Architectural Design of Digital Integrated Circuits
Prof. Indranil Hatai
School of VLSI Technology
Indian Institute of Engineering Science and Technology, Shibpur, Howrah

Lecture – 66
Hardware for Machine Learning: Design Considerations Design Tips

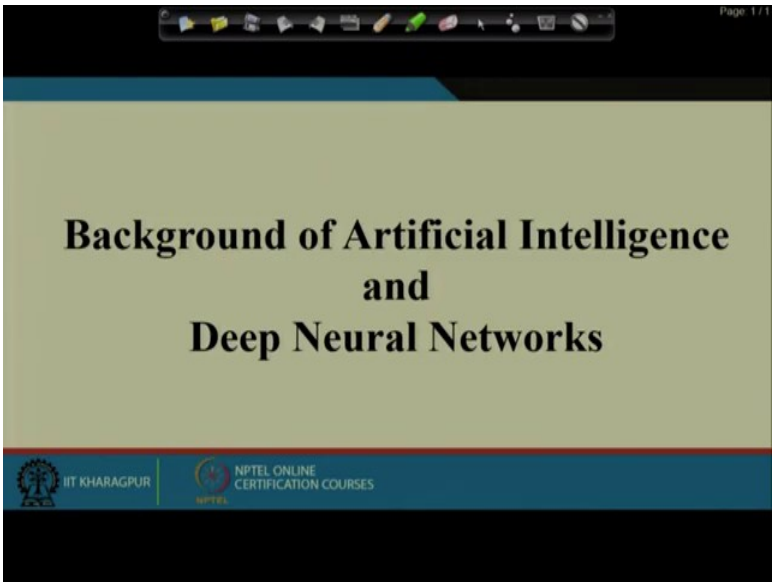
Hello everyone, welcome to the course on Architectural Design of ICs. So, today we will see one of the emerging areas of this or the latest trends in VLSI design, which is nothing but this machine learning or artificial intelligence. So, basically people are trying for different-different algorithms to make these artificial intelligence to be applicable everywhere.

Actually, machine intelligence or this artificial intelligence has the applicability everywhere. So, due to its application in a very wide field, it is the latest hot topic in VLSI design. So, that is why actually I thought that some flavors of this artificial intelligence hardware design aspects, I will just provide to you. So, if anyone is more interested to work on this particular field then, we can just take that offline.

But, the thing is that we will here today, we will discuss some of the features; that means, there are huge scope. And, people are basically trying to develop the hardware as well as that to develop the software; that means the algorithms efficient algorithms for that. So, like crazy so, that is why we will today we will see the hardware for machine learning design consideration and the design tips. What people are basically following, what people have already doing, it is one actually emerging field.

So, that is why people everywhere everyday people are basically working to develop the hardware for machine learning as well as the algorithm. That means, for the first time you have to develop the algorithm then, that algorithm now you have to put into one hardware. So, how far the people have gone and what are there that we will see in this particular lecture series.

(Refer Slide Time: 02:06)



So, the very first we will see the background of artificial intelligence and then deep neural network. The first the thing is that we by the term itself it says that, what is artificial intelligence. So, as a human being whatever intelligence we got that is the natural intelligences we got that which is the gift from our god. But, if we put the machines, if we consider any machines or robots or any machines, they are basically programmed.

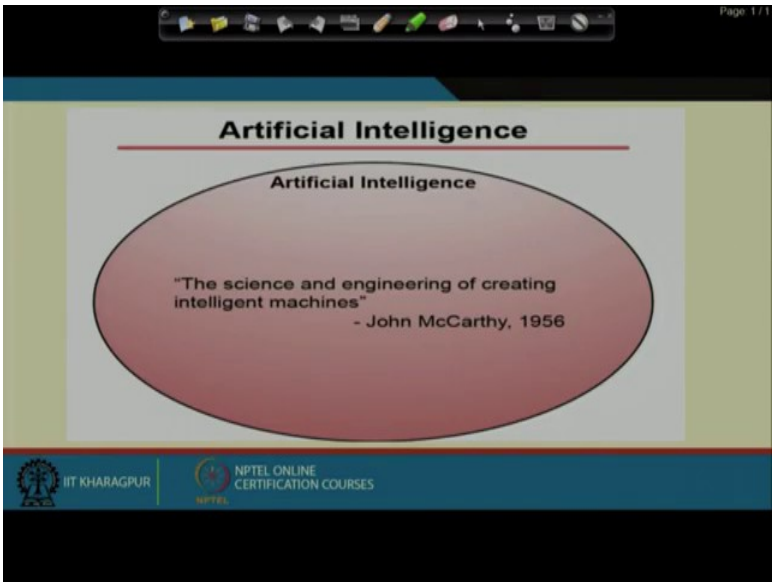
So, they has whenever as a programmer whenever we put that some specific instruction to do. So, that is why actually if I know the term in terminology in VLSI, which is that application specific integrated circuit, that means, in VLSI whenever we are designing particular circuit. They are basically particular to one particular application. They are specific to one particular job to be done on that particular system design.

So, nowadays actually with the artificial intelligence, it will be the circuit; that means, consideration will be not application specific, it will be adaptive in its kind. That means, like day to day as we learn; that means, we experience something we have gone through some of the; that means, situation and then we experience the information day by day. We learn something from that particular situation or that particular experience. So, if we put that in capability in the machines. So, that is nothing but the artificial intelligence ok.

So, I just this is not the bookish way, I said that what is the meaning of artificial intelligence, as such generic way this is the meaning of artificial intelligence. And, that is

why I just said like this is that. Now, as a hardware engineer or as a VLSI engineer, I have to make the hardware; so, that in machine, this basically they the software part as well as the hardware part, both should be there. So, that can work together to make one efficient artificial intelligence hardware.

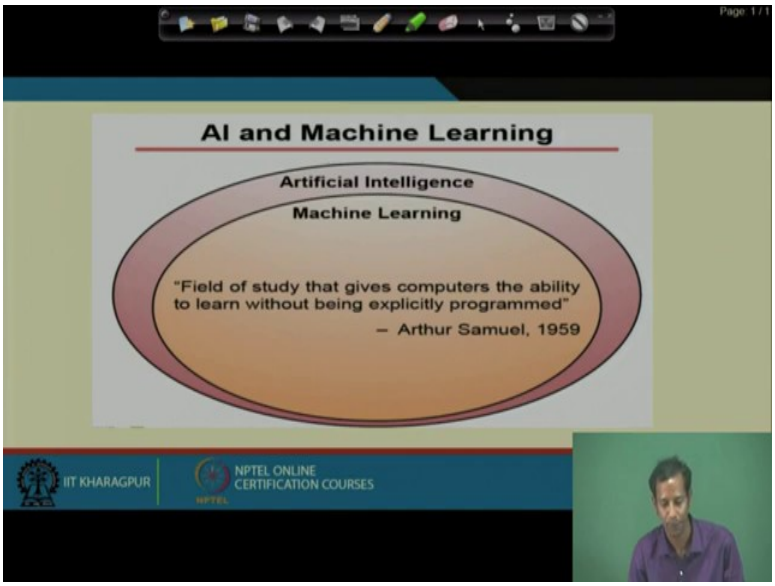
(Refer Slide Time: 04:20)



So, the artificial intelligence basically has been developed by, it is not that just now this terminology came, artificial intelligence it came very earlier in 1956 by John McCarthy. He said that the science and the engineering for creating one intelligence machines, I need the artificial intelligent. That means, as in machines there is no brain we have brain.

So, that is why we are intelligent but, in machines they are not actually, they are not having any brain that is why they are not that much intelligent. So, if you want to make intelligent then, we have to put some brain inside of it. So, that every day it can adopt according to the corresponding situation, which we can do as we have our brain.

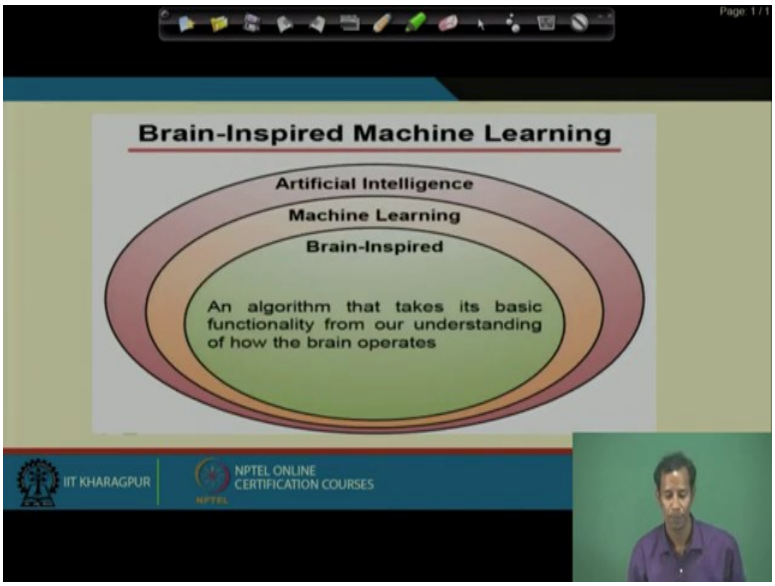
(Refer Slide Time: 05:07)



So, then whenever we just have considered this machine learning at this artificial intelligence. So, at that time in 1959, the Arthur Samuels actually introduced or this terminology is that machine learning. So; that means, as a like we human actually we learned. So, the machine also should have the capability to learn day to day, the field of study that gives computers the ability to learn without being explicitly programmed.

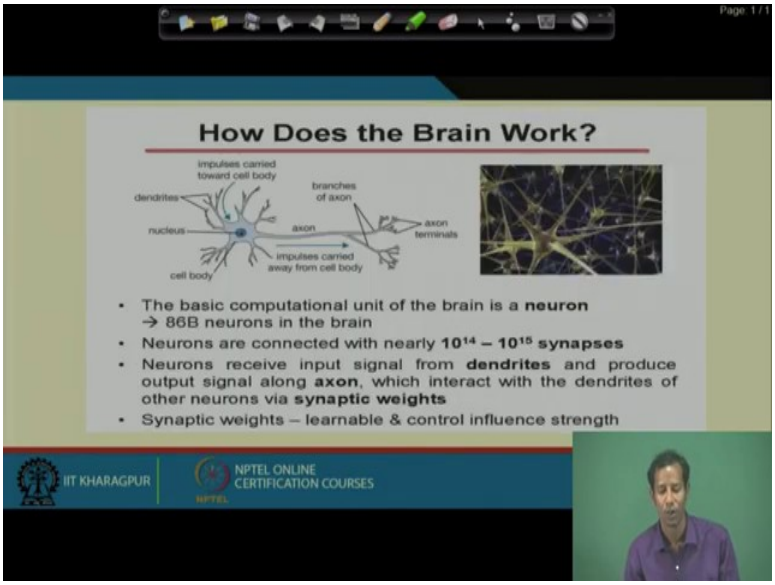
So, application specific means, it is explicitly programmed; that means one particular program. So, it has it can do only a single job but, if you have the capability to learn. So, at that time it is not that it will do the specific job only, it can do the that means, the other job too which will do not particularly specific.

(Refer Slide Time: 05:56)



So, at that time this machine learning if you have to learn this day to day. So, at the time you have this brain inspired computing, inside of the machines. So, what is that? This brain inspired is that one algorithm that takes it basic functionality from our understanding how the brain operates. So, basically we will see that how our brain works and if I can say that. And, from our brain that intelligence comes to our body.

(Refer Slide Time: 06:32)



So, first we will see how does the brain work? Ok. So, in brain we are having 86 billion numbers of neurons. So, what are these neurons? Neurons are the basically this what I

say is that, some cells. So, these 86 billion neurons in the brain, they are basically connected to near about 10 to the power 14 to 10 to the power 15 synapses ok. So, if you see these particular pictures. So, this is one neurons picture where we are having one central processor or which is the known as the nucleus of that neurons.

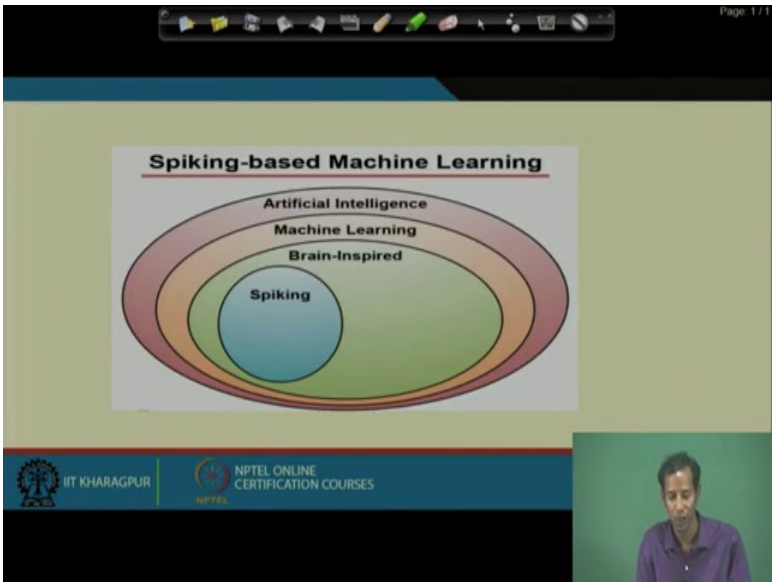
Then, we are having the cell body, we are having the dendrites and then we are having these axons which are basically connected with the corresponding synapses of other or the dendrites of the other neurons. Then, the neurons receive the input signals from dendrites and produce the output signals along axon, which interacts with the dendrites of the other neuron via synaptic weights. So, that means, these dendrites is something like it is the sensor, which is connected with these neurons.

Now, these neurons whenever after; that means, it process the signal which is received from this dendrites. And, then the corresponding decision has been processed to the other neurons via this axon. And, they are being connected with the corresponding dendrites of the other neurons, via some synaptic weights.

So, these synaptic weights are learnable and control the influence strength. So, that means, these synaptic weights are adaptive in nature. So, these synaptic weights they are not fixed. So, now, if I can build up the circuits, some kind of this or if I can build these neurons, inside of this machines. So, at that time it can also have the adaptability; that means, adaptability capability to learn or to; that means, to remember that this is the situation.

So, at that time I will take the decision immediately. I do not have to wait for ok, again reanalyze the situation and then again you take the decision or if the situation is totally new to me. So, at the time what will be my decision; so, that machine itself can take that particular decision depending on the newer situation or the older situation.

(Refer Slide Time: 09:19)

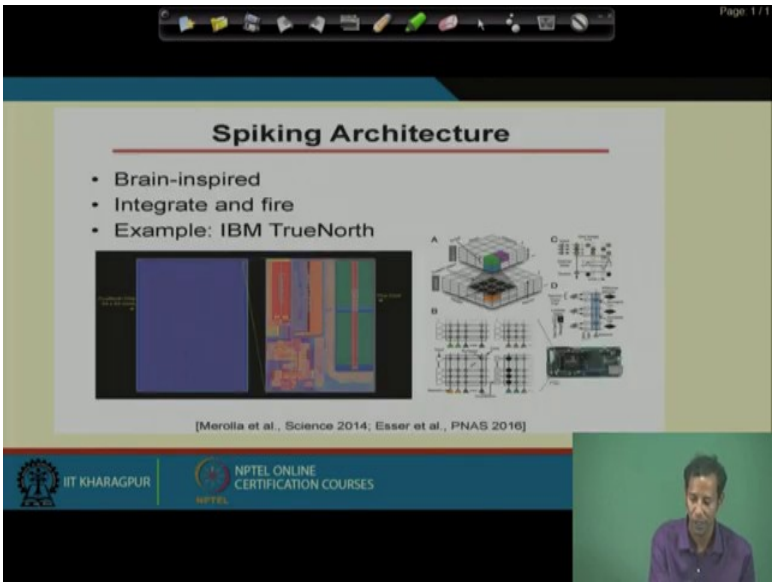


So, whenever we are considering this brain inspired computing. So, at that time the one thing is needed that is the spiking. So, that means, whenever I suppose I touched the fire. So, it creates so, the skin is my sensor. So, whenever this senses something.

So, that it is hot. So, this is a sense which goes to my brain and the create one spike. And, that is why I can feel that I am this is hot, if I touch ice. So, at that time it is the it is cold, that is not like fire. So, these are the feelings which create goes into the brain and then it creates the corresponding spike about the corresponding feelings.

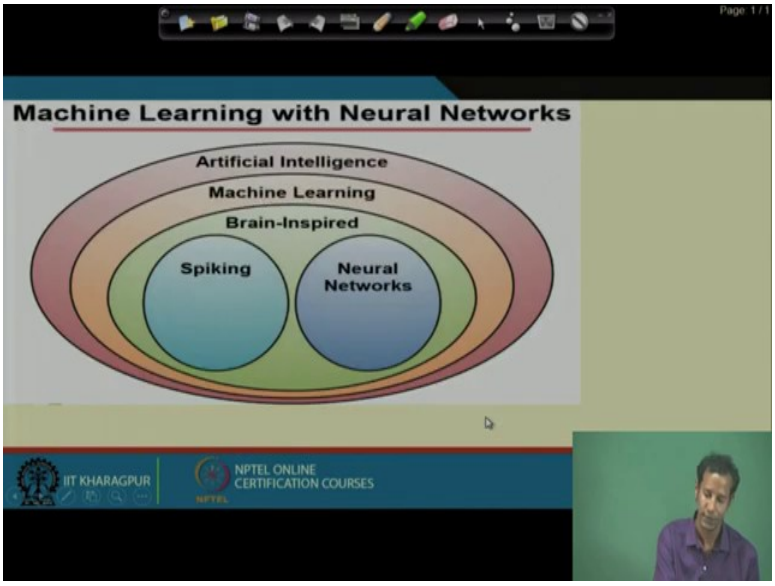
So, suppose through nose. Nose is another sensor to our body. So, through nose sometimes I can smell good, sometimes I can smell bad. So, what is good smell? What is bad smell? So, that information whenever the received signal goes to the brain. So, brain actually consider that this signal is good, this signal is bad. So, this creates a spike based on the spike, now I can get the feelings that. I am getting good feelings or I am getting good bad feelings.

(Refer Slide Time: 10:32)



So, the people hair so, much advance that, they have create the circuit for which we will be just act as a in our inside of our brain, how this spiking happens, they have created this spiking architecture. So, this is the corresponding chip for this spiking circuit. So, and this is being published in 2014 and 2016.

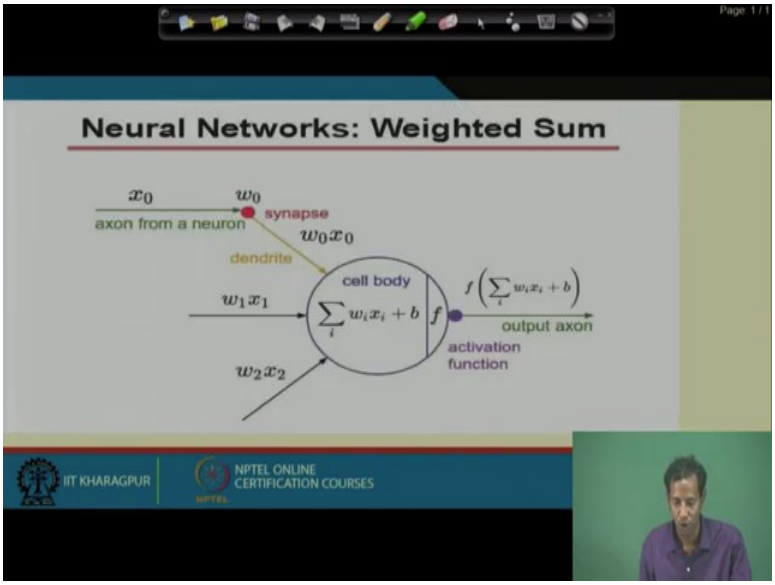
(Refer Slide Time: 11:00)



Then, whenever we are actually considering this brain inspired. So, spiking is one of the things, the other thing is that neural network ok. So, neural network is nothing but this

whatever we called that this neurons in our in our brains. So, these neural networks will be the brain or these neurons inside of your corresponding machines.

(Refer Slide Time: 11:26)

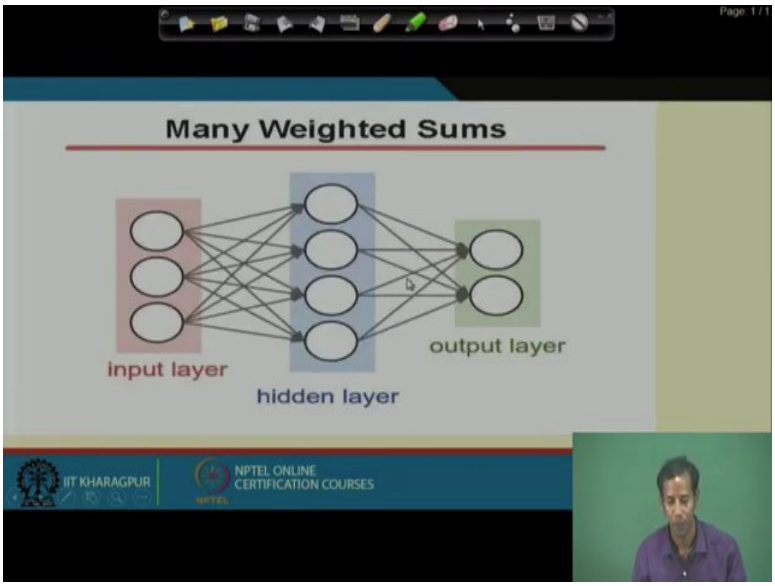


So, the neural network, how this neural network work is that? Suppose, this is our cell body and then each of this is basically these nodes are called as the synapses which are basically connected with the cell body with some of the weights, which are the dendrites. So, this is the input side.

So, that is why this is with some weight like here, this weight is w 0, here w 1, this is w 2 and this is different signal from other neurons or the other that means, just like the connection in the inside of the brain. And, then it develops summits that all this information together, it takes it does some operation and then it basically creates the, it basically generates the decision, which comes out as the output action ok.

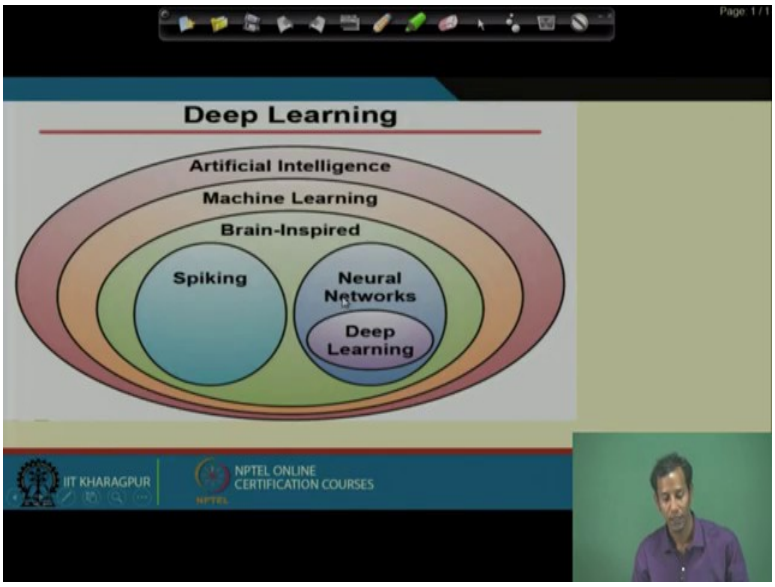
So, the activation basically happens depending on the, what operation is happening inside of this cell body.

(Refer Slide Time: 12:26)



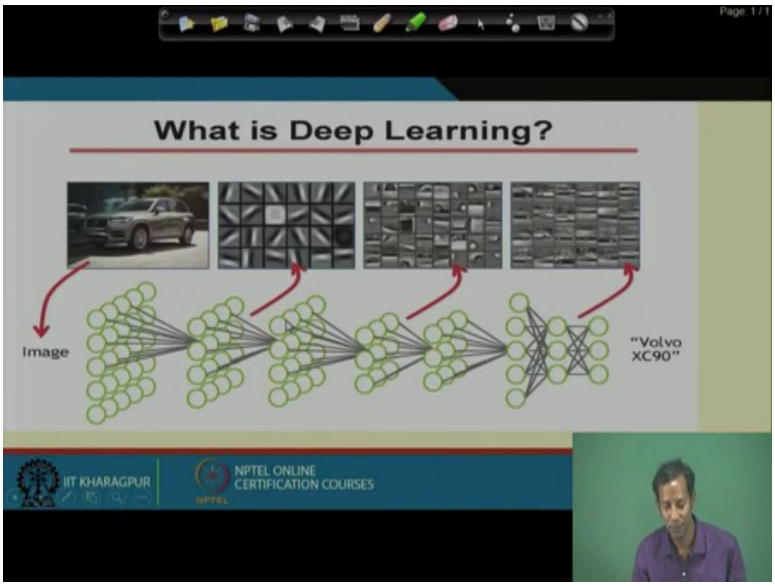
So, if I actually if I consider 3 layer. So, this will be the input layer, this will be the hidden layer and this will be the output layer. And, each of this nodes in this layer, they are basically connected to each other via different-different weights.

(Refer Slide Time: 12:46)



So, this is very much needed whenever, we will just model our brains using this neural networks. So, in neural networks then, we need this deep learning.

(Refer Slide Time: 12:59)

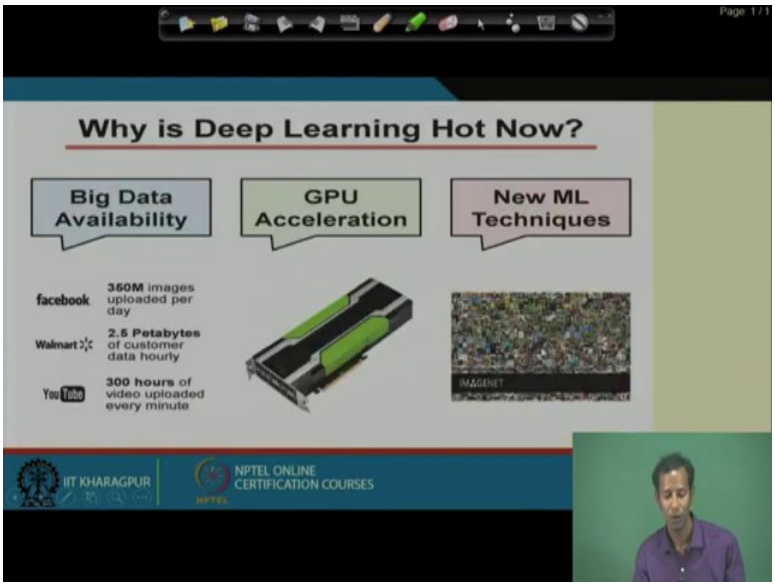


What does it means this deep learning is that. Suppose, how we recognize suppose this is one, suppose this one is one of the car, the model is Volvo XC90. So, now by seeing actually I can visualize that ok, this is one car. If I never seen these particular car, I cannot say that ok, this is Volvo XC90. If I know from the beginning or from my previous experience, whenever, I know that this is the model or how this looks like. This is the model of Volvo XC90 then, only I can say that this is Volvo XC90 model.

So that means, that information is already stored in our memory somewhere. So, that is why that is if I provide that kind of learning; that means technique inside of the machines. Then, how we can put using deep learning method along with this neural network, I can provide that information to the machine's brain.

So, what it does is that we have the image. So, the image has different layers. So, that then the layers is being processed and then we come out as that ok, this is the Volvo XC90 model so, how we do?

(Refer Slide Time: 14:19)



So, as I said that in 1959 the machine and this artificial intelligence this terminology came. But, still the people are actually people are trying but, why nowadays this deep learning or this machine intelligence is hot now.

Because, one of this; that means, major; that means, aspect is that big data availability. So, big data availability means according to the Facebook 350 million images uploaded per day. According to the Wal-Mart 2.5 petabytes of customer data per hour it uploaded in the; that means, in the cloud and, in according to the YouTube, 300 hours of video uploaded every minute.

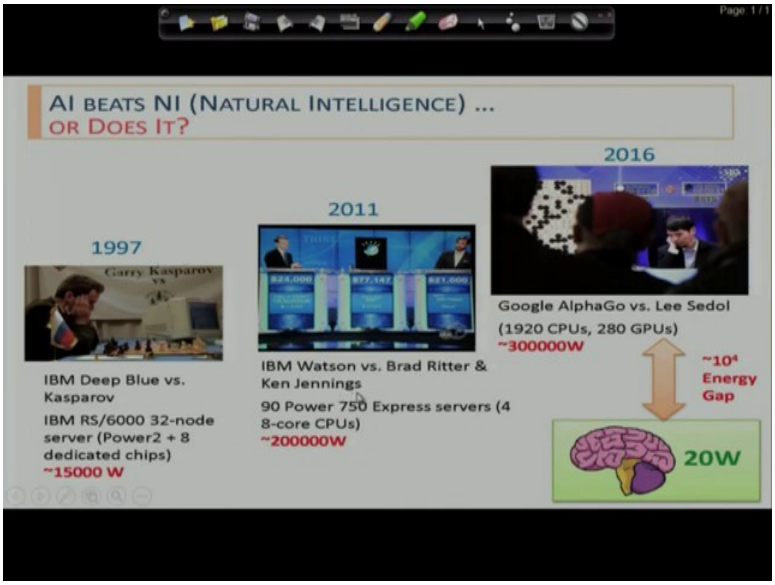
So, all these creates what all these creates a huge learning actually data sets to me. That means, different-different situation, how we actually learn. If we gone through the; that means, the situation. Whether that is bad situation or good situation from that only we learn so, here also, if you have some training modules or the training situations to the machines. Then, only you will be able; that means, you will put the brain or you will be actually your hardware will be so much adaptable. That it can recognize this is the scene. So, where it relates or it correlates to some of this.

Then, for actually whenever we are having this big data availability, we are having this big data or the training sets we are already having. But, at that time I need a huge processing power to process all these data's because, the images is 350 million. The actually the storage is 2.5 petabytes. So, all this information to process or to; that means,

to process all this data's, I need the processing element or the processing power very high.

So, the GPU acceleration now a days is also one of the support system to the machine learning advancement. So, advancement is the GPU acceleration also creates that ok. Now, you can try with or you can process this much of data in the GPU. So, you can work on this deep learning method. And, then the new machine learning techniques, which are evolved or which are basically emerging day by day.

(Refer Slide Time: 16:55)



Now, in 1987 actually this this Kasparov has played against IBM deep blue computer. And, it consumes actually at that chess play, actually the IBM power, actually consumes 15000 Watts. In 2011 again this IBM Watson versus this Brad Ritter and Ken Jennings, they have actually played some games.

And, they are actually the processing has been done on 750 express server, which are having 48 cores CPU's which consumes near about that 200000 Watts of power. And, then in 2016 this Google AlphaGo versus Lee Sedol has again played one game, where it is consuming actually it the processing power of this Google AlphaGo is having 1920 CPU's and 200 GPU's, which consumes more than 300000 Watts power.

But, whenever this human is playing, at that time it consumes only 20 Watt of power inside our brain, as a human brain. So, the energy gap is in the range of 10 to the power

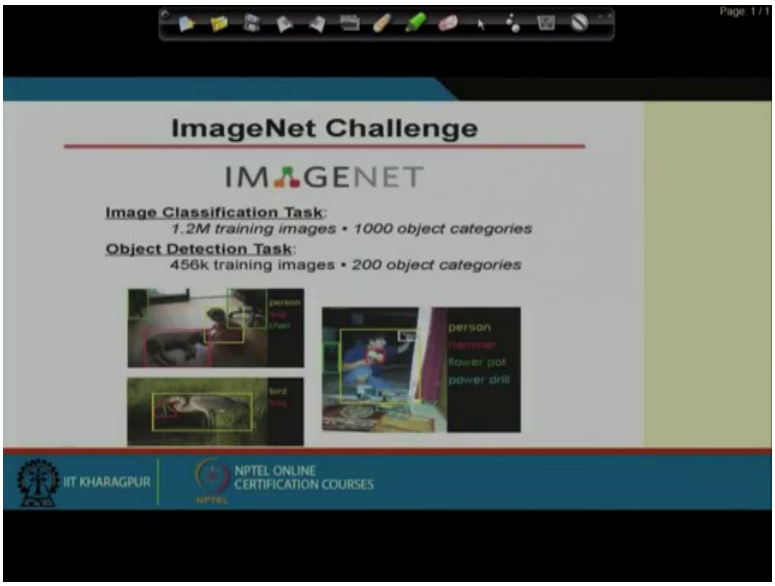
4. Not only the high processing power. So, because of this high processing power requirement, it consume it consumes or it needs one high power also.

(Refer Slide Time: 18:42)

So, that is why whenever we are actually this is one on chip embedded learning. So, depending on that; that means, this machine learning or artificial intelligence you can put in mobiles, wearables, IoT device, it is not that every time you have to upload the data to the cloud. But, on the device itself, if you have the artificial intelligence operation so, then again you do not need the processing powers at very high.

But, whenever you are considering your mobile phones, at that time you have limited processing power. So, within that limited processing powers, how much machine intelligence you can provide. So, that is the major challenge in machine learning hardware design or artificial intelligence hardware design.

(Refer Slide Time: 19:30)

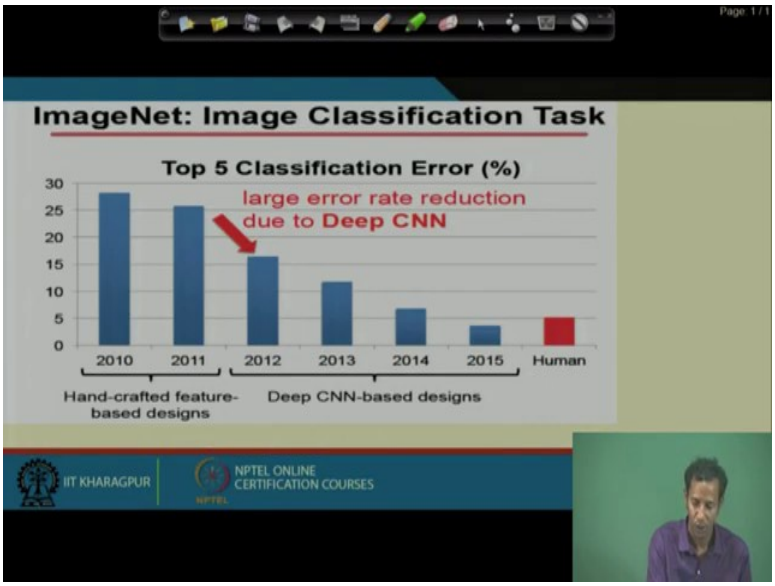


So, the ImageNet is one of the technique in neural network. So, you according to this ImageNet, you can have the 1.2 million training image, where you can object that means 1000 objects, you can basically detects. That means, the detection of this here, if you see these particular image. It can detects that there is a chair, there is a dog, there is a sorry there is a dog, there is a person, if you see these particular image. So, you are having one bird, you are having one frog.

So, if you consider this particular image. So, it is person, hammer, flower pot, power drill. So; that means, these are the objects, which I found in these particular image. So, if I have this training datasets because, hammer that that maybe of different size, different actually it looks like different. But, still I can actually visualizing the image, as a human being how I can percept that ok, this is the image of hammer.

So, the same thing actually can be put inside of the machine. So, that whether the size is different or whether it looks like different. But, still it can recognize this is the person, who is the person? That cannot be actually recognized but, this is a person, this is a frog, this is a bird, this is a dog. Those types of classification you can do by train your neural network. So, if you have more training data. So, we can; that means, you can get the classification or the object detection in a more

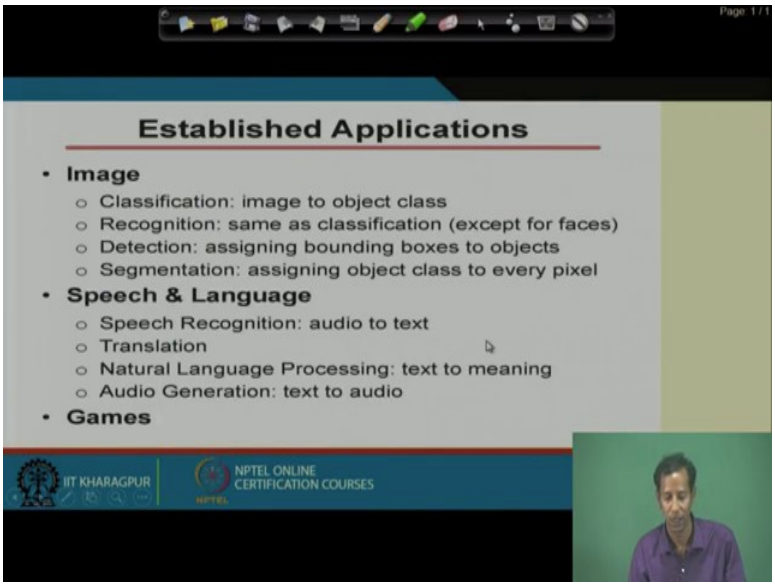
(Refer Slide Time: 21:14)



So, then actually if you see that, use of this image classification using this deep neural network, the error for properly detecting the object, it is much lesser than the human. That means, human also sometimes it can make the error or make the mistakes.

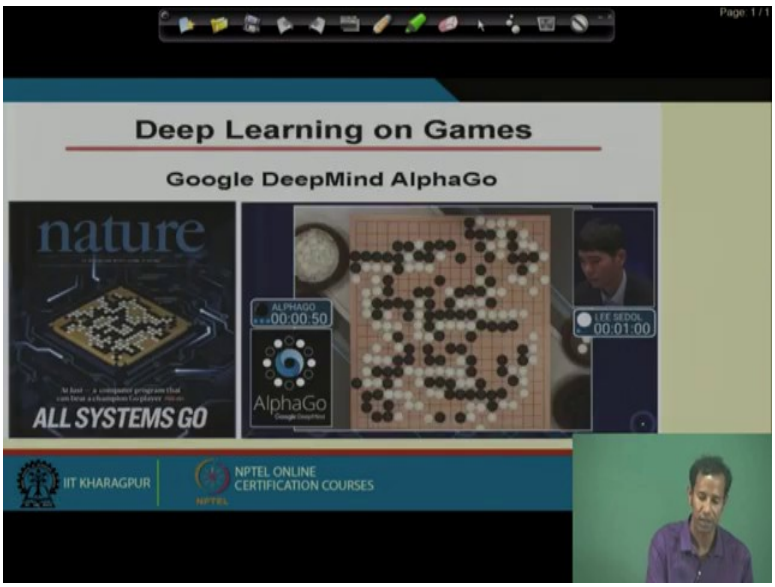
But, this deep neural network they are send with so varieties or so that means whenever we are putting the training sequence means, we are giving them too many of situations. So, that is why their learning is much more than the humans intelligence. So, sometimes the human make; that means, for the object classification or object detection, the human make more than the machines one.

(Refer Slide Time: 22:07)



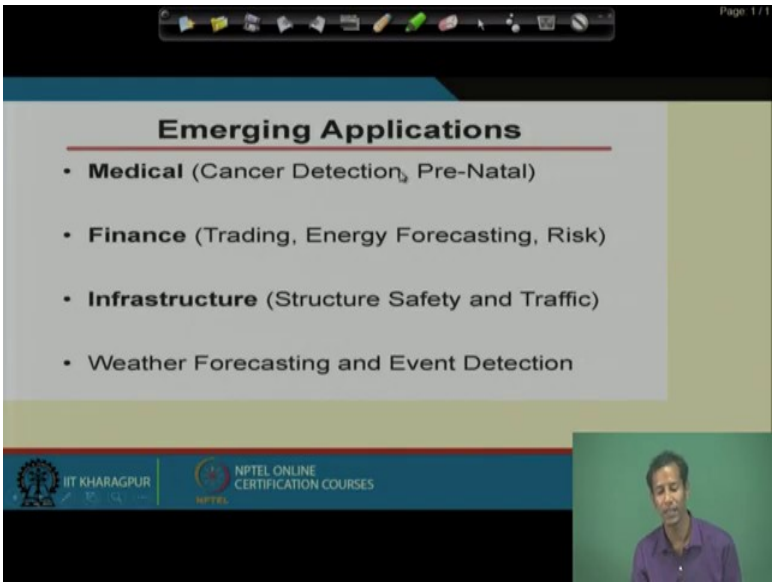
So, this then the application of machine learning is everywhere for image classification, image recognition, image detection, image segmentation in speech and language for speech recognition for translation. That means, from one language to the other language then, the natural language processing then, the audio generation. So, everywhere you can use this artificial intelligence.

(Refer Slide Time: 22:32)



So, in games also; that means, you can play with the computer just like one human being.

(Refer Slide Time: 22:42)



Then, the emerging applications are mainly on the medical applications. Medical application means, you can detect cancer, you can detect any prenatal; that means, the malfunctions. Then, you can put the artificial intelligence in finance; that means that the market share, trends in trading, energy forecasting and the risk factors in mutual funds or any other funds or any other investment.

Then, you can put this artificial intelligence in infrastructure for structure safety and the traffic controlling. Then, we you can put this artificial intelligence in weather forecasting and event detections. It is not that limited to this whatever I have mentioned, it is machine learning is applicable to every of this; that means, area.

(Refer Slide Time: 23:31)

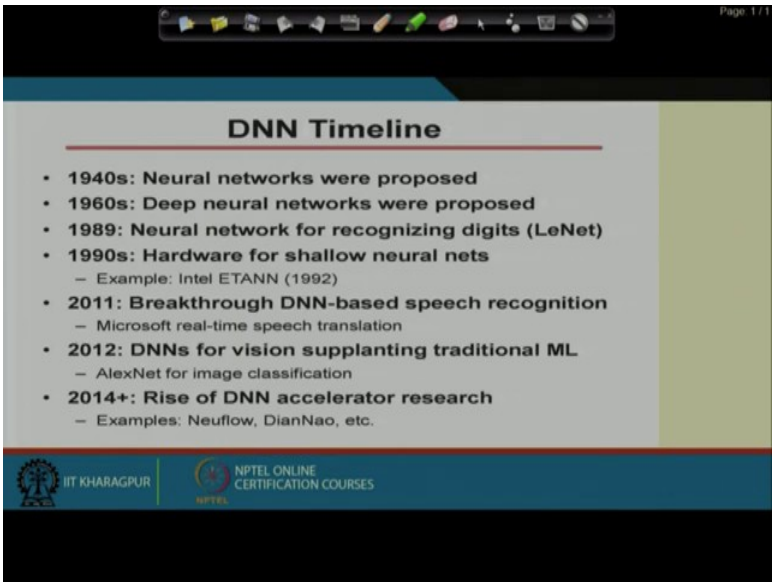
So, nowadays actually we are having these self driving cars. So, whenever this self driving cars whenever, I drive a car. So, at that time I know that these are the; that means, the signals which we have to obey. So, if there is only any obstacles. So, at the time I have to stop, I have to put the brake.

So, or I have to; that means, how much is the open area. So, that open area, I will pass my car. So, everything as a human being whenever I drive the car, I take the decisions. So, if I put all this information inside one machines or inside of the car processors. So, at the time it can detects at the time what it needed is that, what are the signals and what is the meaning of those signals. Then, it has to detect the objects, what are their whenever I am just going through the lane.

So, at the time what are the; that means, traffic light condition and then I have to take the right lane or I have to take the left lane. So, all these information if I provide to the car itself so, then there is no need for the human interference, who will be sitting on the driver seat. So, at that time this car will be self driving; that means, the car automatically has the decision or the capability to take the decisions that I have to go by this way.

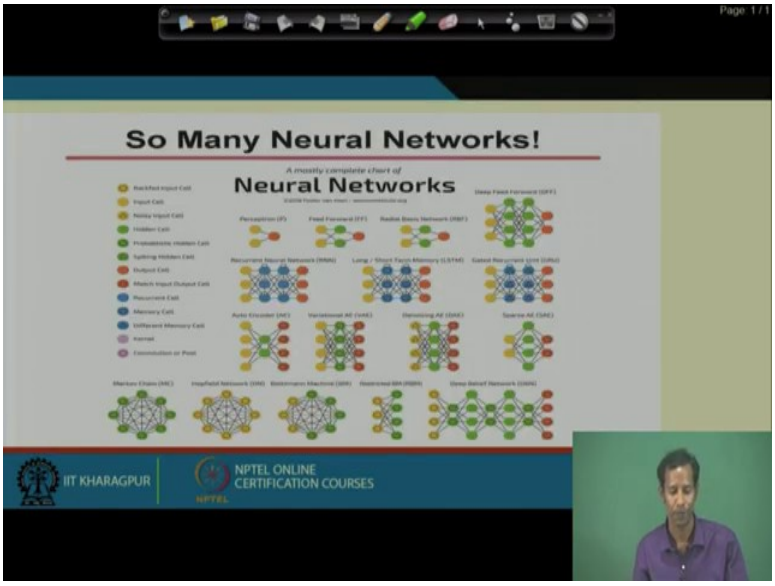
Or at what speed I can do or at what side I will pass my car. So, everything at that time I can make that automated. So, and that is possible through deep learning.

(Refer Slide Time: 25:23)



So, if I actually talk about the overview of this deep neural network and that is why this nowadays this deep neural network is the hot topic of research. So, deep neural network timeline according to this in 1940s, this deep neural network were proposed in 1960, the deep neural network were proposed again. And, then in 2014 actually it is evolving day by day. In 2014 the rise of this deep neural network accelerator research which becomes or which pushes this artificial intelligence in the forward direction.

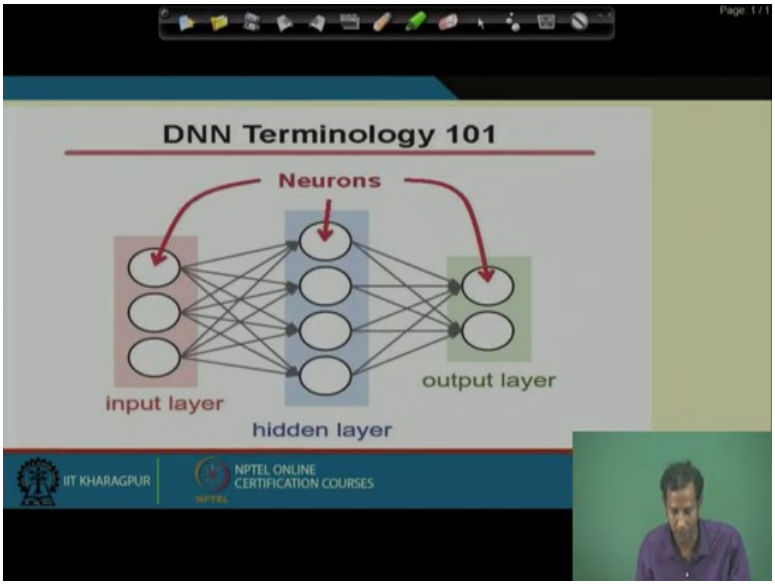
(Refer Slide Time: 26:01)



There are actually so many neural networks which are already there but, still people are trying to develop this efficient neural network algorithm which takes minimum number of datasets for it is training purpose.

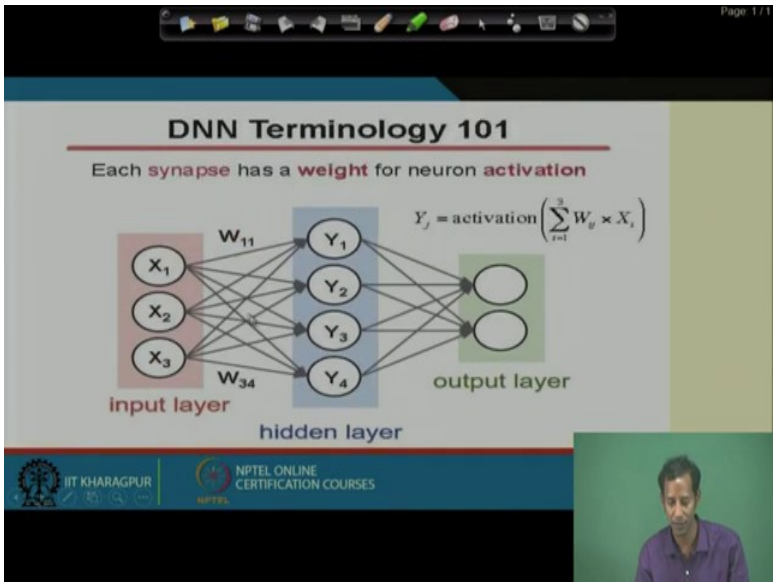
And then otherwise what will happen, if I had to train my model that. So, many datasets; that means, number of datasets if it is let us say; that means, 20 million. So, to 20 million data sets, I need to process somewhere. So, at that time I need the hardware capability very high. So, to avoid that we need the deep neural network algorithm should be so much efficient. So, that the processing power also becomes very much on the lower side.

(Refer Slide Time: 26:51)

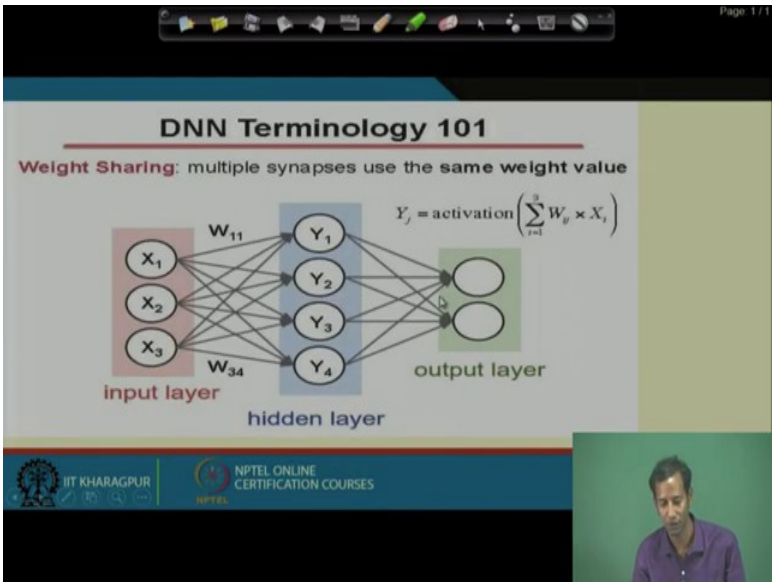


So, we are already in deep neural network we are having 3, we consider for this basic understanding of this deep neural network, we will be having 3 layer input layer, hidden layer and output layer. Depending on that they are basically connected with the synapses. Each of these synapses has different-different weight and which basically activates the corresponding neurons ok.

(Refer Slide Time: 27:08)

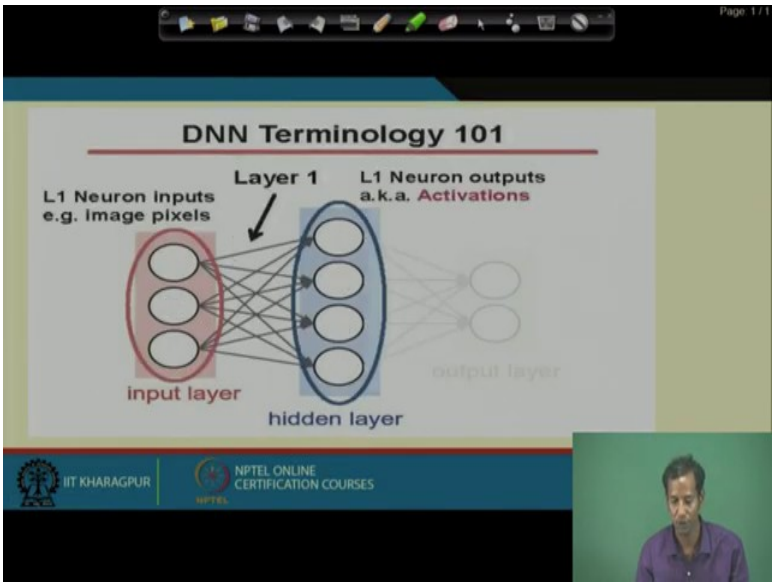


(Refer Slide Time: 27:15)



So, then we are having this weight sharing in deep neural network, I will not go into the details of this because, this lecture is mainly focused on the hardware aspects of the deep neural network.

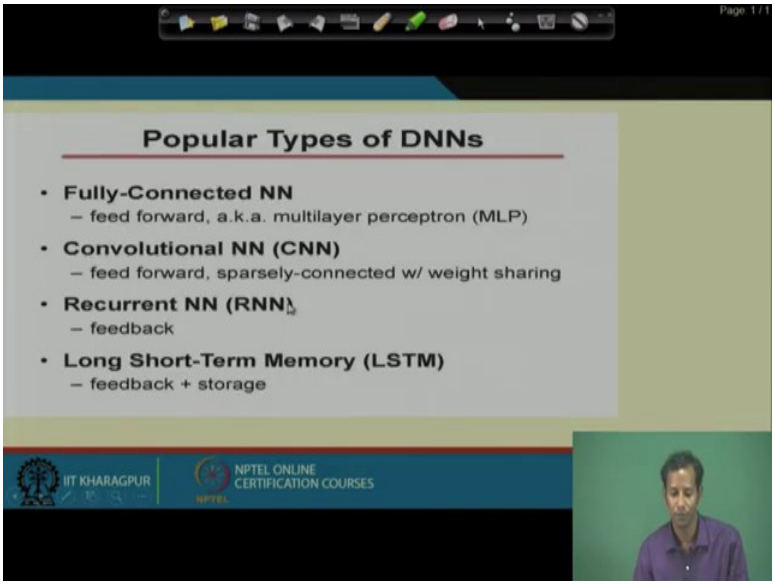
(Refer Slide Time: 27:20)



So, the algorithmic aspect if you want to know more than take any deep neural network or this artificial intelligence course. There you will get the how to classify this you know how to; that means, detect the object, how to classify the image ok.

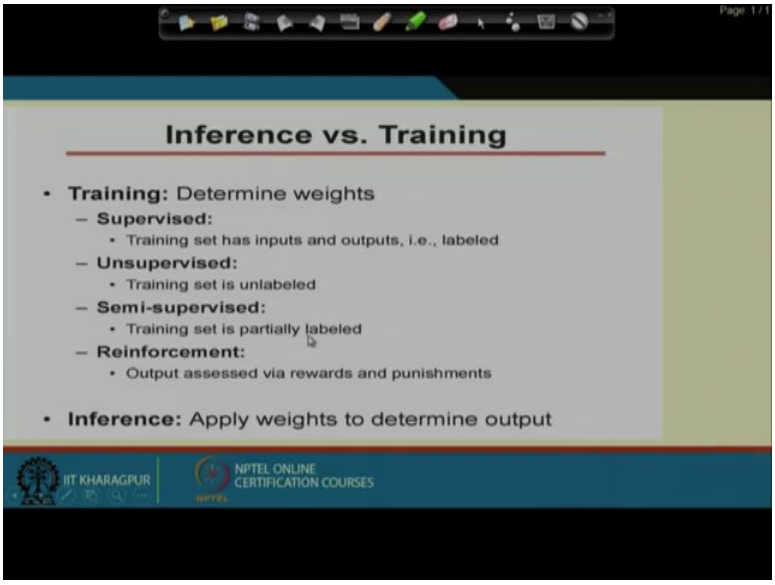
So, what will be the algorithms? So, you will get that information on the deep learning on the machine learning coursework. But, here we are just trying to give you the overview of the hardware, how the hardware will look like or what are the challenges whenever you will design the hardware for machine learning.

(Refer Slide Time: 28:00)



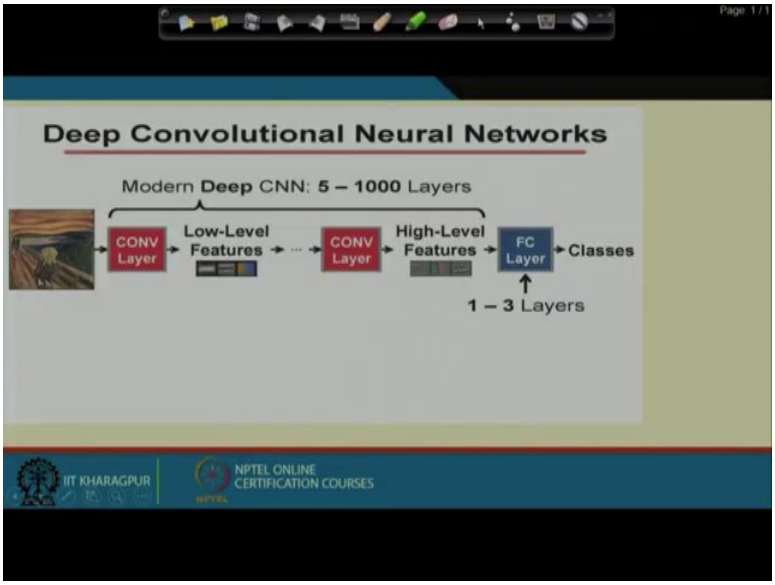
So, there are several deep neural networks among them they are fully connected neural network, convolutional neural network, recurrent neural network and long short term memory, this LSTM neural type neural network.

(Refer Slide Time: 28:15)



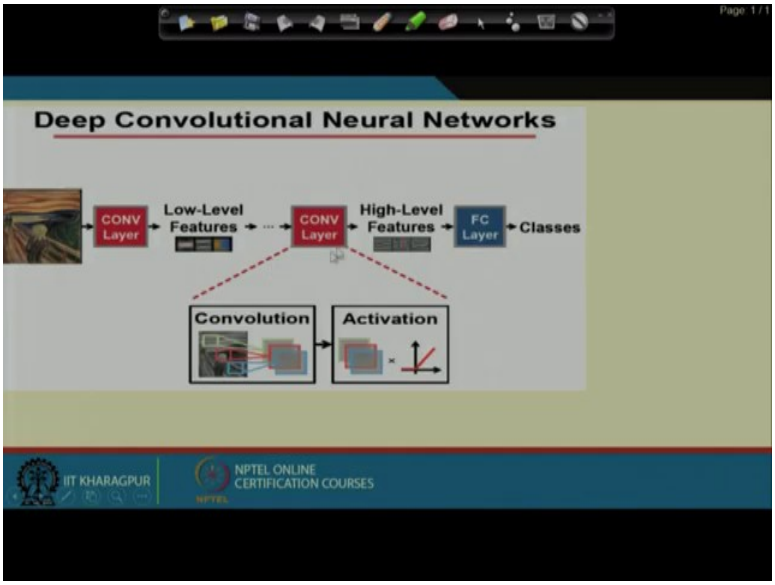
So, you can train your model in using; that means, by different actually technique one is that supervise, another one is unsupervised, then semi supervised and then reinforcements. So, all of these are the; that means, training techniques of your neural networks.

(Refer Slide Time: 28:40)



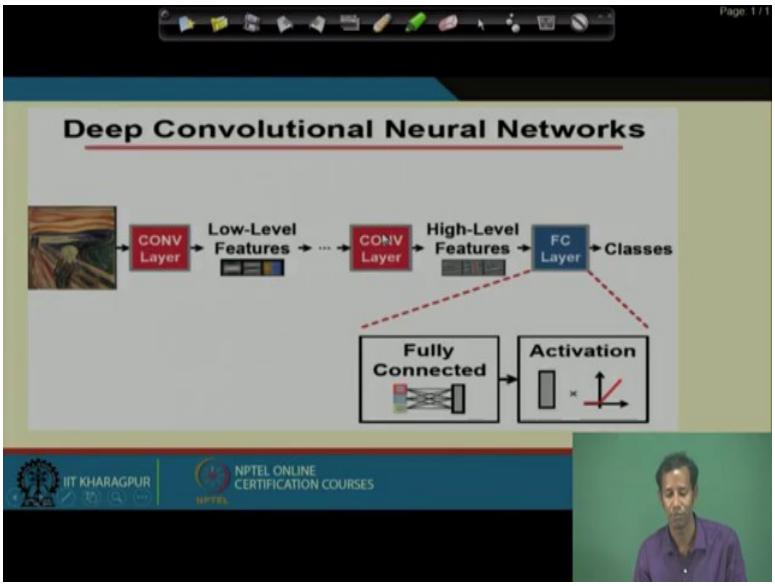
So, this is one of the examples that suppose, this is the one i; that means, image. Now, one this model modern deep convolutional neural network has 5 to 1000 layers. So, in the low actually in the first layer it detects the low level features. And, then if we go on the higher layers actually, it detects the high level features ok. And, then they sum up and then it goes to the image classification or image detection or image or the object detection or for properly recognizes that which object is this. So, it go take the decision and goes there.

(Refer Slide Time: 29:24)



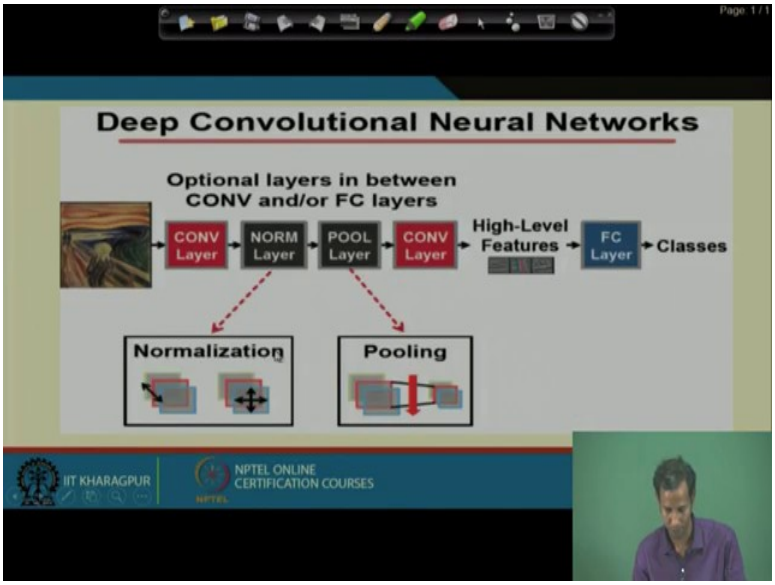
So, this each of this convolution and; that means, layer they are having this convolution operation and then they have this activation operation, which are done inside of this convolutional layer to find out the features, whether that is in low level or they are in high level.

(Refer Slide Time: 29:36)



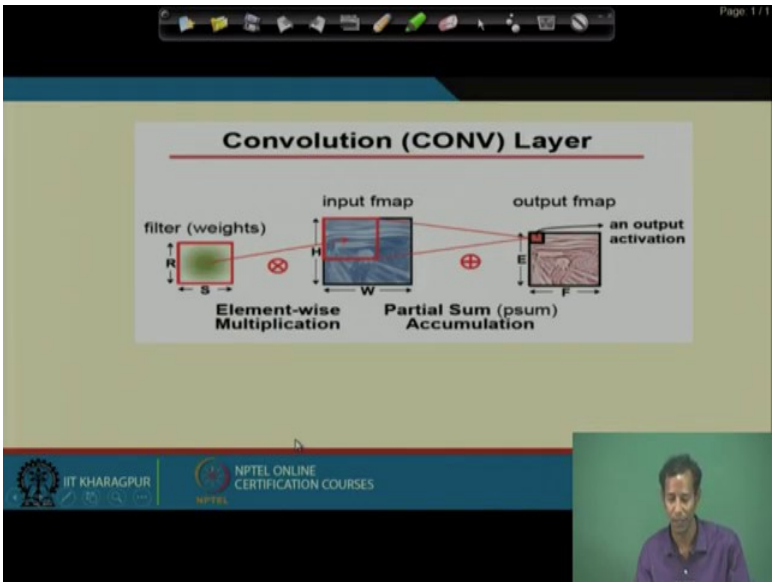
So, in this fully connected layer, we are having this fully connected all the synapses. That means, all these with different weight and then we are having this activation.

(Refer Slide Time: 29:49)



So, inside of them this convolutional layer we can we are having this norm layer or pool layer. In the norm layer, we are just having this normalization and then we do some pooling kind of thing.

(Refer Slide Time: 30:14)

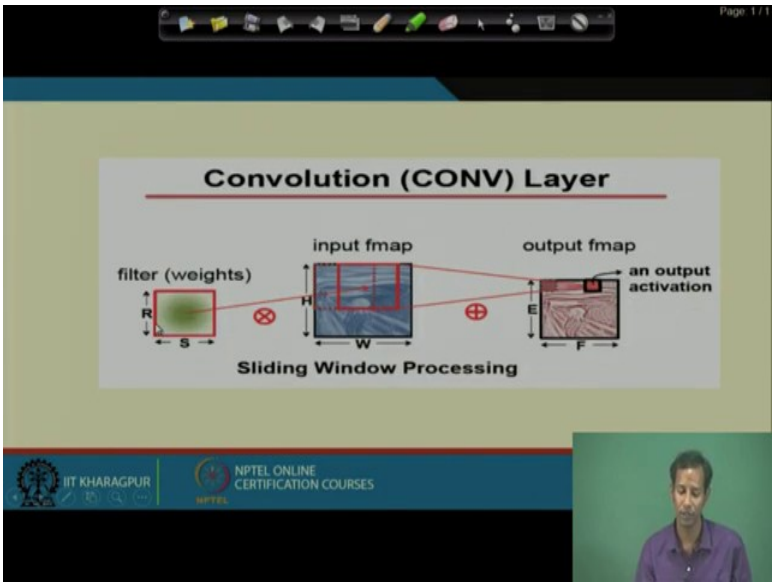


So, these are the techniques which basically used in deep neural network so, then if I consider how this convolutional layer works is that? Suppose, we are having one particular of this is the filter, now this is the input image. Now, I will run this filter in the overall image to find out the element wise. That means, whenever we will run this convolution to find out so, the similarity. Convolution means, between these two particular picture how many similarity I find.

So, if this is my training module that ok. So, this is the information about the classification or the features about the bird or the features about the dog or the features about the cat. So, then if I have the input image so now, if I run that whenever, I will find that this object this features are matching. So, at the time I can detect that this is the object corresponds to one bird, this object correspond to one dog, this corresponds to one cow so, something like that. So, in hardware how we can do the convolution and layer of; that means, I we have to do element wise multiplication ok.

So, then actually these particular things are stored in one partial sum accumulation ok.

(Refer Slide Time: 31:31)



And, how actually this is like sliding window processing. So, this filter first it is done with this portion of the image then, again it slides then, again if with the other image. So, just like the sliding window processing we do and we can detects the object according if there is any match, depending on this convolution operation.

So, thank you for today again, we will discuss what are there in this artificial intelligence hardware development that we will see in the next lecture.