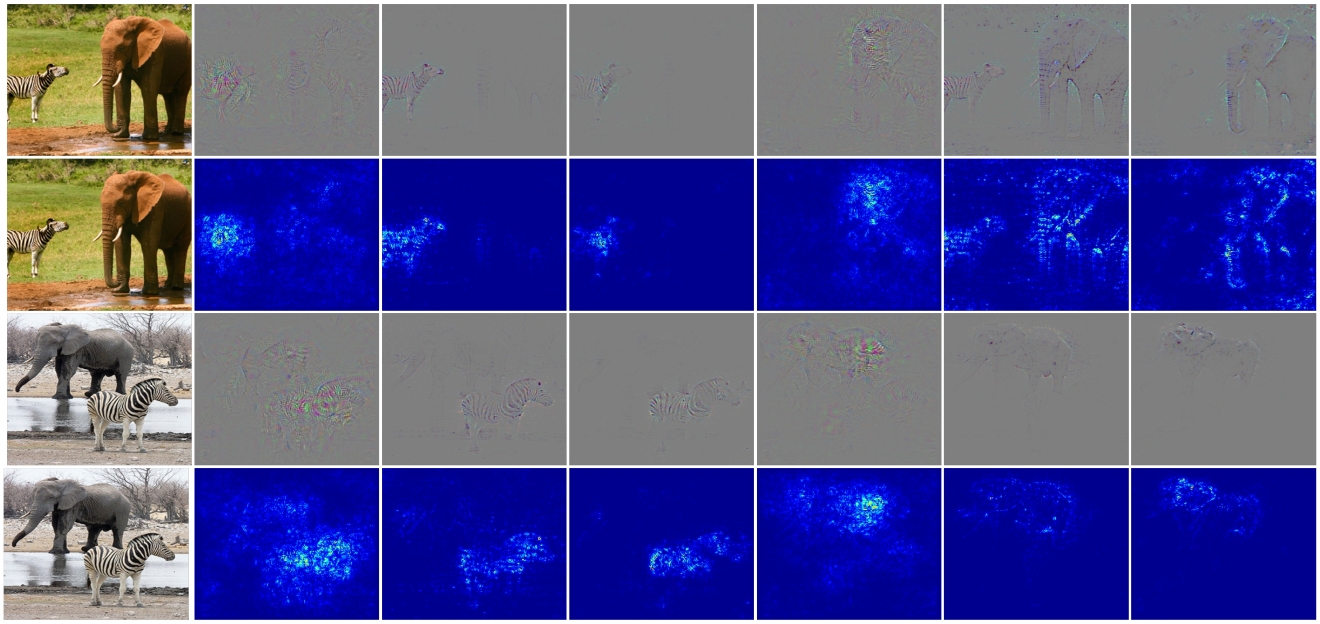
Figure 4. We demonstrate the effectiveness of feedback neural networks for class-specific feature extraction, by comparing the class model visualization results against original gradient [22] and Deconv [29] on selected images with multiple objects. All methods compute visualizations given a pre-trained GoogleNet. 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. 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 focus more on the correponding salient object area. Better viewed in color and zoom in.

[6] R. Desimone and J. Duncan. Neural mechanisms of selective visual attention. *Annual review of neuroscience*, 18(1):193–222, 1995. 1, 2

[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. 2, 3

[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,*

ICCV #1344



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

Figure 5. We visualize the feedback ability of three most popular pre-trained ConvNets: AlexNet, VggNet and GoogleNet, by visualizing the final image gradients and salience maps after feedback. We show the input images in column (a). We show the results of the three models feedbacked by "zebra" in column (b), (c), (d) and by "elephant" in column (e), (f), (g) repsectively. 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 accruacy, we find GoogleNet provides the better class specific feature separations. We suspect the two 4096 fully connected layers in VggNet (which GoogleNetdoes not have) could harm the spatial distinctiveness of image features.

IEEE Transactions on, 32(9):1627–1645, 2010. 2

[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. 2, 3

[10] K. Gregor, I. Danihelka, A. Graves, and D. Wierstra. Draw: A recurrent neural network for image generation. *arXiv preprint arXiv:1502.04623*, 2015. 2

[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. 2

[12] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997. 2

[13] S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*, 2015. 2

[14] A. Krizhevsky, I. Sutskever, and G. E. Hinton. ImageNet classification with deep convolutional neural networks, 2012. 1, 2, 5, 6

[15] 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. 1, 2

[16] 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. 1

[17] T. S. Lee and D. Mumford. Hierarchical bayesian inference in the visual cortex. *JOSA A*, 20(7):1434–1448, 2003. 2, 3

[18] M. Lin, Q. Chen, and S. Yan. Network in network. *arXiv preprint arXiv:1312.4400*, 2013. 2

[19] V. Mnih, N. Heess, A. Graves, and K. Kavukcuoglu. Recurrent Models of Visual Attention. *NIPS*, June 2014. 2, 3

[20] 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. 2

[21] R. Salakhutdinov and G. E. Hinton. Deep boltzmann machines. In *International Conference on Artificial Intelligence and Statistics*, pages 448–455, 2009. 2

[22] K. Simonyan, A. Vedaldi, and A. Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*, 2013. 1, 3, 5, 6, 7

[23] K. Simonyan and A. Zisserman. Very Deep Convolutional Networks for Large-Scale Image Recognition. *arXiv.org*, Sept. 2014. 1, 2, 3, 5, 6

[24] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov. Dropout: A simple way to prevent neural

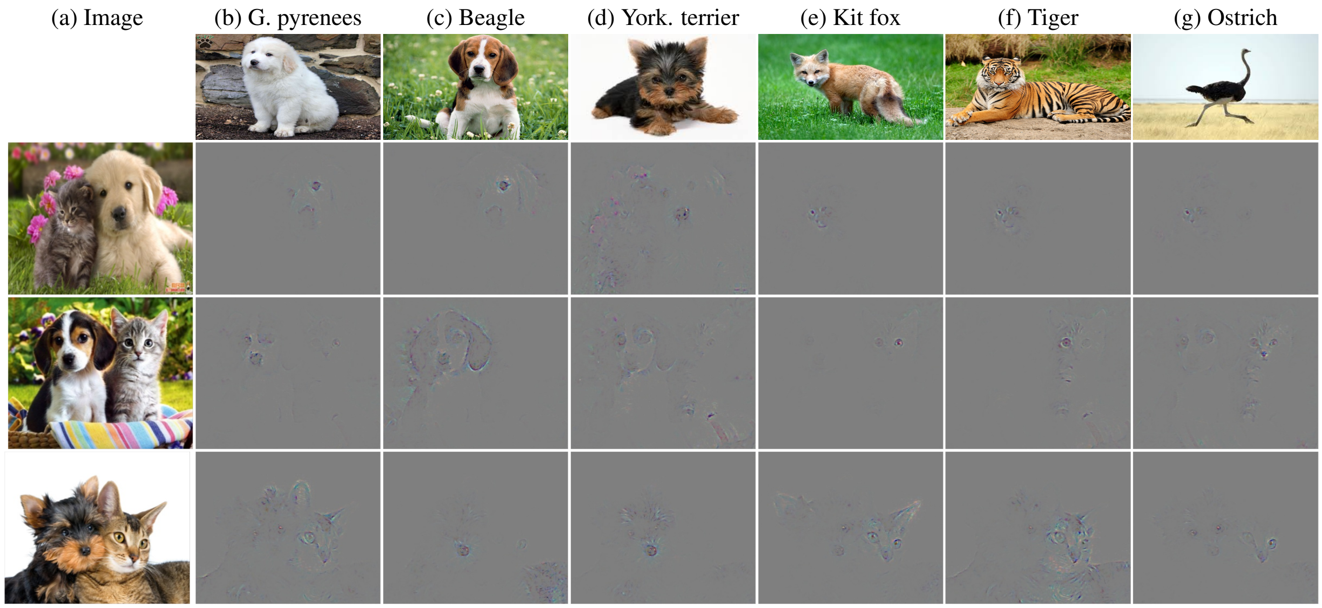


Figure 6. We show some interesting visualizations for the understanding of fine-grained classification by comparsing against the feedback gradients of ground truth labels and other classes. The top row shows the class labels and a representative image for each class for the ease of understanding. Column (a) shows the three examplar input images, their ground truth labels are great pyrenees, beagle and yorkshire terrier respectively. We can see that although (b), (c) and (d) are all dogs, their salient area for distinction are quite different. Noses are one of the most important feature for classifying dogs, but ears are specific feature for beagles, while fluffy is more importatnt to yorkshire terrier. When the top down is from (e) kit fox, features on the cat in the last row is more fox-specific: nose and ears. When top down is from (f) tiger, features on the same at is more tiger-specific: textures. And when it's (g) ostrich, nothing special come out.

networks from overfitting. *The Journal of Machine Learning Research*, 15(1):1929–1958, 2014. 2

[25] M. Stollenga, J. Masci, F. Gomez, and J. Schmidhuber. Deep Networks with Internal Selective Attention through Feedback Connections. *arXiv.org*, July 2014. 2

[26] 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. 2, 5, 6

[27] 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. 3

[28] B. Yuri and J. Marie-Pierre. Interactive graph cuts for optimal boundaryand region segmentation of objects in nd images. In *IEEEInternational Conference on Computer Vision. USA: IEEE*, volume 112, 2001. 5

[29] M. D. Zeiler and R. Fergus. Visualizing and understanding convolutional networks. In *Computer Vision–ECCV 2014*, pages 818–833. Springer, 2014. 1, 3, 5, 6, 7

[30] 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. 2