Deep Learning
Prof. Mitesh M. Khapra
Department of Computer Science and Engineering
Indian Institute of Technology, Madras

Lecture - 94
Visualizing filters of a CNN

So, now, this was Visualizing the Neurons inside the Convolutional Neural Network. So, neurons remember are the outputs right, these are not these are the feature maps. What about the weights itself, What are the weights in a convolutional neural network?

Student: (Refer Time: 00:25).

The filters; the filters themselves are which have you ever tried to visualize weights before, when?

Student: Auto encoders.

Auto encoders and what was a trick there, how did we?

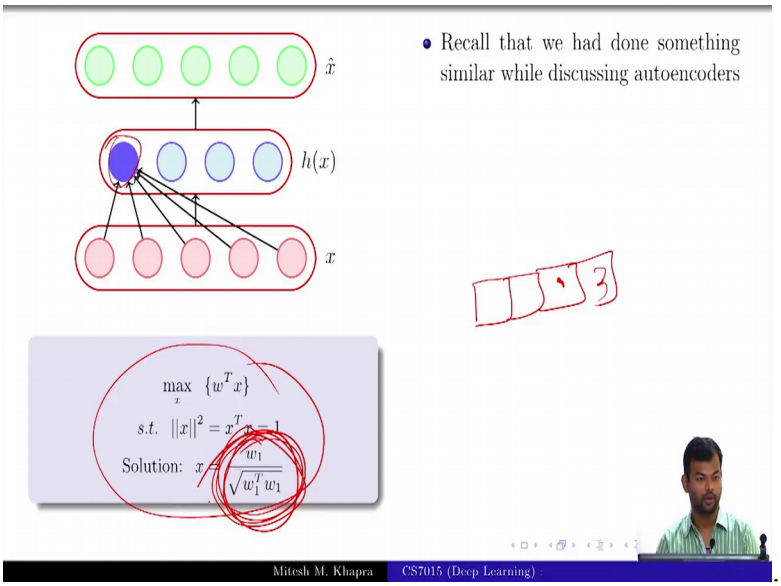
Student: (Refer Time: 00:36).

Visualize what was the optimization problem that we solved?

Student: (Refer Time: 00:40).

How many of you went and looked at the prerequisites? How many if you looked at the prerequisites? ok.

(Refer Slide Time: 00:46)



So, we had done something similar while discussing auto encoders. So, because that we had done something similar while discussing auto encoders right. So, we were interested in knowing that there is a particular hidden neuron inside the auto encoder and we wanted to see that; what does this neuron capture? So, if I give it emnist digits then what kind of patterns does it fire for and if you remember we had solved this optimization problem and realize that, this neuron will fire for an input, which looks like this where w 1 or all the weights which are connecting to this neuron ok. What was the dimension of the input if you are dealing with emnist digits?

Student: (Refer Time: 01:22).

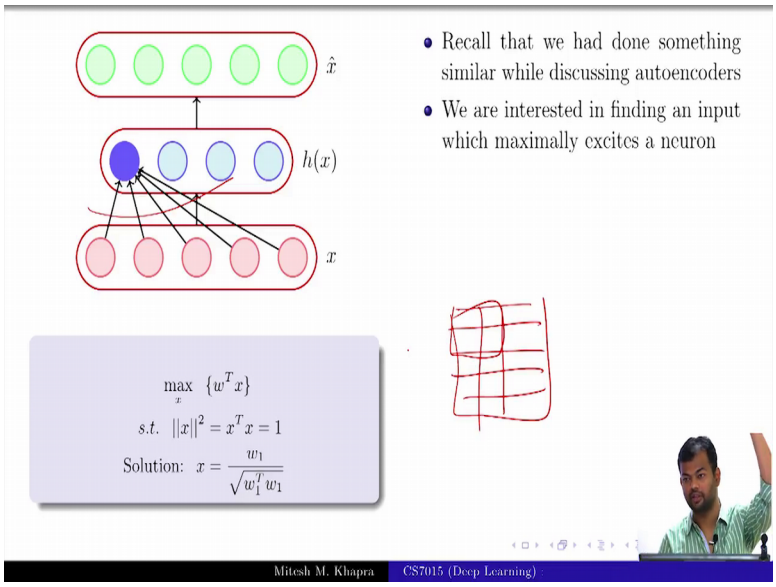
784; what is the dimension of this a one thing which I have circled here?

Student: (Refer Time: 01:27).

784 right; it is written x equal to so, it has to be 784. Why is it 784 because there are 784 weights connecting each of the input pixels to that neuron right so; that means, this weight matrix itself we can visualize it as an image and that is exactly what we had done if you remember we had this grid of images that we were analyzing and in some images we saw that some dark element fires here and each we were arguing that this is the curve which exists in 2 or 9 or 8 and that is the one which is capturing.

And in some cases there was a cusp here which was firing and we were arguing that this could be for the 3 or for a 9 or for a 8 or something like that right. So, we were trying to visualize these things and the way we had plotted it was just treating this weight matrix or weight vector as an image and seeing what causes the neuron to fire right.

(Refer Slide Time: 02:21)



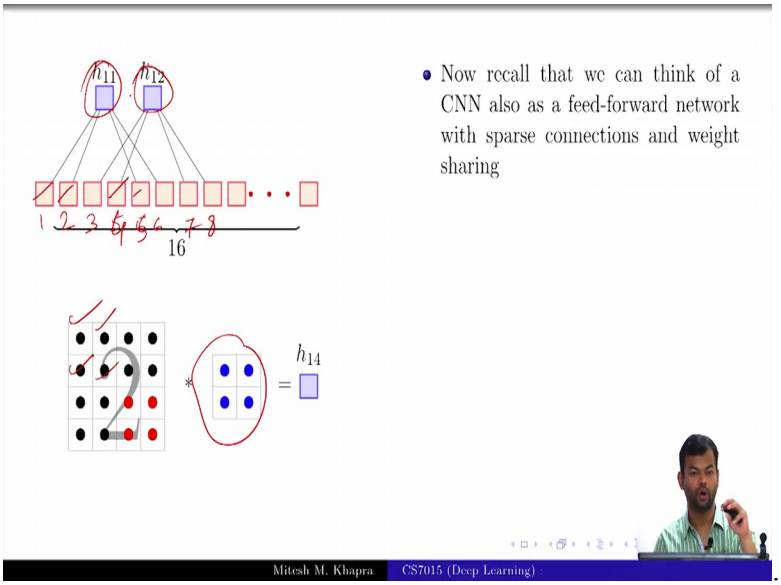
So, we can do something similar for convolutional neural networks. I want you to think how would you do that I will give you some hints; the answer is there on the next slide, but I just want you to think about it right. So, remember here you have dense connections ok; that means your weight vector was the same dimension as the input vector. What about filters in the case of CNN? They are smaller they are 3 cross 3, 5 cross 5 or 7 cross 7 much much smaller than your original image.

So, then what do these filters correspond to? Just think of the animation that we had seen right we had this image and we were taking a filter and applying it at different places. So, what does the filter correspond to? What is the filter overlap with; patches in the image right. So, now, what kind of analysis can you do?

Student: Dense.

What kind of patches does this filter fire for or what kind of patches does the neuron connected to this filter fire for? Does that make sense everyone gets the intuition? How many if you get the intuition please raise your hands; thank you.

(Refer Slide Time: 03:23)

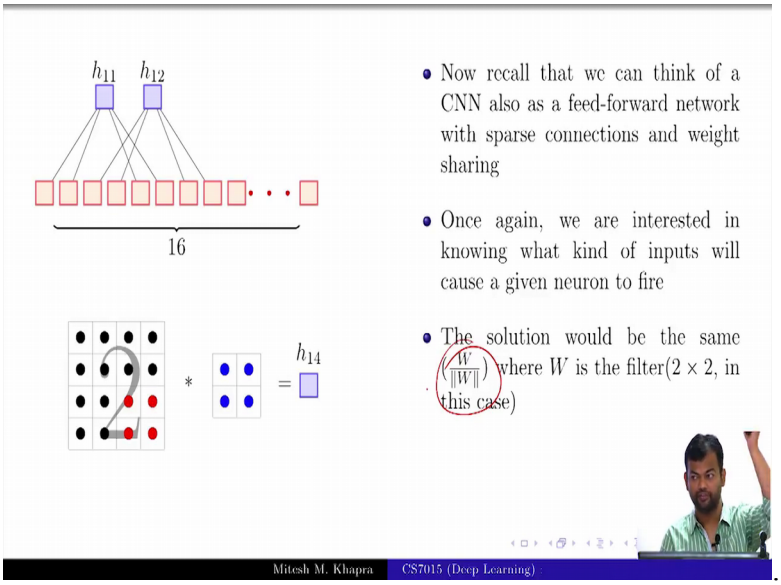


So, now, recall that we can think of a CNN as a feed forward neural network and in particular when you have a filter it actually interacts only with few pixels right. So, interacts with say pixel 1 2 5 and 6. So, that is the patch that it interacts with.

And now I want to see when does this neuron fire. So, that is the same as asking what do I put in 1, 2, 5, 6 for this neuron to fire or similarly what do I put in 3 I do not know this was 1, 2, 5, 6 I guess. So, 3, 4, 7, 8 for the same different neuron to fire right, but all these neurons fire because they are connected to the same filter.

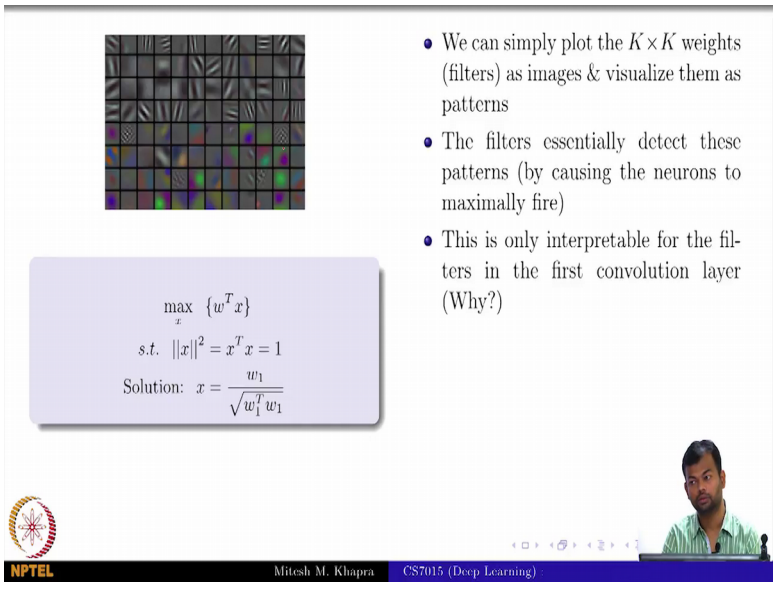
So; that means, I am interested in these patches, which will cause the neuron to fire and those patches can appear anywhere in their image, is that fine that is the whole point of convolution neural networks; wherever there is a nose whether it is at the top corner of the image or the center or the bottom it should be able to detect right that is the whole point of weight sharing and sparse connectivity ok.

(Refer Slide Time: 04:23)



So, we are going to do exactly the same thing. We will have a 3 cross 3 filters or 5 cross 5 filters or 7 cross 7 filters were just going to visualize as them as images, but unlike the earlier case where the image actually correspond to the full emnist image here these images are just corresponding to those 3 cross 3 or 5 cross 5 patches and you want to see what kind of patches causes the neurons to fire ok, and the solution is still the same we will have this W by W the normalized weight filter weight, which is causing the input to fire. How many if you are fine with everything at this point please raise your hands high up.

(Refer Slide Time: 04:55)



So, this is what we get right and we observe certain things which like we had earlier made a case for that these filters try to detect certain types of patterns or textures or edges. So, you can see that right this is capturing these slanting edges this is trying to capture some horizontal sorry vertical edges then some edges oriented differently and also some colored, patterns some texture right. So, you see something like a checkbox here or chess box here and so on.

So, these filters are actually firing for different kinds of patches. So, they are trying to detect different things from the images. So, you could visualize this and unless you see a lot of variety in this; that means, something is wrong right because your filters are not being trained to be discriminative with terms of different patterns that they can detect and so, on right. So, you want these variety of patterns to occur ok, and I am going to make a claim that this is only interpretable for the first layers in the convolutional neural network why is it so, I am seeing some half complete answers. So, I will ask this as a quiz question.