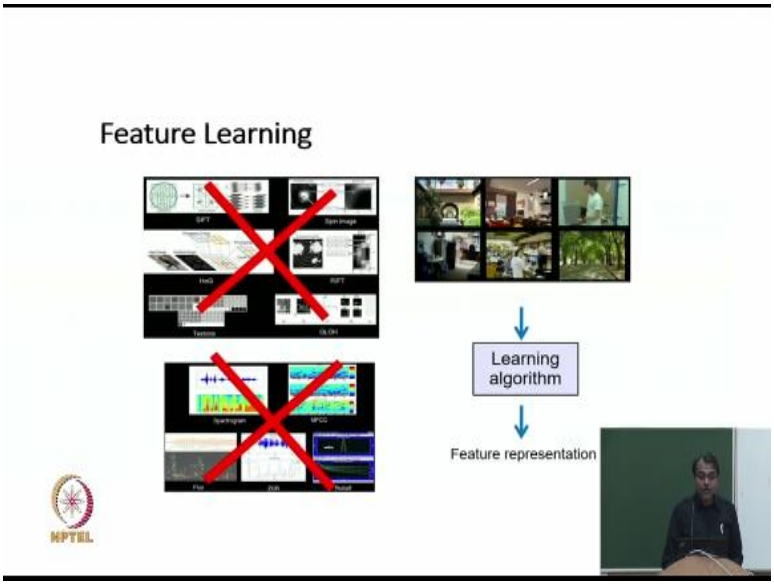
Artificial Intelligence
Prof. Mausam
Department of Computer Science and Engineering
Indian Institute of Technology-Delhi

Lecture - 90
Deep Learning: Convolution Neural Networks

Let us do five minutes of convolution neural networks. There is no such thing as five minutes of CNN. But I just want to introduce this to you, just to give a sense. Because CNNs were the first deep learning models that became super successful. They were for image classification. We can think about why they would be so good by a simple converted into a bit vector and then a fully connected neural network will not be the best idea, okay?

(Refer Slide Time: 00:48)



Again, the old style model was give me the image, I will convert it into a set of features. Give me the audio I will convert it into a set of features that does not happen anymore. I you give raw images and that with the learning algorithm makes the feature representation. Now here is the question for you. Suppose I am giving you the 16 cross 16 bit vector with state a. Let us take an easy one which says 1, right.

Now this 1 could be right in the middle very vertical. This 1 could be slightly slanted still in the middle. This 1 could be slightly shifted, but vertical, right. We need to make those predictions all of them correctly. Whether it is right, straight here, whether it is slightly slanted, whether it is slightly shifted, whether it is suddenly shifted like

this, whether it is this kind of a one, whether it is you know there are different kinds of ones as well.

So technically, there is a universal approximation theorem, technically even a 2d I mean, a 2 layer neural network can do this, technically a fully connected neural network can do this. But let us figure out at least some parts of the structure of the problem. Let us figure out, you know what can we say about our problem at hand, which allows us to make some improvements in the architecture.

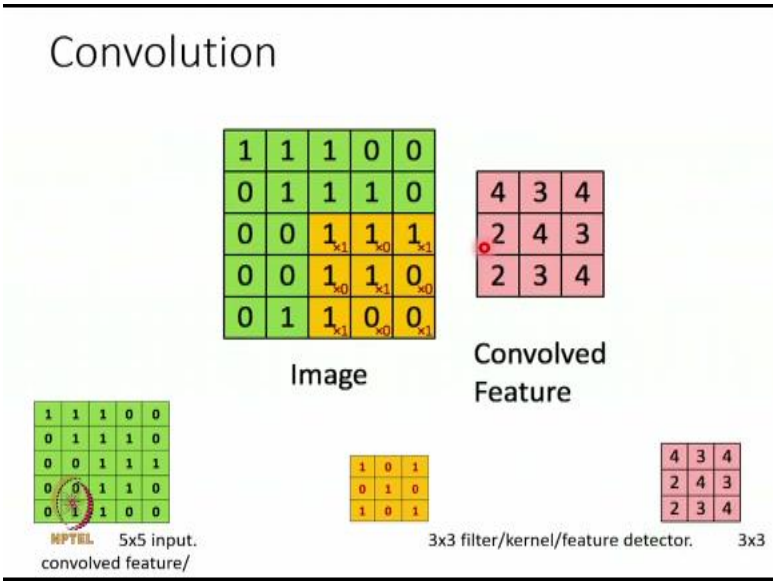
So how do we even start thinking about this? So the intuition that they came up with is that let us come up with invariances. And there are many kinds of invariances. One is called the translational invariance. Let us say I translate every pixel right. So I had a 1 like this. Now I have a 1 like this. It shifted. Is it a 1 still? Absolutely. It is a 1. Or I shifted left, it is still a 1. I reduce the size. It is still a 1.

I blow it up a little bit. It is almost still a 1. I rotate it slightly, it is still a 1. I rotate slightly in the other direction, it is still a 1. But I keep rotating it is it still a one? No, this is not a 1. So people were like okay, so what kind of invariance are they absolutely sure we need. But translation invariance we are pretty much sure we need. That if my eye was exactly in these locations, or it was in these locations or slightly shifted.

Eye still care, right? It is still okay. So they said that let us create an architecture which imposes that whether you have give me the original image or you will give me a slightly shifted image I will still make the same prediction. Now that would not be true in a bit vector representation because every bit is defined and separate. But here we are saying that if the structure is shifted, we should still be okay.

We should be able to deal with it. So for that they came up with this general principle of convolution.

(Refer Slide Time: 04:09)

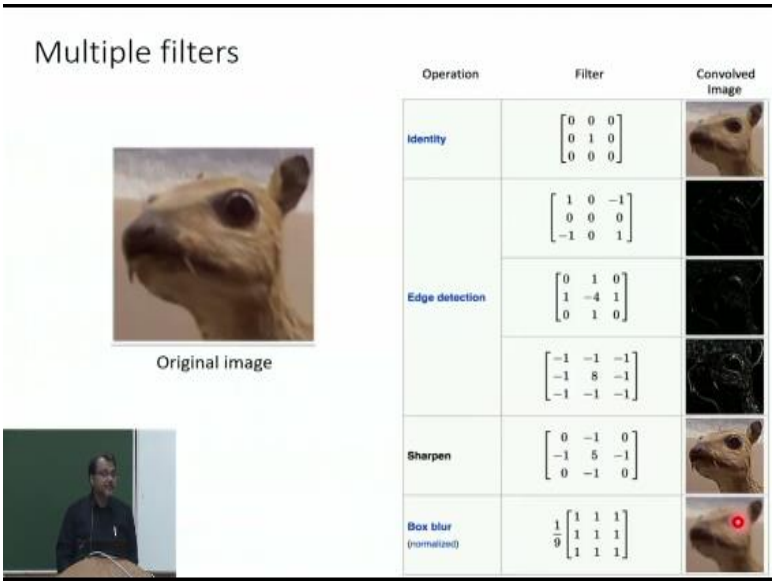


So what does a convolution do? It takes the original image. This is my original bit vector image and it takes a filter. This is called a filter. Now these filters are very common. Have you used any kind of image editing software? I am sure you have used Photoshop or IrfanView or whatever, some photo editing softwares whichever is your favorite. You must have seen lot of filters, low pass filter, edge detection filter, this filter that filter.

What is that doing? Well, it is nothing but an element wise multiplication with an addition. So let us say my filter is 101010101, this filter. I will apply it to each 3 cross 3 subset, not subset but subsequence right sub image. So when I apply it to the very first image here, it will be 1 times 1 plus 1 times 0 plus 1 times 1 and so on so forth and I will multiply these 9 numbers with these 9 numbers and add them.

I will get a 4 because this I believe this and this and this, these 4 terms will contribute. So I will write down the number 4 here. Then I will shift everything by 1, I will still do the same, I will write down the number 3 here. And I will keep doing this. If I had a 5 cross 5 image and I had a 3 cross 3 filter, I will get a 3 cross 3 a convolved feature image or feature matrix whatever. What is the intuition? The intuition is that this is a convolution, think of it as a feature.

(Refer Slide Time: 05:53)



Why do I say think of it as a feature because people have studied many kinds of filters in the past. If I run every image by this original identity filter, I get the original image back. But if I run it through any of these I get edge detections. Can you see these are the edges that are getting detected depending upon what feature values we have.

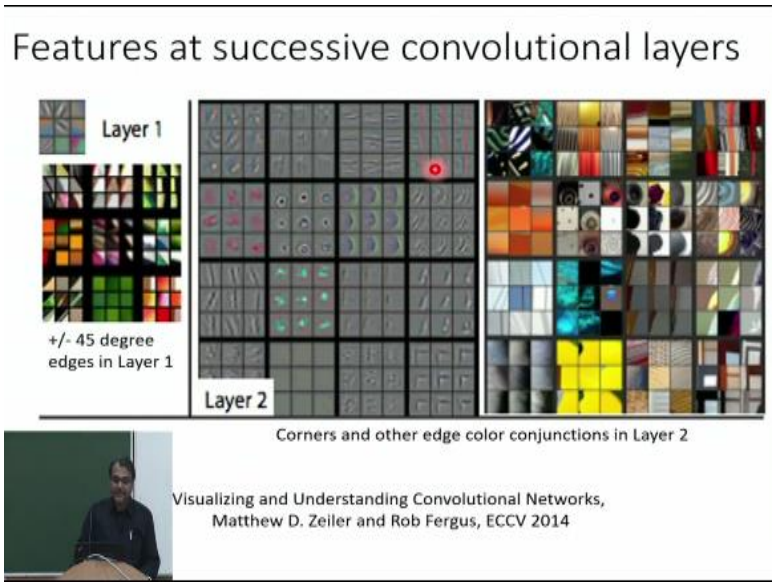
I can also have a sharpened filter, which makes all the edges sharper. I can have a blurr filter, which makes all the edges blurred. In fact, there were computer vision researchers or image processing researchers which did this for a living. They will come up with new filters and they will show the effect of those filters on the images. So therefore, think about it.

What is it saying it is saying that applying a filter is like finding a feature in that image, in that point in that subspace of the image. And I because I am doing it for every part of the image, I will come up with slightly smaller image but which will have a lot of features. It will have the number will indicate whether the feature is present or strong or my negative number will indicate that it is absent or weak or opposite has a negative influence and so on and so forth.

But I do not have to do just one filter, I can do many such filters. These are by the way called convolution filters. I can do many such convolution filters. Each filter creates an image of numbers. Many filters creates many images of numbers. You can again think of it as so what is an original image? An original image is technically x cross y cross RGB values. So you can say 3 channels R channel G channel B channel.

Now after I do a convolution, I will have a 2d image with one filter has one channel. Then if I have many filters, I have many channels like this. Now basically I convert some x cross y cross z image into an x cross y cross z image after doing lot of convolutions on top of it.

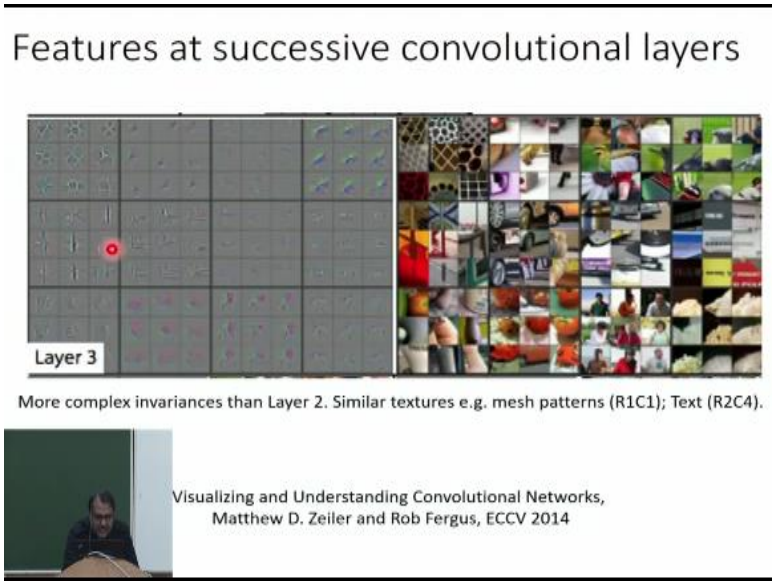
(Refer Slide Time: 07:56)



And when you train this automatically for object detection, these are the kinds of filters you learn for layer 1. Let us look at this is the filter you learn for layer 1. This is sort of saying do I have a slant line pointing downwards right to the right. Do I have a slant line pointing downwards to the left and so on so forth, right. More oblique line, a horizontal line and so on so forth. And this feature is on for these kinds of images.

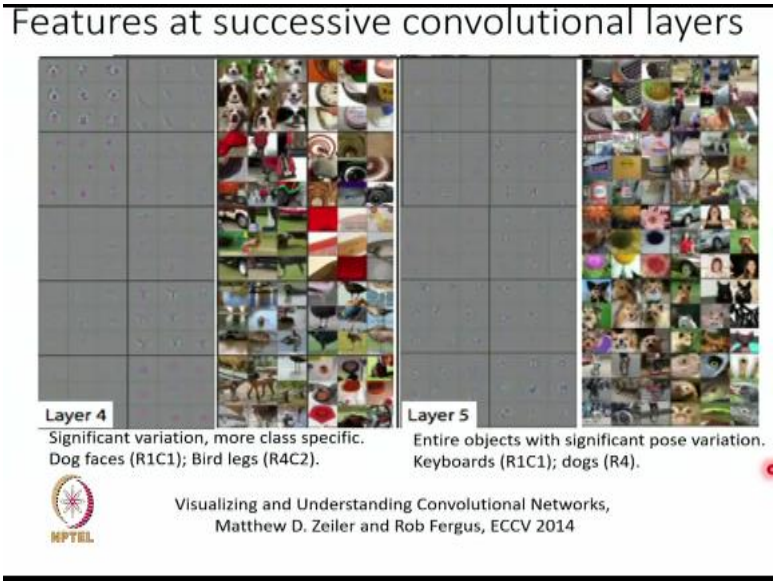
This particular feature is on for these kinds of images and so on so forth. Notice that for object detection, I am learning the very low level information in the first layer. What happens in the second layer? I learn slightly better image, slightly better filters. Notice that I am learning like a full round thing here in the middle, right? I am learning more interesting shapes because now I can combine these together.

(Refer Slide Time: 08:58)



If I keep doing this in layer 3, you can see that I am starting to learn some textures or mesh patterns, right? These kinds of textures correspond to this kind of like say tiger skin or things like that, right? You can have more interesting combinations here.

(Refer Slide Time: 09:19)

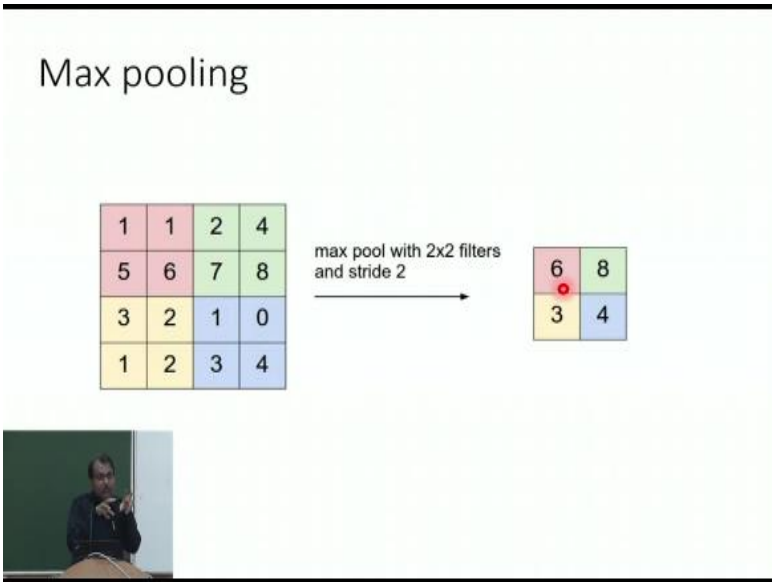


Then if you go to layer 4 and layer 5, look at layer 4. In layer 4, I am almost starting to learn dog faces here, right. Small nose two eyes, you know this kind of a structure, I can start to learn some dog faces, I can start to learn more interesting patterns. And finally, I am able to learn the entire objects with significant pose variation.

So it sort of gives you the handle that if I put in this filter on the input, create another image that learn some useful information then I create more filters on that as input then I create you know next level features, more next level features more next level

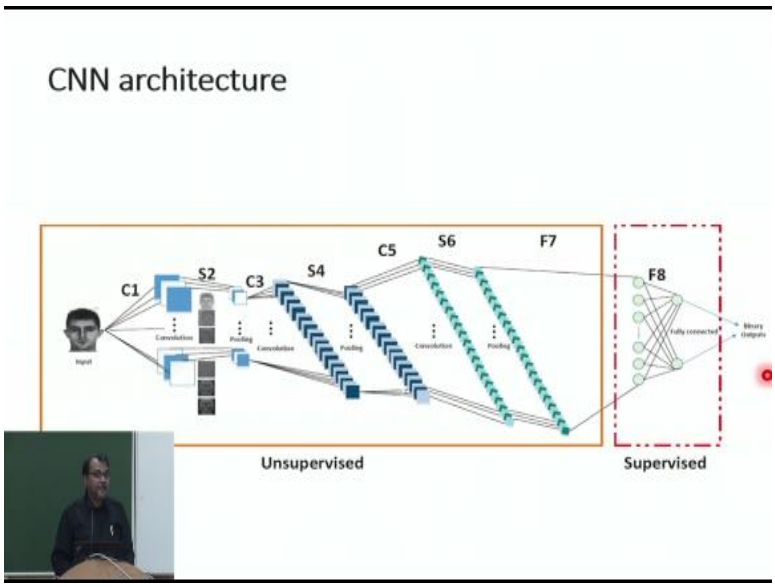
features. So I can keep doing layer by layer convolutional neural network. Now is this translation invariant? That is a question you should ask. And that comes in because of max pooling.

(Refer Slide Time: 10:19)



Now what does Max pooling do? Max pooling says I will take these two and I will just convert it into Max. I will take these two curves too and I will just take a mx and convert it into a 1 cross 1. So now what am I saying irrespective of whether my 6 is here, here, here or here, I will still output 6. In other words, it is saying that I will keep the strongest influence irrespective of where it was in this small sub image. And I will do it at every phase. So I get small local translation invariances.

(Refer Slide Time: 10:50)



So my final scene in architecture becomes something like this. Do a convolution layer, then do a max pooling, this is called max pooling. Do another convolution layer, then do a max pooling. Do another convolution layer, do another max pooling and every time you do this, you are reducing the size of the image. Finally, you are left with a very small image. It may be 1 cross 1 image, but there are many channels.

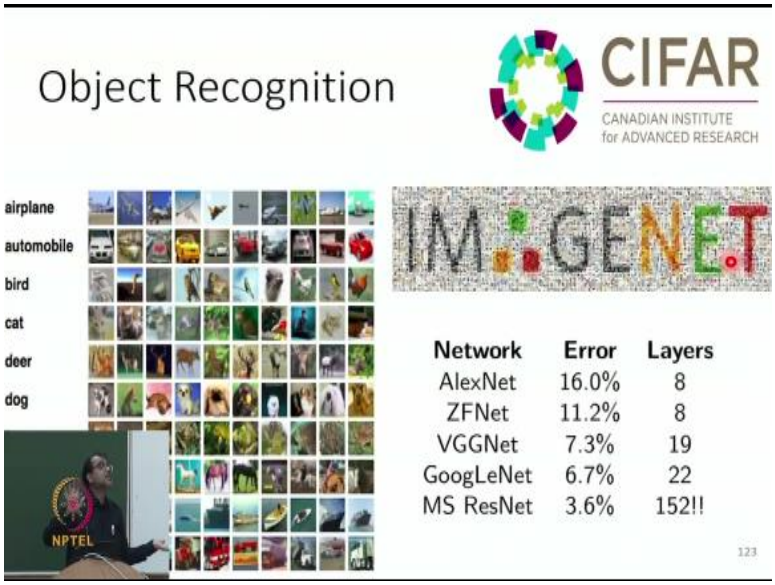
So you are left with a big vector, and then you can run it through a fully connected neural network to make the final output class prediction. Now the biggest problem is how do we figure out what should the filters be? I just showed you some filters that humans had created. But what do I do? Where do which filters do I use? Any guesses? We do not give any filter. We let the algorithm learn it. What is a filter? It is a matrix.

It is a matrix of parameters. Let us define this matrix of parameters as free parameters. And let us just take derivative with respect to those free parameters and backpropagate. And this is going to be a very interesting way of thinking down the line. You think of an intuition. You say okay, if I have this kind of a you know matrix, this will do this. If I have that kind of a vector, it will do that.

You define it as if you were doing it. But then you leave your hands out and say okay, I am not going to give the features, I am only going to give the structure of features. The structure of features is that I will do convolution and they will have this 3 cross 3 weights or 4 cross 4 weights whatever it is. But then the model will learn it completely independently.

So we gave it the structure for the features, which is a filter, but the model figure out which filters which features were necessary for the task at hand. And here is here are the results.

(Refer Slide Time: 12:56)



I think I believe I have shown you the slide on the ImageNet data set. It is an object recognition data set. In this object recognition data set, we have an aeroplane class, an automobile class, a bird class, a cat class, a deer class, a truck class. And we have lots of examples. It is a large data set. And I had told you that until 2011, the non-neural solutions the state of the art was 25% error.

When AlexNet came in, in 2012 I believe it showed, and every year they would make improvement and the improvement would be 0.2 points, 0.4 points, extremely small improvements. The field was moving very slowly, the errors were going down very slowly. The first time AlexNet came from the University of Toronto, it brought the error down to 16% a jump of about 9%. People were shocked, people could not believe it.

People were like this is not possible. But they figured this out, they started working pretty fast and then every year new architectures came which basically went deeper and deeper. I believe ResNet is 156 layers. So I could be wrong in my exact number. And finally, I am told that the human performance on this task is about 2%. And we have been able to beat that. Oh 152 layers it says on the slide.

Now beating humans in this data set does not mean that image recognition is soiled or object recognition is soiled. No. It only knows these classes and no other class, right? It cannot learn a new class quickly. If I showed you the seagull for the very first time

in your life, but you have seen many other birds. How many data points would you need to learn what is a seagull?

How many birds would you have to see to learn what is a seagull? Jay says one, somebody says two, nobody has said three so far. You have so much conviction in your ability to learn what a seagull is that you think one or maximum two examples of a seagull, and I would figure it out. And you are probably right. Do you think how many examples of a neural networks will need to learn a new class?

Thousands maybe. So we have not solved the problem. By the way, this is called zero short learning. Sometimes I do not even show you a seagull. I just say that a seagull in English I say a seagull is like a is a white bird with a fat this and a long beak and blah. I just give the description and maybe you would not be able to recognize seagull. But there is some chance you would be without even seeing the first seagull in life, right.

If you think about our medical professionals, people who do surgeries people you every once in a while come up with a rare abnormality. Have you been trained on that abnormality? No, you have never seen that in life. You may have read it in the book. But because you have read it in the book when you see it, you can make the connection. However, machines cannot do this. Right?

So there are a lot of things that our machines cannot do in computer vision today. So vision is not a solved problem. But the amount of progress we have made is by leaps and trucks. Yes, there was a question. Right. So Parth says that I could, I might if I see a seagull in practice I will see them like a video, not like an image. But I can show you images also I can do the test, where you have to learn a new concept.

And I am showing you images only. And then it is equivalent to a machine, right? It is not very different. But yes, you can also do a video and I am not sure that our machine will be very good with videos either. Right? So I think this is a good time to stop today. We are left with one small part of this topic, which is deep reinforcement learning. So now what have we done? We have learned how to approximate a function.

But we have not learned how to use it in reinforcement learning. Now that is a full course in and of itself. Just deep reinforcement learning is a full course actually, believe it or not. But we will cover 10 minutes of that at the beginning of the next class to complete the story that we started some time back with the RL class. After completing that story, that would be the last mathematical thing that we will study and then we will move into ethics of AI.

We will talk about if AI systems are really becoming that successful as they are, they will start showing up in our world in our real world very soon. And what are the challenges that we will need to solve in order for them to be successful in different applications? Where is it a good use of AI? Where is it not a good use of AI? Would AI lead to better fairness or worse fairness?

Would AI lead to new kinds of mistakes, but less number of them? Would AI lead to, what would AI in our day to day life lead to? That is the question that we would ask. And we would discuss this in the next class. Maybe we will have a part of that conversation in the last class. And in the last class, we will wrap up so we will go over the full course that we studied.

We will talk about what have we studied, what are the take home principles and what is the big picture of the full AI course.