# 

[**9.3: Visualising CNNs**](#_gqktgwusys14) **2**

[9.3.1: Receptive field of a neuron](#_nfl7zny9qcnr) 2

[9.3.2: Identifying images which cause certain neurons to fire](#_qww6kjo92kth) 4

[9.3.3: Visualising filters](#_ha9jwydcvvll) 5

[9.3.4: Occlusion experiments](#_ystnhil3ywxu) 7

# 

# Visualising CNNs

## Receptive field of a neuron

How does the receptive field increase across layers?

1. What does a filter learn? What kind of images cause certain neurons to fire? How good are the hidden representations?
2. To answer these questions, we need to talk about the receptive field.
3. Consider a 3 layered CNN, where in each of the layers, we apply a 3x3 filter
4. The **Receptive Field or Region of influence** of the neuron/pixel in a subsequent layer refers to all the neurons/pixels from the previous that were involved in the convolution operation to produce said output neuron/pixel.
5. Let’s look at layer 3, and express the receptive field in terms of layer 1.
6. Thus, we can see that the 3x3 grid in the receptive field of the layer 3 pixel translates to a 5x5 grid receptive field in the input image.
7. We can see a sort of pattern, whereby if we have more layers, the receptive field becomes larger and larger in the input image.
8. If there was another Layer 4 with the same 3x3 filter, it would correspond to a 7x7 receptive field in the input layer.
9. The following diagram illustrates how the receptive field in the input layer is larger and larger as we move down the layers.
10. The increase in receptive field size through the layers increases with a larger filter size.

## 

## Identifying images which cause certain neurons to fire

How do you check what is causing a neuron to fire?

1. From the previous section, why is the concept of an increasing receptive field important?
2. Consider the same diagram from the before
3. Now, what does Firing mean? In the case of sigmoid/tanh neurons, it refers to when the output approaches 1. In the case of ReLU and leaky ReLU, it refers to any large positive value, greater than some threshold.
4. Here are the results of an experiment to understand the firing of different neurons
5. In the above image, the firing of 6 different neurons was observed across the training set of 1000 images. Each neuron is shown to be sensitive to particular inputs as shown in the figure above.
6. This analysis helps us to understand whether neurons in different layers are learning some meaningful patterns, and having some discriminatory power, instead of all of them firing for all types of inputs..

## 

## Visualising filters

What does a filter capture?

1. We have dealt with filters in all our CNN models so far. Now the question is, what exactly does a filter capture?
2. Let’s look at the working of a 2x2 filter on a 4x4 input image
   1. Here, the input image is 4x4 while the fiter is 2x2
   2. The red input pixel vector
   3. The weight vector
   4. By convolving the input pixels with the filter, we get the output
   5. Output
   6. (This is the same as the dot product between the two vectors)
   8. Now for certain inputs, we want the filter to fire (give a high value).
   9. Now, will be high when is high (cos(θ) = 1), i.e. when θ is 0. This implies the two vectors w and x are in the same direction.
   10. So, we can say that an input vector which aligns with a filter vector yields maximum output.
   11. The neuron will fire maximally when (x is a unit vector in the direction of w)
   12. Thus, when we **slide the** **2x2 filter w** across the 4x4 input region, whenever we **reach a 2x2 region x** that looks exactly like the filter, we get a high output. For all other regions which do not align with the filter, the output is low.
3. Now, let us visualize the filters in AlexNet
   1. The above image shows us how different patterns are identified by different filters.

## 

## Occlusion experiments

Which patches in the image contribute maximally to the output?

1. We can determine the maximally contributing regions in an image using occlusion experiments
2. Occlusion experiments refer to applying occlusions to the input image and seeing how they alter the predicted output distribution.
3. Then we take the difference between the un-occluded distribution and occluded distributions for occlusions placed in all patches of the images.
4. These differences allow us to generate a heatmap of maximally contributing patches.
5. Consider the following images with occlusions applied to them
   1. The true distribution for any of these images would be say
   2. The un-occluded predicted distribution
   3. Occlusion placed at position 1

...

* 1. Occlusion placed at position L

1. Now, calculating the difference between , we get differences in probilites which can be plotted as a heatmap over the input image with colder regions being more highly contributing.
2. Thus, the **higher the difference between un-occluded and occluded distributions**, the **more significant the contribution of that particular occluded region** to the output.