


# Global Perception Feedback Convolutional Neural Networks

Chaoyou Fu , Xiang Wu, Jing Dong, and Ran He

Institution of Automation, Chinese Academy of Sciences, Beijing, China  
fuchaoyou2017@ia.ac.cn

**Abstract.** Top-down feedback mechanism is an important module of visual attention for weakly supervised learning. Previous top-down feedback convolutional neural networks often perform local perception during feedback. Inspired by the fact that the visual system is sensitive to global topological properties [1], we propose a global perception feedback convolutional neural network that considers the global structure of visual response during feedback inference. The global perception eliminates “Visual illusions” that are produced in the process of visual attention. It is achieved by simply imposing the trace norm on hidden neuron activations. Particularly, when updating the status of hidden neuron activations during gradient backpropagation, we get rid of some minor constituent in the SVD decomposition, which both ensures the global low-rank structure of feedback information and the elimination of local noise. Experimental results on the ImageNet dataset corroborate our claims and demonstrate the effectiveness of our global perception model.

**Keywords:** Feedback · Global perception · Weakly supervised learning

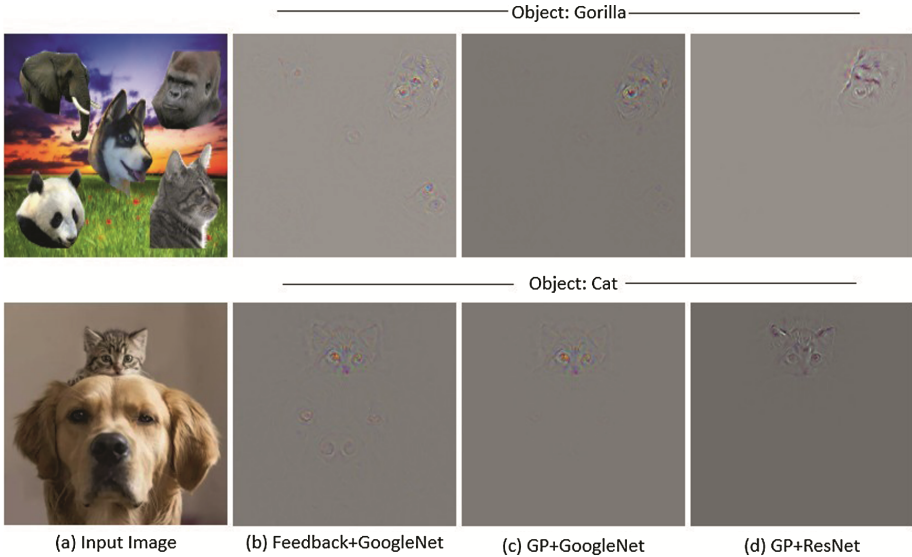
## 1 Introduction

In recent years, although deep convolutional neural networks (CNNs) have achieved great performance in computer vision and pattern recognition, these successful algorithms are mainly based on feedforward neural networks and neglect the top-down feedback mechanism that is important for the visual processing [2–5], especially for weakly-supervised or semi-supervised learning.

The top-down feedback mechanism of CNNs has attracted some research attentions recently, which focus on utilizing the top-down feedback mechanism to further increase generalization ability of CNNs. Zeiler et al. [6] proposed a deconvolution technique that projects feature responses back to the input pixel space for visualizing and understanding CNNs. Simonyan et al. [7] obtained a class saliency map by a single back-propagation pass with a given label. Springenberg et al. [8] got a clearer class saliency map by preventing the forward and backward pass of negative gradients. Cao et al. [9] inspired by “Biased Competition Theory” [10–12] and proposed an original feedback model to simulate visual attention by inferring the status of hidden neuron activations. Other top-down feedback methods that realize visualization or localization include [13–15].

Although these top-down feedback methods have got encouraging achievements, current methods only perform local perception and ignore the global structure during feedback. In the process of top-down local perception, neurons cannot be completely suppressed or activated because of the complex relationship between neurons. The status of a neuron is not only decided by external stimuli but also influenced by surrounding neurons. Cognitive neuroscientists explain this phenomenon as the “Visual illusions” [16], which increases the chance of recognition and detection being interfered with distractive patterns.

Considering the global topological properties of human visual system [1] and some works of global view [17, 18] or structural constraint [19–21], we present a novel framework towards a feedback CNNs to eliminate “Visual illusions” in this paper. Our key innovation lies in introducing a Global Perception (GP) algorithm, which explicitly constrains the structure of inter-layers in a global way. By combining the global perception with local perception, the distribution of active neurons in hidden layers is compulsively constrained and the phenomenon of “Visual illusions” almost disappeared, as shown in Fig. 1.



**Fig. 1.** An illustration of global perception feedback convolutional neural networks. First, we compare the image gradient after the GP process against the Feedback [9], by using the same pre-trained GoogleNet trained on ImageNet 2012 classification dataset. Column (a) shows the input images. Column (b) and (c) show the Feedback results and GP results, respectively. Comparing against Feedback, the GP method filters out more local noise. Then, we demonstrate the more powerful discrimination of ResNet. Column (d) shows the GP results based on pre-trained ResNet. Comparing to GoogleNet, the ResNet has better results.

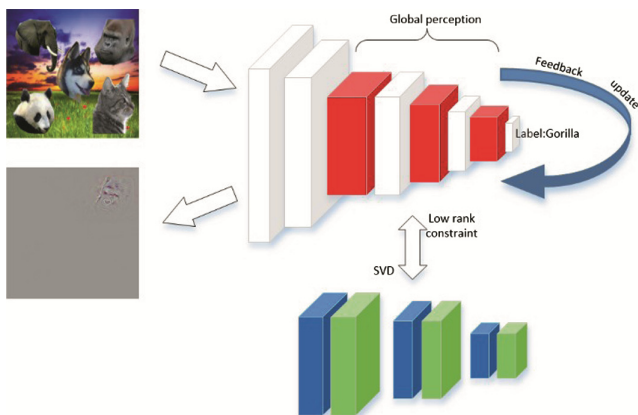
In practice, we maximize the score of the target class to suppress non-relevant neurons and minimize the trace norm of hidden neuron activations to maintain the low-rank structure of hidden layers. Subsequently, we use the gradient algorithm via back-propagation to update the status of hidden neuron activations.

The proposed method is evaluated on the ImageNet object localization dataset, with two widely used CNNs, i.e., GoogleNet [22] and ResNet [23]. We have demonstrated that our model can get better performance compared with previous feedback models.

## 2 Model

Current feedforward CNNs [24] mainly consist of convolutional layers, activation functions (such as ReLU) and pooling layers. Among them, ReLU and pooling layers play the role of “gates” [9], which filter out signals with minor contributions to final classification during the bottom-up propagation of input images. However, because these gates serve for all classes in the final fully-connected classification layer instead of a specific category, the activated neurons involve too many noises for a specific category.

In order to merely let the information of the target class pass through, [9] introduced a “feedback layer”, which is stacked upon each ReLU layer and consists of binary neuron activation variables  $z \in \{0, 1\}$ . The output of the feedback layer  $y$  is equal to the Hadamard product of the input  $x$  and binary variables  $z$ . The activation of these variables  $z$  are decided by top-down information passed from the target class label. However, this method just performs local perception during feedback and ignores the global properties. Our model adds an extra global perception on the foundation of [9], as shown in Fig. 2.



**Fig. 2.** Framework of our global perception feedback convolutional neural networks. First, given an input image with cluttered background and multiple objects, our networks perform in a bottom-up manner. Then, via a global perception, our networks inversely propagate the given label information and updates the status of hidden neuron activations in a top-down manner. Finally, we get a saliency map that includes class and location information corresponding to the given label.

## 2.1 Image-Specific Class Saliency Visualization

Given an image  $I$ , a class  $k$  and the hidden neuron activation status  $z$ , the class score of final fully-connected layer  $S_k(I, Z)$  is a highly non-linear function of  $I$ . Yet, by computing the first-order Taylor expansion, we can approximate  $S_k(I, z)$  in the neighborhood of  $I_0$ :

$$S_k(I, z) \approx G_k(z)^T I + b \quad (1)$$

where  $b$  is the bias term and  $G_k(z)$  is the derivative of  $S_k(I, z)$  with respect to the image at the point of  $I_0$  and  $z$ :

$$G_k(z) = \left. \frac{\partial S_k(I, z)}{\alpha I} \right|_{I_0, z} \quad (2)$$

The size of  $G_k(z)$  represents the relevancy between input pixels and relevant categories. Meanwhile, by the backpropagation method, the  $G_k(z)$  can be calculated and passed to pixel space to realize the visualization of  $G_k(z)$ . We also adopt the guided backpropagation method, which was proposed in [8]. The guided backpropagation method makes the visual images clearer by masking out the values corresponding to negative entries of the top gradient that prevents backward pass of negative gradients.

## 2.2 Optimization of Feedback Layers

The phenomenon of ‘‘Visual illusions’’ seriously affects the effect of top-down suppression. In consideration of the low-rank structure of attention map, we introduce a global perception method for the optimization of feedback layers.

Given image  $I$ , a pre-trained classification CNN trained on ImageNet dataset and a class  $k$ , we define activation variables  $z$  as  $z_{i,j,c}^l$  at every neuron  $(i, j)$  of channel  $c$ , on feedback layer  $l$ . Then we can define an optimization function in the following form:

$$\begin{aligned} \ell(z) &= S_k(I, z) \\ \text{s.t. } 0 &< z_{i,j,c}^l < 1, \forall l, i, c \end{aligned} \quad (3)$$

where  $S_k(I, z)$  is the score of the class  $k$ .

Since the core of top-down suppression is to eliminate the irrelevant background, we impose a global perception method to the optimization function:

$$\begin{aligned} \ell(z) &= S_k(I, z) - \lambda \|z\|_* \\ \text{s.t. } 0 &< z_{i,j,c}^l < 1, \forall l, i, c \end{aligned} \quad (4)$$

where  $\|z\|_*$  is the trace norm of  $z$ , which is used to enforce the low-rank of the feedback information.

Since the trace norm is difficult to directly optimize, we introduce an iterative minimization method for the trace norm [25].

Let  $z \in R^{i \times j}$  in the channel  $c$  of feedback layer  $l$ . The trace norm of  $z$  can be shown as

$$\|z\|_* = \sum_{n=1}^{\min(i,j)} \sigma_n \quad (5)$$

Where  $\sigma_n$  denotes the  $n$ -th singular value of  $z$ . The trace norm can also be represented as

$$\|z\|_* = \frac{1}{2} \inf_{g \geq 0} \text{tr}(z^T g^{-1} z) + \text{tr}(g) \quad (6)$$

The infimum is attained for  $g = (zz^T)^{1/2}$ .

By using this lemma, the previous optimization function Eq. (4) can be reformulated as

$$\begin{aligned} \ell(z) = S_k(I, z) - \frac{1}{2} \lambda \text{tr}(z^T g^{-1} z) - \frac{1}{2} \lambda \text{tr}(g) \\ \text{s.t. } 0 < z_{i,j,c}^T < 1, \forall l, i, c \end{aligned} \quad (7)$$

According to [25], the infimum over  $g$  is then attained for

$$g = (zz^T + \mu I)^{1/2} \quad (8)$$

In order to optimize the Eq. (7) in CNN, we use an alternating optimization method to update the parameters  $z$  and  $g$ . For the  $S_k(I, z)$ , we can calculate  $\partial S_k / \partial z$  by pre-trained CNN and back-propagation method, while the weights are fixed and parameters  $z$  are updated. For the matrix  $g$ , we update it via Eq. (8). For the trace norm, according to the Eq. (6), the derivation of  $z$  is equal to

$$\frac{\partial \|z\|_*}{\partial z} = g^{-1} z + (g^{-1})^T z \quad (9)$$

Hence, the gradient of the Eq. (4) is

$$\frac{\partial \ell(z)}{\partial z} = \frac{\partial S_k}{\partial z} - \frac{1}{2} \lambda g^{-1} z - \frac{1}{2} \lambda (g^{-1})^T z \quad (10)$$

The singular value decomposition of  $z$  is  $U \text{Diag}(\gamma_k) V^T$ . We get rid of minor constituent of  $\gamma_k$  and get  $\gamma'_k$ . Hence, the inverse of matrix  $g$  is

$$S^{-1} = V \text{Diag} \left( \frac{1}{\sqrt{\gamma'_k + \mu}} \right) U^T \quad (11)$$

We use the gradient ascent algorithm to update parameters  $z$  with the learning rate  $\alpha$ :

$$z^{t+1} = z^t + \alpha \left. \frac{\partial \ell(z)}{\partial z} \right|_{z^t} \quad (12)$$

### 3 Experiment

In this section, we evaluate the effectiveness of our GP feedback model. First, we compare the visualization results against the previous one [9] from qualitative perspective. Then, we conducted experiments of weakly supervised object localization on the ImageNet 2014 validation dataset from quantitative perspective. Every picture needs 10–50 iterations of suppression process, which is the same as [9]. Implementation details are included in our subsequent introduction.

#### 3.1 Qualitative Experiments

In this section, we compare the image gradient after the GP process against the previous one [9] on a set of images with multiple objects. Both of methods are given the same pre-trained GoogleNet with ground truth class labels. We also compare different visualization results between GoogleNet and ResNet. All results are shown in Fig. 1.

**Comparison of visualization methods.** We compare our global perception feedback method with the local perception feedback method [9] on a set of images with multiple objects. Both of methods are given the same pre-trained GoogleNet [22] with ground truth class labels. Without global perception, the status of hidden neurons are only suppressed by local perception and the visualization results are seriously disturbed by irrelevant background or objects, as shown in Fig. 1, column (b). Compared with local perception approach, our global perception effectively eliminates local noise, as shown in Fig. 1, column (c).

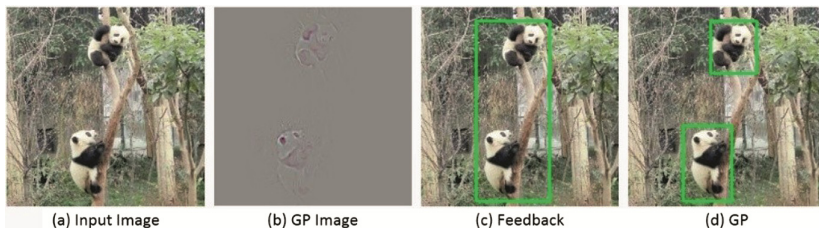
**Comparison of CNN classifiers.** Since [9] has demonstrated that GoogleNet has better feature extraction ability than AlexNet [26] and VggNet [24], we just consider these two popular CNN architectures: GoogleNet [22] and ResNet [23]. Both of them are downloaded from the Caffe Model Zoo [27]. We evaluate our GP feedback method on GoogleNet and ResNet respectively. The visualizing results are shown in Fig. 1. From visualizations, we find that ResNet better captures the salient map of target label than GoogleNet, suggesting that deeper networks have more powerful discrimination.

#### 3.2 Quantitative Experiments

In this section, we demonstrate the effectiveness of our GP feedback model on the ImageNet 2014 validation dataset, which contains ~50,000 images and corresponding class and position information. As shown in Fig. 1, given an image, our GP feedback model has the ability to determine the positions of the target objects. In the localization task, we get the category of an input image by bottom-up manner and get the bounding

boxes of the identified category by top-down manner. A bounding box is considered as correct if its overlap with the ground truth bounding box is over 50%.

Given an image and its corresponding saliency map, [9] merely calculates a tightest bounding box by simply thresholding to let the foreground area cover 95% energy out of the whole saliency map. This localization method will fail when there are multiple same objects, as shown in Fig. 3, column (c). Different from [9], we get every target object position by external contour detection and calculate every accurate bounding box, as shown in Fig. 3, column (d), which respectively identifies the position of two pandas.



**Fig. 3.** We select an example to demonstrate the effectiveness of our localization method. Column (a) shows the original image with two pandas. Column (b) shows the visualization result of GP. Because Feedback [9] merely calculates a tightest bounding box by simply thresholding to let the foreground area cover 95% energy out of the whole saliency map, the bounding box covers all pandas, as shown in column (c). We get every target object position by external contour detection and respectively calculate every accurate bounding box, as shown in column (d).

**Comparison of visualization methods.** We compare our global perception feedback method against the original gradient (GT) [7] and the guided backpropagation (GB) [8] and the local perception feedback method (FB) [9] on the ImageNet 2014 validation dataset. We first respectively use our external contour detection method and the localization method of [9] to conduct FB experiment. Our localization method obtains 61.2% localization error and outperforms the localization method of [9] (62.6%), suggesting that our localization method is better. Hence, all methods in Table 1 use our localization method instead of localization method of [9]. The results in Table 1 show that our global perception feedback method significantly outperforms other visualization methods, all on the GoogleNet architecture.

**Table 1.** Comparison of visualization methods.

	GT [7]	GB [8]	FB [9]	GP
Localization error (%)	65.9	64.8	61.2	59.6

**Comparison of CNN classifiers.** We also compare the weakly supervised localization accuracies of GoogleNet and ResNet in Table 2, based on our localization method. The results suggest that ResNet significantly outperforms GoogleNet, which agrees with the visualization results in Fig. 1.

**Table 2.** Comparison of CNN models.

	GoogleNet [22]	ResNet [23]
Localization error (%)	59.6	58.8

## 4 Conclusion

In this paper, we proposed a global perception model for feedback convolutional neural networks, which further eliminates irrelevant information by forcing the low-rank structure of the responses for hidden layer neurons during the feedback inference. Using GP, we get more discriminative saliency maps correspond to high level semantic labels. Good performance of the method has been demonstrated experimentally on the ImageNet 2014 object localization challenge with weakly supervised information.

## References

1. Chen, L.: Topological structure in visual perception. *Science* **218**(4573), 699–700 (1982)
2. Koch, C., Ullman, S.: Shifts in selective visual attention: towards the underlying neural circuitry. In: Vaina, L.M. (ed.) *Matters of Intelligence. Synthese Library (Studies in Epistemology, Logic, Methodology, and Philosophy of Science)*, vol. 188, pp. 115–141. Springer, Dordrecht (1987). [https://doi.org/10.1007/978-94-009-3833-5\\_5](https://doi.org/10.1007/978-94-009-3833-5_5)
3. Anderson, C.H., Van Essen, D.C.: Shifter circuits: a computational strategy for dynamic aspects of visual processing. *Proc. Natl. Acad. Sci.* **84**(17), 6297–6301 (1987)
4. Tsotsos, J.K., Culhane, S.M., Wai, W.Y.K., Lai, Y., Davis, N., Nuflo, F.: Modeling visual attention via selective tuning. *Artif. Intell.* **78**(1–2), 507–545 (1995)
5. Wolfe, J.M.: Guided search 2.0 a revised model of visual search. *Psychon. Bull. Rev.* **1**(2), 202–238 (1994)
6. Zeiler, M.D., Fergus, R.: Visualizing and understanding convolutional networks. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) *ECCV 2014. LNCS*, vol. 8689, pp. 818–833. Springer, Cham (2014). [https://doi.org/10.1007/978-3-319-10590-1\\_53](https://doi.org/10.1007/978-3-319-10590-1_53)
7. Simonyan, K., Vedaldi, A., Zisserman, A.: Deep inside convolutional networks: visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034* (2013)
8. Springenberg, J.T., Dosovitskiy, A., Brox, T., Riedmiller, M.: Striving for simplicity: the all convolutional net. *arXiv preprint arXiv:1412.6806* (2014)
9. Cao, C., Liu, X., Yang, Y., Yu, Y., Wang, J., Wang, Z., Huang, Y., Wang, L., Huang, C., Xu, W., et al.: Look and think twice: capturing top-down visual attention with feedback convolutional neural networks. In: *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2956–2964 (2015)
10. Desimone, R., Duncan, J.: Neural mechanisms of selective visual attention. *Annu. Rev. Neurosci.* **18**(1), 193–222 (1995)
11. Desimone, R.: Visual attention mediated by biased competition in extrastriate visual cortex. *Philos. Trans. R. Soc. B Biol. Sci.* **353**(1373), 1245 (1998)
12. Beck, D.M., Kastner, S.: Top-down and bottom-up mechanisms in biasing competition in the human brain. *Vis. Res.* **49**(10), 1154–1165 (2009)
13. Zhang, J., Lin, Z., Brandt, J., Shen, X., Sclaroff, S.: Top-down neural attention by excitation backprop. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) *ECCV 2016. LNCS*, vol. 9908, pp. 543–559. Springer, Cham (2016). [https://doi.org/10.1007/978-3-319-46493-0\\_33](https://doi.org/10.1007/978-3-319-46493-0_33)



14. Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., Torralba, A.: Learning deep features for discriminative localization. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2921–2929 (2016)
15. Hu, P., Ramanan, D.: Bottom-up and top-down reasoning with hierarchical rectified Gaussians. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5600–5609 (2016)
16. Coren, S., Girgus, J.S.: Visual illusions. In: Held, R., Leibowitz, H.W., Teuber, H.L. (eds.) Perception. Handbook of Sensory Physiology, vol. 8, pp. 549–568. Springer, Heidelberg (1978). [https://doi.org/10.1007/978-3-642-46354-9\\_16](https://doi.org/10.1007/978-3-642-46354-9_16)
17. Lin, S., Ji, R., Guo, X., Li, X., et al.: Towards convolutional neural networks compression via global error reconstruction. In: International Joint Conferences on Artificial Intelligence (2016)
18. Huang, R., Zhang, S., Li, T., He, R.: Beyond face rotation: global and local perception GAN for photorealistic and identity preserving frontal view synthesis. arXiv preprint [arXiv:1704.04086](https://arxiv.org/abs/1704.04086) (2017)
19. Wu, X., Song, L., He, R., Tan, T.: Coupled deep learning for heterogeneous face recognition. arXiv preprint [arXiv:1704.02450](https://arxiv.org/abs/1704.02450) (2017)
20. He, R., Tan, T., Wang, L.: Robust recovery of corrupted low-rank matrix by implicit regularizers. *IEEE Trans. Pattern Anal. Mach. Intell.* **36**(4), 770–783 (2014)
21. He, R., Sun, Z., Tan, T., Zheng, W.S.: Recovery of corrupted low-rank matrices via half-quadratic based nonconvex minimization. In: 2011 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2889–2896. IEEE (2011)
22. Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A.: Going deeper with convolutions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1–9 (2015)
23. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)
24. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint [arXiv:1409.1556](https://arxiv.org/abs/1409.1556) (2014)
25. Grave, E., Obozinski, G.R., Bach, F.R.: Trace lasso: a trace norm regularization for correlated designs. In: Advances in Neural Information Processing Systems, pp. 2187–2195 (2011)
26. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Advances in Neural Information Processing Systems, pp. 1097–1105 (2012)
27. Caffe Model Zoo. <https://github.com/BVLC/caffe/wiki/Model-Zoo>