Deep Feedback Neural Networks for Weakly Supervised Object Localization

Anonymous ICCV submission

Paper ID 1344

Abstract

Deep convolutional neural networks have been proven to be a very powerful method in computer vision. 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 in the past several weeks. The feedback philosophy helps us to visualize the neural network and understand deeper on how deep neural network works, especially for the deep convolutional neural networks. The feedback framework is also extended to re-train the neural networks to better explore the properties of the natural images to avoid overfitting as well as improve the image recognition accuracy. We show will discuss the plans on future improving the feedback neural network architectures.

1. Introduction

[**?**] We present a novel feedback neural networks for joint reasoning the class node and hidden layer information. The network is powerful to be applied on model class visualization and object localization even in cluttered scenes with multi objects. The framework is novel and

Deep Learning and Deep Convolutional Neural Networks, Feedforward Strcture Deep Convolutional networks (ConvNets) have recently enjoyed a great success in large-scale im- age and video recognition (Krizhevsky et al., 2012; Zeiler Fergus, 2013; Sermanet et al., 2014; Simonyan Zisserman, 2014) which has become possible due to the large public image reposito- ries, such as ImageNet (Deng et al., 2009), and high-performance computing systems, such as GPUs or large-scale distributed clusters (Dean et al., 2012). In particular, an important role in the advance of deep visual recognition architectures has been played by the ImageNet Large-Scale Visual Recog- nition Challenge (ILSVRC) (Russakovsky et al., 2014), which has served as a testbed for a few generations of large-scale image classification systems, from high-dimensional shallow feature encodings (Perronnin et al., 2010) (the winner of ILSVRC-2011) to deep ConvNets (Krizhevsky et al., 2012) (the winner of ILSVRC-2012).

Psychological feedback, inference top-down and bottom-up While we have outlined in this paper a hierarchical feedfor- ward view on visual processing, it is important to remember that within the visual cortex there are generally more feedback connections than forward connections. 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 feedforward fashion. However, vision is inherently ambiguous and benefits from any prior knowledge available. This may even imply that the knowledge of how the tower of Pisa looks influences the perception of an edge on the level of V1. It also means that a system should be able to produce several hypotheses that are concurrently considered and possibly not resolved [102].

This paper Main Contribution In this paper, we address the visualisation of deep image classification ConvNets, trained on the large-scale ImageNet challenge dataset [2]. To this end, we make the following three contributions. First, we demonstrate that understandable visualisations of ConvNet classification models can be obtained using the numerical optimisation of the input image [5] (Sect. 2). Note, in our case, unlike [5], the net is trained in a supervised manner, so we know which neuron in the final fully-connected clas- sification layer should be maximised to visualise the class of interest (in the unsupervised case, [9] had to use a separate annotated image set to find out the neuron responsible for a particular class). To the best of our knowledge, we are the first to apply the method of [5] to the visualisation of ImageNet classification ConvNets [8]. Second, we propose a method for computing the spatial support of a given class in a given image (image-specific class saliency map) using a single back-propagation pass through a classification ConvNet (Sect. 3). As discussed in Sect. 3.2, such saliency maps can be used for weakly supervised object localisation. Finally, we show in Sect. 4 that the gradient-based visualisation methods generalise the deconvolutional network reconstruction procedure [13].

Yurgen's feedback neural networks, Attention neural networks, Deep Boltzman Machines

DPM Top-down, weakly supervised object detection, localization and parsing We describe an object detection system that represents highly variable objects using mixtures of multiscale de- formable part models. These models are trained using a discriminative procedure that only requires bounding boxes for the objects in a set of images. The resulting system is both efficient and accurate, achieving state-of- the-art results on the PASCAL VOC benchmarks [11] [13] and the INRIA Person dataset [10]. Our approach builds on the pictorial structures frame- work [15], [20]. Pictorial structures represent objects by a collection of parts arranged in a deformable configu- ration. Each part captures local appearance properties of an object while the deformable configuration is charac- terized by spring-like connections between certain pairs of parts. Detections obtained with a single component person model. The model is defined by a coarse root filter (a), several higher resolution part filters (b) and a spatial model for the location of each part relative to the root (c). The filters specify weights for histogram of oriented gradients features. Their visualization show the positive weights at different orientations. The visualization of the spatial models reflects the cost of placing the center of a part at different locations relative to the root.

Comparing against Oxford and Deconv

ConvNet Implementation Details ConvNet implementation details. Our visualisation experiments were carried out using a single deep ConvNet, trained on the ILSVRC-2013 dataset [2], which includes 1.2M training images, labelled into 1000 classes. Our ConvNet is similar to that of [8] and is implemented using their cuda-convnet toolbox1, although our net is less wide, and we used additional image jittering, based on zeroing-out random parts of an image. Our weight layer configuration is: conv64-conv256- conv256-conv256-conv256-full4096-full4096-full1000, where convN denotes a convolutional layer with N filters, fullM a fully-connected layer with M outputs. On ILSVRC-2013 validation set, the network achieves the top-1/top-5 classification error of 39.7

2. Related Work

Deep Feed-foward Convolutional Neural Networks Feedforward multilayer neural networks have achieved good performance in many classification tasks in the past few years, notably achieving the best performance in the ImageNet competition in vision([21] [7]). However, they typically give a fixed outcome for each input image, therefore cannot naturally model the influence of cognitive biases and are difficult to incorporate into a larger cognitive framework. The current frontier of vision research is to go beyond object recognition towards image understanding [16]. Inspired by neuroscience research, we believe that an unified module which integrates feedback predictions and interpretations with information from the world is an important step towards this goal.

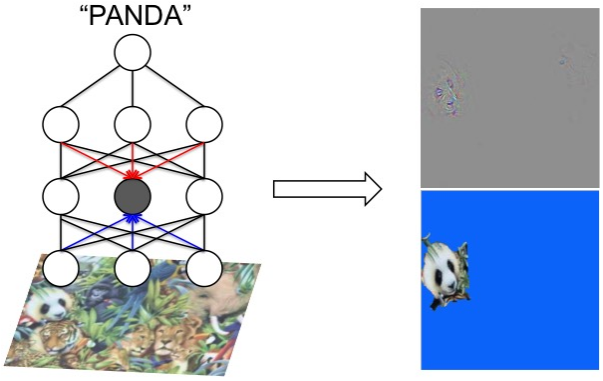


Figure 1. We propose a novel feedback convnet for weakly supervised object localization in complex scenes with cluttered background. The feedback net is able to utilized both bottom-up image features and top-down semantic labels to infer the hidden layer neuron status to match the localize the correponding salient area in the image.

Feedback Neural Networks and Attention Models

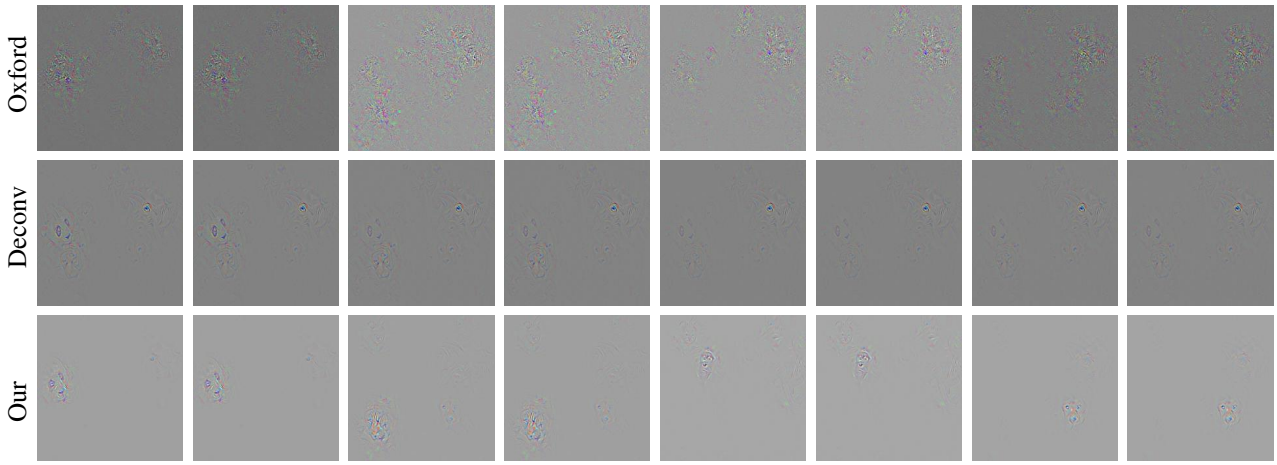
Deep Boltzman Machine and Deep Belief Networks Generative models have been a popular approach([5, 13]). They are typically based on a probabilis- tic framework such as Boltzmann Machines and can be stacked into a deep architecture. They have advantages over discriminative models in dealing with object occlusion. In addition, prior knowl- edge can be easily incorporated in generative models in the forms of latent variables. However, despite the mathematical beauty of a probabilistic framework, this class of models currently suffer from the difficulty of generative learning and have been mostly successful in learning small patches of natural images and objects [17, 22, 13]. In addition, inferring the hidden variables from images is a difficult process and many iterations are typically needed for the model to converge[13, 15]. A recent trend is to first train a DBN or DBM model then turn the model into a discriminative network for classification. This allows for fast recognition but the discriminative network loses the generative ability and cannot combine top-down and bottom-up information.

Incorporating Top-down for Part Localization

Visualizing and Understanding Neural Networks In previous work, Erhan et al. [5] visualised deep models by finding an input image which max- imises the neuron activity of interest by carrying out an optimisation using gradient ascent in the image space. The method was used to visualise the hidden feature layers of unsupervised deep ar- chitectures, such as the Deep Belief Network (DBN)

(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". (b) - (f) demonstrate the powerfulness of our model understanding the image given paticular object labels. We visualization the gradient of each label w.r.t. image after ther convergence of the feedback neural nets (g) shows the localization power for different objects in this complex image based on the gradient. Note that the weights in the net is obtained from a pre-trained feedforward GoogleNet model for image classification.



(a) Panda (b) Tiger (c) Gorilla (d) Lion

Figure 3. We demonstrate the effectiveness of our method by comparing the class model visualization results against Oxford [] and Deconv []. The input image is the same as Figure 1 (a). We show both the visualization results as well as the saliency map. While both Oxford and Deconv have the same input: the image and an object class label (i.e. tiger, panda, etc.), the gradients computed are often salient on one particular object (i.e. elephant). Our feedback framework allows for the model to focus on the most important image area that improves the class confidence.

[7], and it was later employed by Le et al.[9] to visualise the class models, captured by a deep unsupervised auto-encoder. Recently, the problem of ConvNet visualisation was addressed by Zeiler et al.[13]. For convolutional layer visualisation, they proposed the Deconvolutional Network (DeconvNet) architecture, which aims to approximately reconstruct the input of each layer from its output. In this paper, we address the visualisation of deep image classification ConvNets, trained on the large-scale ImageNet challenge dataset [2]. To this end, we make the following three contributions. First, we demonstrate that understandable visualisations of ConvNet classification models can be obtained using the numerical optimisation of the input image [5] (Sect. 2). Note, in our case, unlike [5], the net is trained in a supervised manner, so we know which neuron in the final fully-connected clas- sification layer should be maximised to visualise the class of interest (in the unsupervised case, [9] had to use a separate annotated image set to find out the neuron responsible for a particular class). To the best of our knowledge, we are the first to apply the method of [5] to the visualisation of ImageNet classification ConvNets [8]. Second, we propose a method for computing the spatial support of a given class in a given image (image-specific class saliency map) using a single back-propagation pass through a classification ConvNet (Sect. 3). As discussed in Sect. 3.2, such saliency maps can be used for weakly supervised object localisation. Finally, we show in Sect. 4 that the gradient-based visualisation methods generalise the deconvolutional network reconstruction procedure [13].

Weakly Supervised Object Localization

Feedback Neural Networks and Attention Neural Networks Special form of Recurrent Neural Networks The dasNet Network. Each image in classified after T passes through the network. After each for- ward propagation through the Maxout net, the output classification vec-

tor, the output of the second to last layer, and the averages of all feature maps, are combined into an observation vector that is used by a deterministic policy to choose an action that changes the weights of all the feature maps for the next pass of the same image. After pass T , the output of the Maxout net is finally used to classify the image.

3. Model

3.1. Convolutional Neural Networks

3.2. Infering the Hidden Neurons, the Discriminative Framework

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.

In practice, we treat the inference process as discriminatively optimize the final class node.

Optimizing such function results in an integer programming, if we treat h as binary variables.

During optimization, we proposed two ways to deal with it, 1. coordinate descent 2. continuous relaxation

Hard optimization: The coordinate descent frameworks stand in this way: 1. Initialize h as all 1 meaning the gate is open, compute feedfoward messages to the class node, then given the current activation status, optimize the last layer h to maximize the class output, given the updates last layer h, keep optimizing lower layers. And reiterate this process.

Soft optimization: The continuous relaxation falls in below way, compute the gradient of class node y given h, use gradient descent update h, and keep until this until convergence.

3.3. Class Model Visualization

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

Relationship to Oxford and Deconv

3.4. Object Localization

In this section we describe how a classification ConvNet can be queried about the spatial support of a particular class in a given image. Given an image I0, a class c, and a classification ConvNet with the class score function Sc(I), we would like to rank the pixels of I0 based on their influence on the score Sc(I0). We start with a motivational example. Consider the linear score model for the class c: where the image I is represented in the vectorised (one-dimensional) form, and wc and bc are respectively the weight vector and the bias of the model. In this case, it is easy to see that the magnitude of elements of w defines the importance of the corresponding pixels of I for the class c. In the case of deep ConvNets, the class score Sc(I) is a highly non-linear function of I, so the reasoning of the previous paragraph can not be immediately applied. However, given an image I0, we can approximate Sc(I) with a linear function in the neighbourhood of I0 by computing the first-order Taylor expansion: Another interpretation of computing the image-specific class saliency using the class score deriva- tive (4) is that the magnitude of the derivative indicates which pixels need to be changed the least to affect the class score the most. One can expect that such pixels correspond to the object location in the image. We note that a similar technique has been previously applied by [1] in the context of Bayesian classification.

Given an image I0 (with m rows and n columns) and a class c, the class saliency map M Rmn is computed as follows. First, the derivative w (4) is found by back-propagation. After that, the saliency map is obtained by rearranging the elements of the vector w. In the case of a grey-scale image, the number of elements in w is equal to the number of pixels in I0, so the map can be computed as Mij = —wh(i,j)—, where h(i, j) is the index of the element of w, corresponding to the image pixel in the i-th row and j-th column. In the case of the multi-channel (e.g. RGB) image, let us assume that the colour channel c of the pixel (i, j) of image I corresponds to the element of w with the index h(i,j,c). To derive a single class saliency value for each pixel (i,j), we took the maximum magnitude of w across all colour channels: Mij = maxc —wh(i,j,c)—. It is important to note that the saliency maps are extracted using a classification ConvNet trained on the image labels, so no additional annotation is required (such as object bounding boxes or segmentation masks). The computation of the image-specific saliency map for a single class is extremely quick, since it only requires a single back-propagation pass. We visualise the saliency maps for the highest-scoring class (top-1 class prediction) on randomly se- lected ILSVRC-2013 test set images in Fig. 2. Similarly to the ConvNet classification procedure [8], where the class predictions are computed on 10 cropped and reflected sub-images, we computed 10 saliency maps on the 10 sub-images, and then averaged them. The weakly supervised class saliency maps (Sect. 3.1) encode the location of the object of the given class in the given image, and thus can be used for object localisation (in spite of being trained on image labels only). Here we briefly describe a simple object localisation procedure, which we used for the localisation task of the ILSVRC-2013 challenge [12]. Given an image and the corresponding class saliency map, we compute the object segmentation mask using the GraphCut colour segmentation [3]. The use of the colour segmentation is motivated by the fact that the saliency map might capture only the most discriminative part of an object, so saliency thresholding might not be able to highlight the whole object. Therefore, it is important to be able to propagate the thresholded map to other

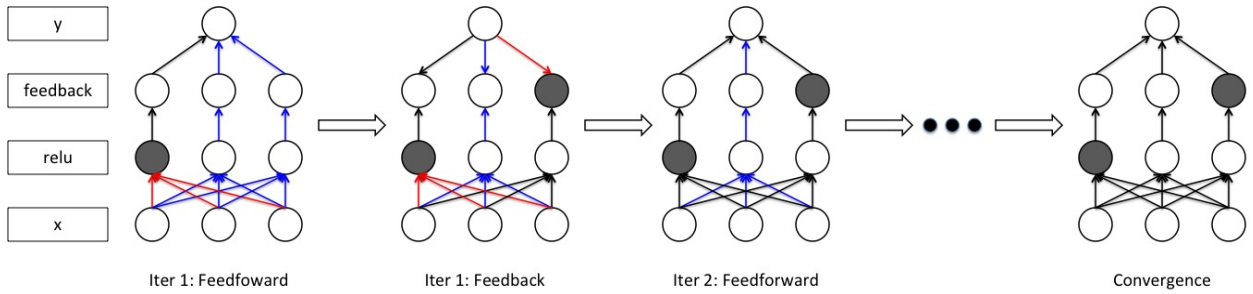


Figure 4. Illustration of our feedback model and its inference process. At the first iteration, the model performs as a feedforward neural net. Then given the top signal neuron, the hidden layers update their gates to maximize the confidence for the top neuron. This process continues until convergence.

parts of the object, which we aim to achieve here using the colour continuity cues. Foreground and background colour models were set to be the Gaussian Mixture Models. The foreground model was estimated from the pixels with the saliency higher than a threshold, set to the 95We entered our object localisation method into the ILSVRC-2013 localisation challenge. Consid- ering that the challenge requires the object bounding boxes to be reported, we computed them as the bounding boxes of the object segmentation masks. The procedure was repeated for each of the top-5 predicted classes. The method achieved 46.4

Relationship to Oxford and Deconv

4. Experiments

4.1. Datasets

4.2. Single Model Visualization on ImageNet

4.3. Class Model Visualization on Multiple Object Images

4.4. Construction Visualization

4.5. Localization

5. Conclusion & Discussion

We proposes a Feedback Convolusional Neural Networks for class model visualization and object localization. Our Feedback Neural Networks can infer the hidden neuron status given the bottom level input image and top level class labels. Experiments on ImageNet localization challengeindicates that our model is superior in weakly supervised object localization, and further experiments demonstrate its powerfulness in distinguishing objects, even under cluttered backgrounds with multiple objects.

(1) Robust (2) Multi-task (3)

Figure 1: We show the powerfulness of feedback neural networks on class model visualization and object localizations, even when an image contains very cluttered background and lots of salient objects. Note that we simply use

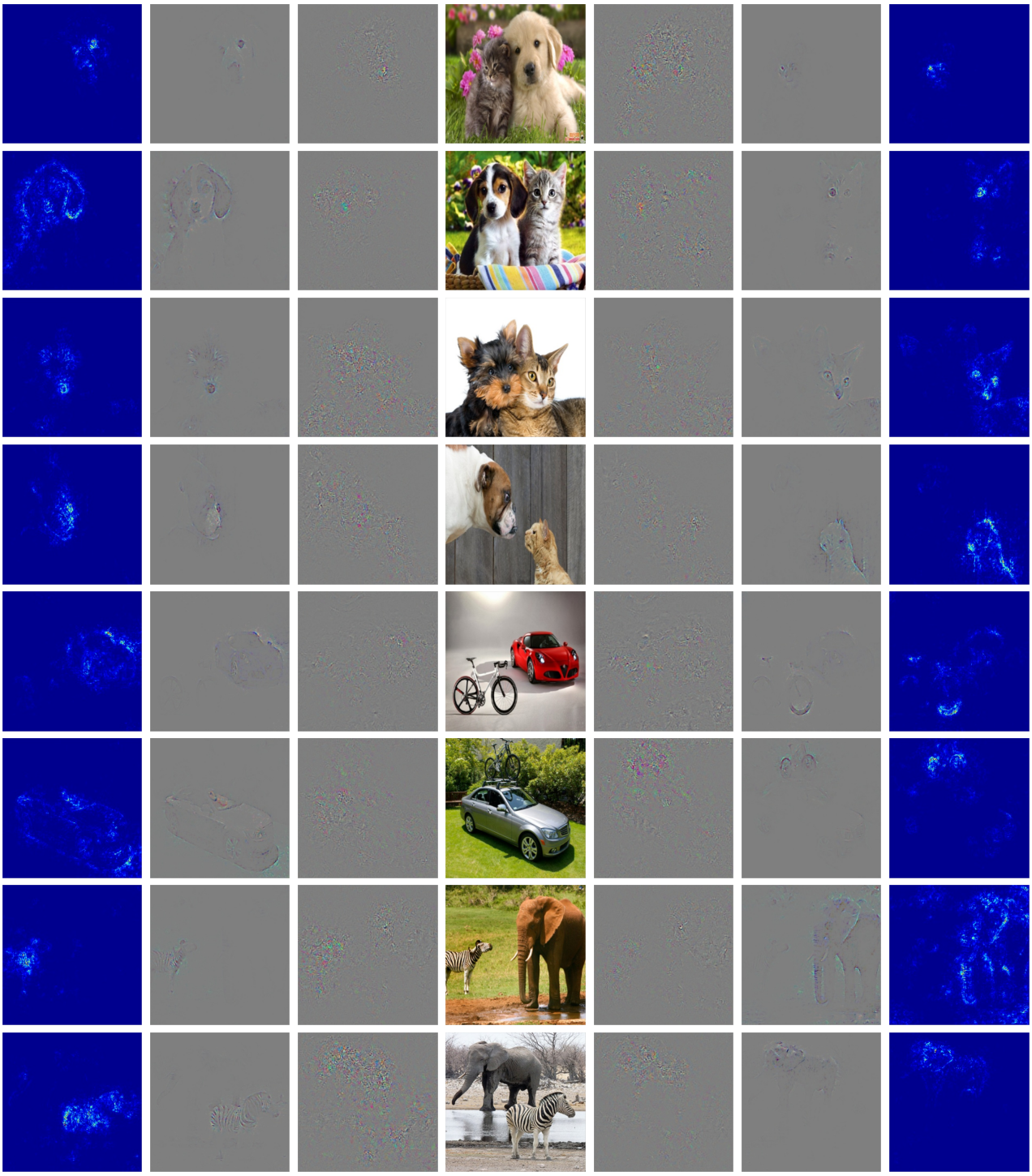
Evaluation on Imagenet 2014 localization test set

Method	Localization error	Classification error
VggNet-Supervised[]	25.3	7.4
GoogleNet-Supervised []	26.4	14.8
AlexNet-Weakly[]		
AlexNet-Weakly Feedback		
VggNet-Weakly Feedback		
GooglNet-Weakl Feedback		

Table 1. We show the localization evaluation on ImageNet 2014 localization competition. Our model clearly outperforms the weakly supervised approach based on []. Notably, we compare even fairbly well against supervisedly trained localization model, where an extra localization bouding boxes dataset is used. We demonstrate that when learning for classifications objectives, the Deep ConvNet already integrate powerful class specific features for attentioning on the important areas.

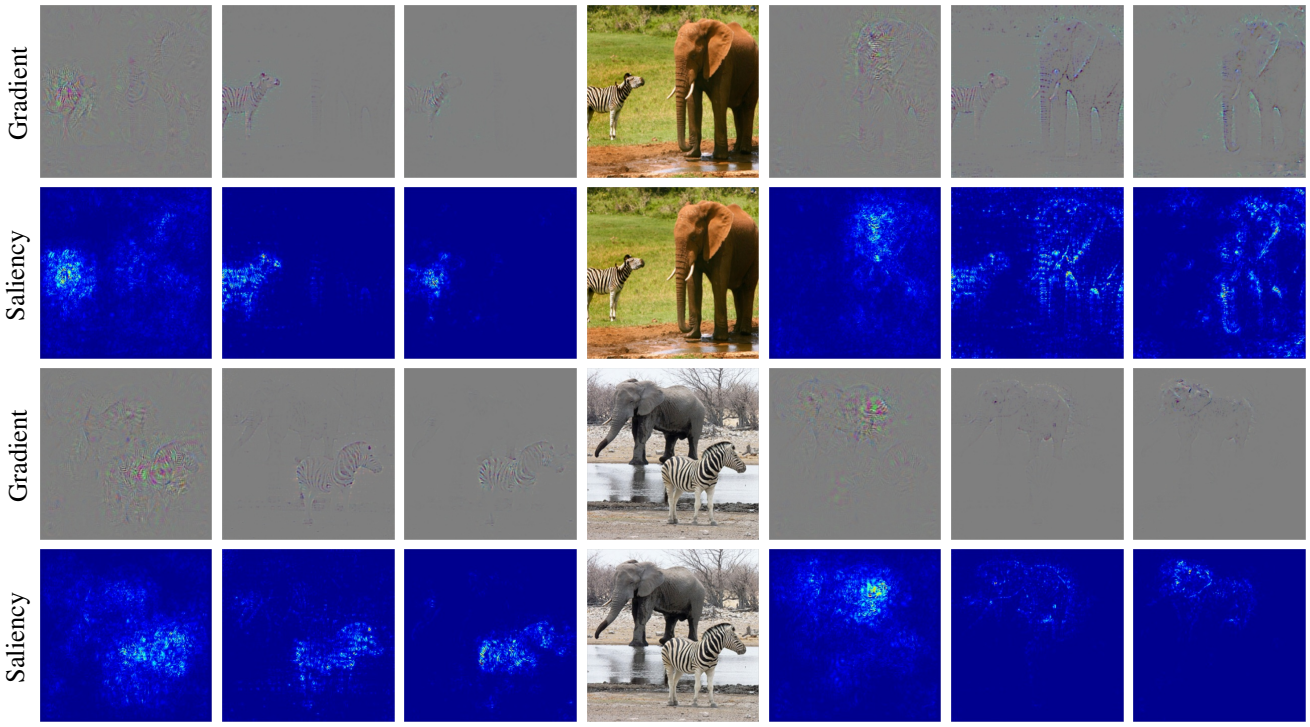
a pertained feedforward multi-class convnets (GoogleNet) model [] trained with ImageNet dataset where each training image only contains one object and no further training is involved. The feedback neural nets are able to adapt its middle-level hidden layers (usually represents object parts) by combining the bottom-up feedforward image features as well as top-down feedback semantic information. (a) input image (b) class model visualization given the class label: panda, elephant, gorilla, tiger, lion, (c) final object localization results based on the class model visualization.

References



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

Figure 5. We show more qualitative results of our feedback neural network on natual images. For each image, we show the original image gradient w.r.t. the class labels and the updated image gradients w.r.t. class labels after feedback. We select a few images for comparing bike v.s. car, zebra v.s. elephant, dog v.s. cat, indicating that our approach can find the targeted objects given only a trained feedforward convnets and class labels.



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

Figure 6. We show more qualitative results of our feedback neural network on natual images. For each image, we show the original image gradient w.r.t. the class labels and the updated image gradients w.r.t. class labels after feedback. We select a few images for comparing bike v.s. car, zebra v.s. elephant, dog v.s. cat, indicating that our approach can find the targeted objects given only a trained feedforward convnets and class labels.

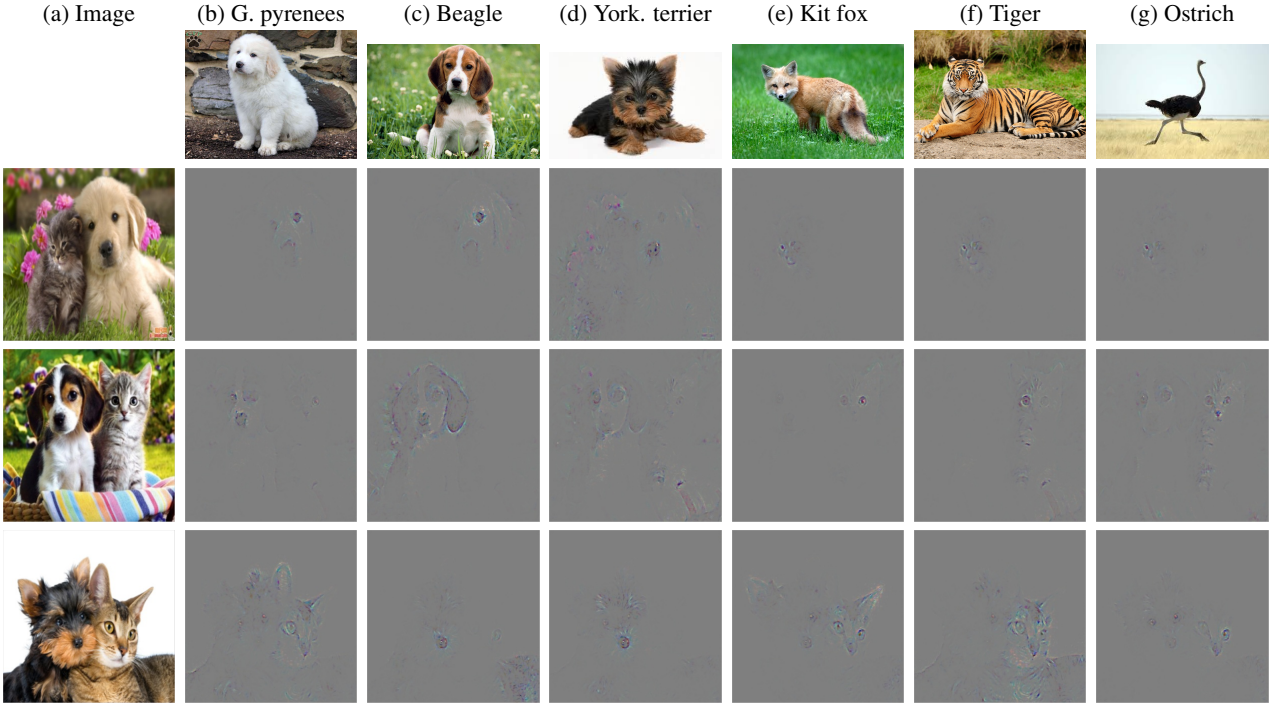


Figure 7. We show more qualitative results of our feedback neural network on natual images. For each image, we show the original image gradient w.r.t. the class labels and the updated image gradients w.r.t. class labels after feedback. We select a few images for comparing bike v.s. car, zebra v.s. elephant, dog v.s. cat, indicating that our approach can find the targeted objects given only a trained feedforward convnets and class labels.