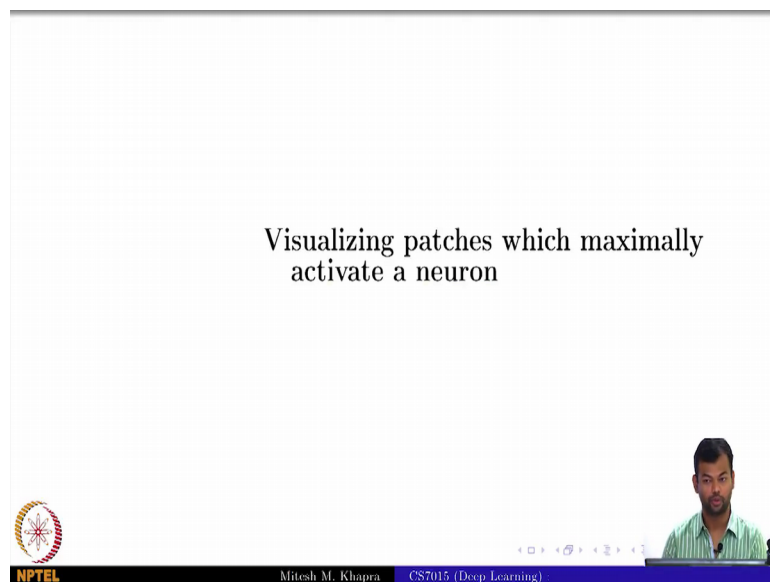


Deep Learning
Prof. Mitesh M. Khapra
Department of Computer Science and Engineering
Indian Institute of Technology, Madras

Lecture – 12
Visualizing Convolutional Neural Networks, Guided Backpropagation, Deep Dream, Deep Art, Fooling Convolutional Neural Networks

So in this lecture we will look at various ways of Visualizing Convolutional Neural Networks and although it is not very obvious at this point as we go along we will see what I mean by that. So, let us start this lecture.

(Refer Slide Time: 00:29)

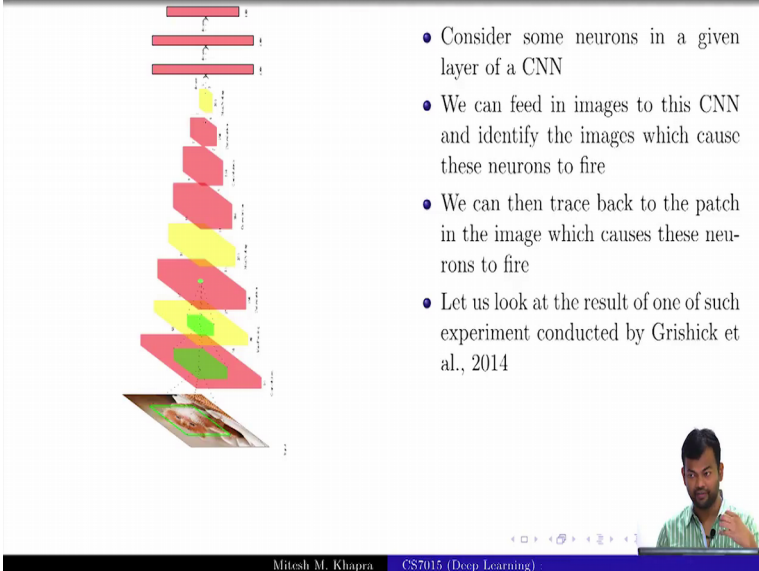


So, I forgot to add the acknowledgments slide. So, a lot of the material that I am going to cover today is based on some content by Andrey Karpaty in his online course or the Stanford course we will add the appropriate acknowledgments and a link to the course ok.

So, with that I will start module 1, which is visualizing patches which maximally activate a neuron ok. So, what are we trying to do here is we are trying to the quest today largely is going to be able to understand what a CNN has actually learned right and what I mean by that is we said that there are these filters, which try to detect edges, which try to detect blurs and so on and then there are these neurons which fire for certain things and so, on.

So, we want to see different ways of finding out what a convolution neural network has actually learned or what have the filters actually learned or what are the different neurons in the convolutional neural network actually capturing? What do they fire for what are the kind of images that make them trigger and so, on right. So, that is the first thing that we are going to look at how do you visualize patches which are causing a neuron to fire.

(Refer Slide Time: 01:32)



- Consider some neurons in a given layer of a CNN
- We can feed in images to this CNN and identify the images which cause these neurons to fire
- We can then trace back to the patch in the image which causes these neurons to fire
- Let us look at the result of one of such experiment conducted by Grishick et al., 2014

So, this is again our VGG network just put it vertically, say have passed an image to that, and then at every layer you are applying convolutions and then max pooling and so on, right up to the last layer right. Now we consider some neurons in one of these layers. So, I am considering this neuron and I want to find out what exactly is this neuron trying to do right and which is the same as asking what kind of images does this neuron fire for.

So, I have thousand different classes I have cats, dogs, cars, trucks and so on. I am interested in figuring out what are the different kinds of classes that this neuron fires. And this is more from say I am already getting some output accuracy and I am either happy with it or not happy with it in either case I just want to see what is it that my network is learning is there any scope for improving. Is that that there are no neurons in the network which actually fire for the dog class, did not should I do something differently was it that most of the neurons fire for all classes; that means, they do not have any discriminative power. So, what exactly is going on right?

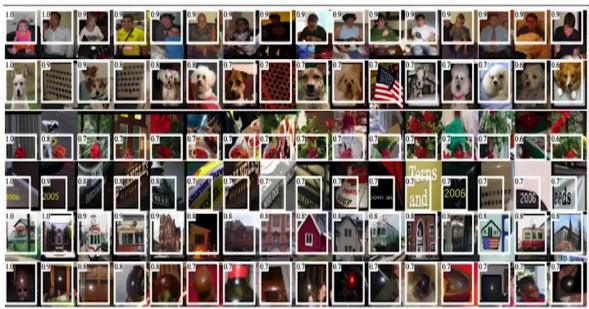
So, that is why we are that is why this study is interesting and you will do something of this sort in your CNN assignment ok. So, and by now we are clear that if I am focusing on any neuron and any layer, I can always go back and trace the patch to which it corresponds in the input image everyone is fine with that right. So, we saw that if I am somewhere here, then every neuron here corresponds to some 16 cross 16 patch in the original image and the same is true for every layer right. I can always this is a deterministic process, I can just find out which are the original image pixels which contributed to the computation of this particular neuron in any layer ok.

So, I can do that. So, now, what I am going to do is, I will send as many images as possible, whatever images are there in my training data, test data whatever images I have, I will pass these images through the convolutional neural network ok, and for the neuron of interest I will note down, which when does it fire and where ever it fires and by fire I mean it is a output is close to one or it is a output is high because these are ReLU neurons, I look for high output they do not saturate at one right. So, this I look which images for which this neuron had an high output and for those cases I will go back and trace the image and see which patch of the image actually caused this to fire.

So, I want to see whether my neurons are actually learning things like noise detector or eye detector or something right.

(Refer Slide Time: 04:01)

- They consider 6 neurons in the pool5 layer and find the image patches which cause these neurons to fire
- Another neuron fires for shiny surfaces



The slide displays a grid of 60 small image patches, arranged in 6 rows and 10 columns. These patches represent the input images that cause specific neurons in the pool5 layer to fire. The patches show a variety of objects and textures, including faces, animals, buildings, and abstract patterns. The grid is labeled with 'L5' in the top-left corner of each row. Below the grid, there is a small circular logo with a red and white design. At the bottom of the slide, there is a footer with the NPTEL logo, the name 'Mitesh M. Khapra', and the course title 'CS7015 (Deep Learning)'. The slide number '4/51' is also visible in the bottom right corner.

So, let us look at the results of one such experiment done by a group of researchers. So, they considered some neurons in the pool 5 layer and they did this experiment that they pass a lot of images and whenever this neuron fired they went back and saw what was the patch in the image, which was causing this neuron to fire.

So, that they found that one set of neurons is actually fires for people places. So, if you go back and trace which is the image, which caused is to fire or which is the patch, then it is largely centered around a persons face or which is something which is very clearly a person ok. Another set of neurons fires for dogs, another set of neurons fires for flowers all sorts of flowers and different orientations different maybe colors are same here, but they are all different thing right somewhere inside a bouquet somewhere inside a flower pot some somewhere on a table and so, on, but expected of that these neurons are firing for any flowers that appear in your input image and the fire only for that patch nothing around it ok.

So, it is very is actually able to localize and fire. There are some images which fire for this images the digits and alphabets written in the image. So, these are some addresses or dates or billboard signs or something like that and whenever there are these characters or numerals there and this neurons fire.

And some neurons fire for houses and then some neurons fire for shiny surfaces. So, there is this different sets of neurons which fire for different sets of things right. So, also; that means, your convolutional neural network is trying to learn specific characters of the input characteristics of the input and this is one way of visualizing. So, this is not like anything tricky here it is just that its good you can think of this as debugging tools for your convolutional neural network right, because in your you I guys are used to programming where you give different inputs and see what is the output and then try to debug it.

So, this is one way of trying to figure out whether your network has learned does it really need more training is there a certain class of images for which it is not firing at all or is it confusing between two classes and so, right. So, that is one way of visualizing.