# Deep saliency: What is learnt by a deep network about saliency?

**Sen He** [1]  **Nicolas Pugeault** [1]

## Abstract

Deep convolutional neural networks have achieved impressive performance on a broad range of problems, beating prior art on established benchmarks, but it often remains unclear what are the representations learnt by those systems and how they achieve such performance. This article examines the specific problem of saliency detection, where benchmarks are currently dominated by CNN-based approaches, and investigates the properties of the learnt representation by visualizing the artificial neurons' receptive fields. We demonstrate that fine tuning a pre-trained network on the saliency detection task lead to a profound transformation of the network's deeper layers. Moreover we argue that this transformation leads to the emergence of receptive fields conceptually similar to the centre-surround filters hypothesized by early research on visual saliency.

## 1. Introduction

Deep convolutional neural networks have achieved great success in dealing with computer vision problems, such as image classification (Krizhevsky et al., 2012), object detection (Girshick et al., 2014) and semantic segmentation (Girshick et al., 2014), etc. However, deep convolutional neural networks are complex non-linear system, and what is learnt by intermediate layers remain in most case mysterious. In addition, despite high performance on many benchmarks, recent published research has demonstrated that despite high performance on benchmark measures, deep networks could be easily fooled by small perturbations of the original signal (Moosavi-Dezfooli et al., 2016), begging the question what are the representations learnt by the networks and how they are used to answer the chosen task (ie,

do we recognise a bunny by its ears or by the texture of its fur?). For this reason, it becomes increasingly important for scientists to investigate what is learnt by such networks and what features deep artificial neurons are attuned to, in a way not dissimilar to what neuroscientists did for the human visual cortex.
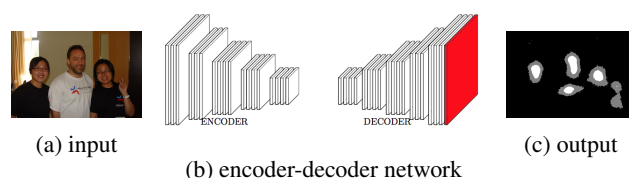


(a) input
(b) encoder-decoder network
(c) output

*Figure 1.1.* The architecture (the encoder part is fine tuning from VGG-19 convolutional part) we developed for saliency prediction, which is competitive with the state-of-the-art on MIT300 saliency benchmark (Bylinskii et al., 2015).

In this article, we are concerned with the task of predicting image saliency. Saliency can be defined as how likely a visual pattern is to attract a human viewer's gaze when observing the image. Visual saliency has been the subject of intense study over the last decades, both in psychology and computer vision (Borji & Itti, 2013), and recent publications have demonstrated that deep neural networks can achieve very high performance on this task (Bylinskii et al., 2016). We will use a recently developed architecture (see Figure 1.1) for saliency detection based on a standard CNN encoding (so-called VGG19 (Simonyan & Zisserman, 2014)), and visualise the receptive fields of the artificial neurons before and after fine-tuning (the CNN encoder is pre-trained on a standard ImageNet classification task). We demonstrate that after fine-tuning the network for the task, the deeper neurons have evolved vastly different receptive fields to the pre-trained neurons, and display characteristic patterns that evoke the centre-surround difference paradigm hypothesized by early psychophysical research (Treisman & Gelade, 1980).

## 2. Related Work

With the rise in popularity of deep convolutional neural networks, several groups have recently attempted to visualise which features a network has learned. Zeiler & Fergus (2014) proposed to back-propagate feature maps obtained
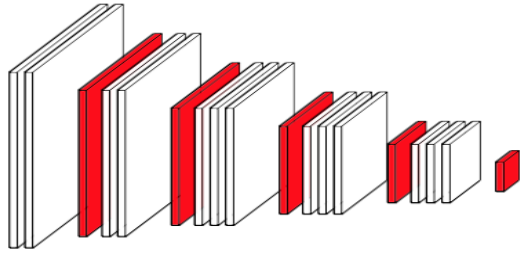
---

[1]Department of Computer Science, College of Engineering, Mathematics and Physical Sciences, University of Exeter, UK.. Correspondence to: Sen He <sh752@exeter.ac.uk>, Nicolas Pugeault <N.Pugeault@exeter.ac.uk>.

*Figure 1.2.* The model used in this paper, we try to visualise the general patterns of the encoder part in Figure 1.1 that would activate the red pooling layers

by processing a specific image through the network, in order to visualise the image content that activated the feature maps. Yosinski et al. (2015) follow a similar approach to develop a tool for deep visualisation, and additionally proposed an approach to visualise features of each layer via regularised optimisation in image space. Nguyen et al. (2015), they found that the deep neural networks are easily fooled, and use evolutionary algorithm and gradient ascent method to derive a pattern that the network has a high confidence to determine the derived pattern is belong to a specific class.

In contrast, we manually clamp the value of single neurons selected from intermediate layers in the network, and back-propagate the activation to the image space, thus deriving the optimal activation pattern for individual selected neurons. Hence, this visualisation provides us with an understanding of what patterns the deep representations have become attuned to.

## 3. Methods

In this work we will be concerned with visualising the input patterns most strongly related to individual neurons in the network. In the following we will call these patterns as the neurons' *receptive fields*, in analogy to biological neurons.

In deep convolutional neural networks, the forward pass typically consists of three main processes: convolution, non-linearity (usually a ReLU function) and pooling. Similarly, the visualisation of neural patterns are produced by three similar processes, in reverse order: that is, upsampling, deconvolution, and non-linearity (again, ReLU). We will describe the three processes in turn.

### 3.1. upsampling

The purpose of upsampling is to recover the gradually reduced resolution caused by pooling in the forward pass.

The classic upsampling method in feature visualisation is unpooling, using the pooling indices in the forward pass to do unpooling—see Figure 3.1.
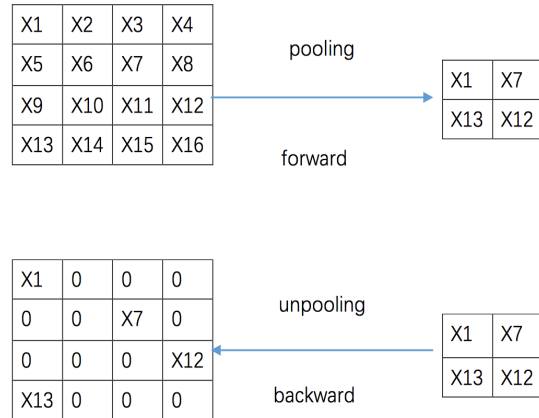


*Figure 3.1.* classic upsampling method

Because pooling indices only exist when processing an actual image through the network, these indices are not available when visualizing a neuron's receptive field in abstraction from any input. Hence, in order to visualize individual neuron's receptive fields, we set the pooled feature map as a sparse matrix (with only one non-zero value) and do upsampling by repeating this sparse matrix—see Figure 3.2 .
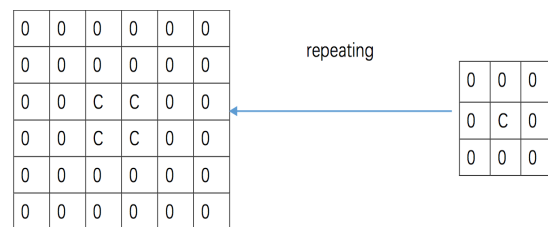


*Figure 3.2.* upsampling by repeating the sparse feature map, c is a constant

### 3.2. Deconvolution

Convolution is the key process in the forward pass of the convolutional neural network, as it is the one that is tuned by back-propagation during training. It can be formulated

as:

$$O = I * F \qquad (1)$$

where $O$ is the extracted feature map, $I$ is the input and $F$ is the learnt filter. Reconstructing the input pattern $I$ which activated an extracted feature map $O$, can be formulated as follows:

$$A = O * F^T \qquad (2)$$

where $O$ is the extracted feature map in the forward pass, $F^T$ is the transpose of the learned filter and $A$ is the content in the input $I$ which activated $O$.

### 3.3. Relu

the ReLU function in feature visualisation is the same as that in the forward pass of deep convolutional network, which only leave the positive components of the input, and can be formulated as:

$$f(x) = \max(0, x) \qquad (3)$$

## 4. pattern visualisation

In this part, we show the general patterns learnt by fine-tuning the network on a saliency prediction task, as well as the patterns for the original VGG-19 network, pre-trained on classification on ImageNet. Additionally, we also visualise individual neurons' receptive field by clamping them and back-projecting to the input domain as described above.

### 4.1. VGG pattern and salient pattern

Figures 4.1 to 4.5 contrast the receptive fields learnt by neurons in various layers of the network, both before and after fine-tuning on the saliency detection task.
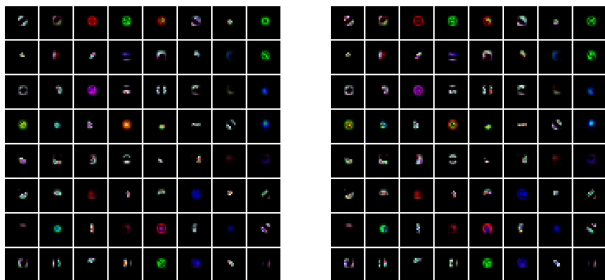


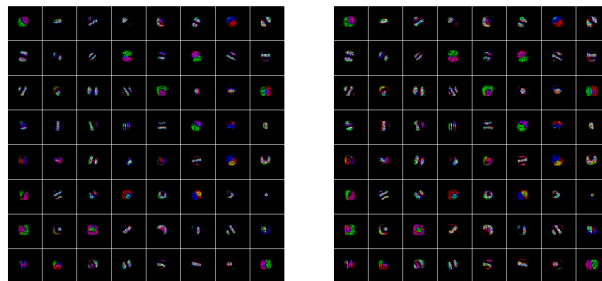*Figure 4.1.* the 64 general salient(left) and vgg(right) patterns for the first pooling layer



*Figure 4.2.* the first 64 (128 in total)general salient(left) and vgg(right) patterns for the second pooling layer
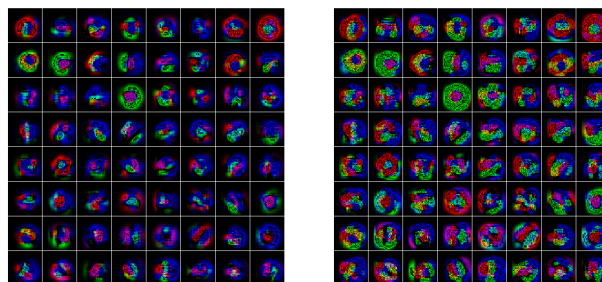


*Figure 4.3.* the first 64 (256 in total) general salient(left) and vgg(right) patterns for the third pooling layer
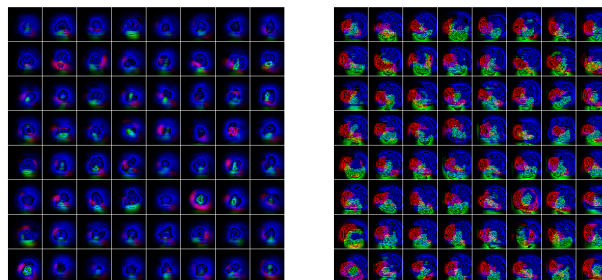


*Figure 4.4.* the first 64 (512 in total)general salient(left) and vgg(right) patterns for the fourth pooling layer
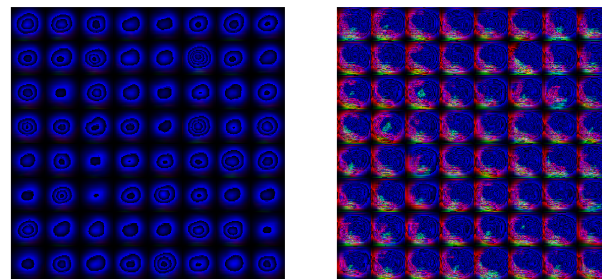


*Figure 4.5.* the first 64 (512 in total)general salient(left) and vgg(right) patterns for the fifth pooling layer

In those figures, we can see little differences between the neurons' receptive fields in the first three pooling layers af-

ter fine-tuning. Some patterns are the same as the edge pattern. However, when considering deeper layers, fundamentally different patterns arise in the neurons' receptive field after fine-tuning for the saliency task: after fine-tuning the deep neurons all appear to have attuned to variations of central-surround patterns. Interestingly, such patterns emerge solely through the process of fine-tuning the network, starting from vastly different receptive fields, and they appear to be consistent with theoretical and experimental research on saliency by psychologists.

### 4.2. Pattern Propagation

In a second experiment, we illustrate how the different patterns yield different levels of activation to specific deep neurons. This is achieved by clamping a neuron-specific output to a constant value from 0.5 to 3. These results are recorded in Figure 4.6. From the figure above, when in-
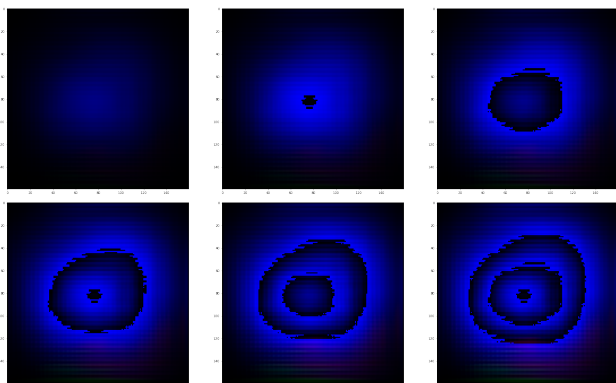


*Figure 4.6.* the pattern propagation by increasing the constant value (from left to right,up to down)

creasing the neuron's output value its receptive field shows patterns that propagates like the water wave propagation. We argue that the constant in the sparse matrix is the energy of the pattern, the higher energy of a pattern, the wider it will propagate.

### 4.3. Pattern Validating

In the previous sections, we have shown that the proposed visualisation strategy can be used to illustrate the patterns learned by deep neurons in a network. In this section we verify that those back-propagated patterns actually activate the selected neuron. An additional question is how such patterns affect other neurons in the same layer. This is tested in a straightforward manner by feeding the back-propagated pattern as an input to the network. Note that due to the pooling process in the forward pass, the resulted pooling feature map may not the same as the clamped sparse matrix used to generate the pattern. We check the activation by the summation of the pooled feature map—
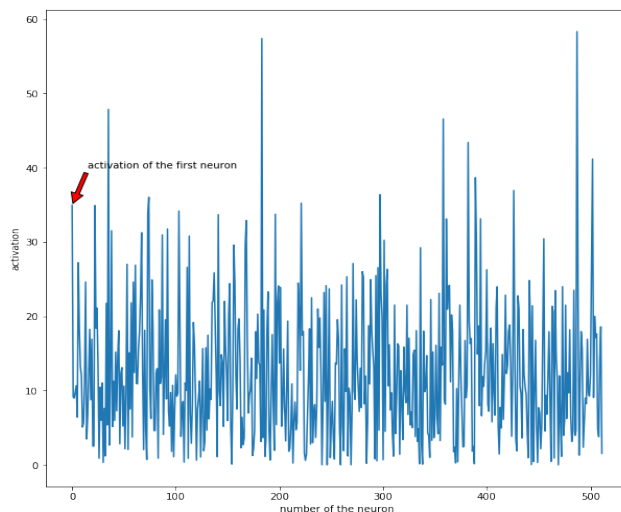
see Figure 4.7.



*Figure 4.7.* the activation of the first general pattern to the fifth pooling layers(512 neurons)

From this figure, we can see that the first pattern indeed activate the first neuron, as expected. Furthermore, some other neurons have higher activation, demonstrating that the network has developed some redundancy in its coding, whereas other neurons are inhibited. We argue that this is because most of the learned general patterns are similar (central surround difference), some neurons are more sensitive to the central surround difference pattern; and for those neuron that are inhibited is due to the lateral inhibition (Ratliff et al., 1967) which is also used in local response normalisation (Krizhevsky et al., 2012) when training the deep neural network.

## 5. Conclusion

In this article we proposed a novel approach for visualising the representations learnt by deep neural networks, and specifically to visualise the receptive fields of individual deep neurons. We have demonstrated this approach to a VGG-19 network pre-trained on ImageNet classification and fine-tuned for the task of saliency detection. Importantly, we demonstrate that this approach can reveal important insights in what is learnt by the network to achieve its high performance: receptive fields are shown to change drastically from the original VGG-19 representation to characteristic centre-surround patterns. Interestingly, these emergent patterns are consistent with the psychological theories of saliency. This demonstrates that such a visualisation offers an important tool for interpreting the workings of deep neural networks.

To sum up, by manually set the feature map as a sparse ma-

trix, we derive a set of general patterns for the deep neural network. We also double check the resulted general patterns by forwarding it into the network, which show those general patterns are not illogical and follow the evidence in neurobiology.

## 6. Acknowledgements

## References

Borji, Ali and Itti, Laurent. State-of-the-art in visual attention modeling. *IEEE transactions on pattern analysis and machine intelligence*, 35(1):185–207, 2013.

Bylinskii, Zoya, Judd, Tilke, Borji, Ali, Itti, Laurent, Durand, Frédo, Oliva, Aude, and Torralba, Antonio. Mit saliency benchmark, 2015.

Bylinskii, Zoya, Recasens, Adrià, Borji, Ali, Oliva, Aude, Torralba, Antonio, and Durand, Frédo. Where should saliency models look next? In *European Conference on Computer Vision*, pp. 809–824. Springer, 2016.

Girshick, Ross, Donahue, Jeff, Darrell, Trevor, and Malik, Jitendra. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 580–587, 2014.

Krizhevsky, Alex, Sutskever, Ilya, and Hinton, Geoffrey E. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pp. 1097–1105, 2012.

Moosavi-Dezfooli, Seyed-Mohsen, Fawzi, Alhussein, Fawzi, Omar, and Frossard, Pascal. Universal adversarial perturbations. *arXiv preprint arXiv:1610.08401*, 2016.

Nguyen, Anh, Yosinski, Jason, and Clune, Jeff. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 427–436, 2015.

Ratliff, Floyd, Knight, BW, Toyoda, Jun-ichi, and Hartline, HK. Enhancement of flicker by lateral inhibition. *Science*, 158(3799):392–393, 1967.

Simonyan, Karen and Zisserman, Andrew. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

Treisman, Anne M and Gelade, Garry. A feature-integration theory of attention. *Cognitive psychology*, 12 (1):97–136, 1980.

Yosinski, Jason, Clune, Jeff, Nguyen, Anh, Fuchs, Thomas, and Lipson, Hod. Understanding neural networks through deep visualization. *arXiv preprint arXiv:1506.06579*, 2015.

Zeiler, Matthew D and Fergus, Rob. Visualizing and understanding convolutional networks. In *European conference on computer vision*, pp. 818–833. Springer, 2014.