

Deep Learning for Visual Computing
Prof. Debdoot Sheet
Department of Electrical Engineering
Indian Institute of Technology, Kharagpur



Lecture – 06
Introduction to Deep Learning with Neural Networks

Welcome. So, this is our week two and this is where we get started with understanding on very basic and preliminary concepts of deep learning, and that is what we call as deep learning with neural networks. So, as you know that in the first week what we have learned is on the basics and classical ways of visual ways of computing, and where we were getting down images and then relating each image to features itself. And as we relate down images to it is features and these which are more compact representations of how images are represented and from there we get down to something called as a classification problem or associating a categorical label to it.

And we have done it through the whole classical way which is get down an image on the labs side you had also learned on how to code down and extract on certain features which we had studied in the previous preceding lectures, and then subsequently going down and using a neural network for classification purposes. So, a very simple perceptron model, and then using it to classify and that that is together completing it to the lab.

Now, today what we entered into is the first introductory lecture itself and as a lecture is known it is called as introduction to deep learning with neural networks, and what happens within this introduction to deep learning with neural networks is of the way that, we would be starting down understanding as to what we have as a relation between these neural networks, and why this word has deep comes into it, but before starting down any of these aspects over there the first introduction which we need to have very clear in our minds is about what do we define as something called as learning.

(Refer Slide Time: 02:06)



NPTEL ONLINE
CERTIFICATION COURSES
Indian Institute of Technology Kharagpur | Department of Electrical Engineering

Learning?

A computer program is said to **learn** from **experience E** with respect to some class of **tasks T** and performance **measure P**, if its performance at tasks in **T**, as measured by **P**, improves with experience **E**

-Tom Mitchell

Introduction to Deep Learning with Neural Networks [Debdoot Sheel] 2

Now, if you go down by the very classical definition of learning by professor Tom Mitchell. So, you will get down that case is classical textbook on machine learning is what outlines it out, and the outline is something like this that a computer program is said to learn from certain experience E with respect to a certain class of task T and a performance measure P. So, if you see there are three attributes to this activity of what is called as learning or what he calls has learned. So, there is one factor which is called as an experience E there is a particular task which it has to perform T and there is a performance measure P.

So, if you go down through our earlier first week lectures and that experience you would get done, that what happens is that we were trying to do up one particular task of image classification right. So, we are not doing any other task. So, this this was like if I want to just find out whether there is a ball in the image or there is not a ball then that is the only task I am trying to solve over there.

Now, in order to solve this task I was gaining certain experience E, and that is by looking into multiple number of images. So, the more the images I looked through the more is the experience the more the number of epochs over, which I translate the more is the experience over there. Now from this basis of T and E I was using a certain measure called a P and this performance measure P which I was using over here, there is something which is my error function.

So, that was the cost function which I was using down over there, now as you see that as my experience was increasing which is my number of epochs over which I was translating and the number of images, which I was looking down over there my error was coming down and; that means, that I am somehow able to measure my performance and see that the performance is increasing. So, as the performance increases it becomes more and more accurate and accordingly my error keeps on going down ok.

So, that is what is saying that if it is it said to learn if this performance on a task T which is of my classification as measured by this performance P improves with experience. And now this definition if you really try to introspect onto this one, this definition is in no way very different from how human learning is all centered around. In fact, human learning is also quite similar there also, as human beings when we say that we are learning about something then the whole task of learning is when we are able to really getting more.

And as we get more and more experience and that is what practice makes I mean man perfect so, with that one. So, you get more experience and then your measure becomes higher and higher with respect to a certain class of task itself. So, that is what goes down by the very classical standard definition. Now once we have been able to define that one let us look into trying to demystify what this would mean.

(Refer Slide Time: 04:58)

The slide features a header with the NPTEL logo and text: 'NPTEL ONLINE CERTIFICATION COURSES Indian Institute of Technology Kharagpur | Department of Electrical Engineering'. The main title is 'Demystifying Learning'. Below the title is a photograph of four men standing in front of a Great Wall logo and tower. The men are labeled 'Man 1' (Debdoot), 'Man 2' (Kim), 'Man 3' (Jung), and 'Man 4' (Wang). A red dashed line outlines the Great Wall tower, and a blue dashed line outlines the Great Wall logo. To the right of the photo is a graph with 'Performance (P)' on the vertical axis and 'Experience (E)' on the horizontal axis. The graph shows a blue curve that starts at the origin and increases with a positive slope, representing the relationship between experience and performance.

Debdoot, Kim, Jung and Wang are standing near the Great Wall logo and the Great Wall tower is behind them.

Introduction to Deep Learning with Neural Networks [Debdoot Sheet] 3

Now, let us get down with a very basic problem and that basic problem is that I have an image, and now I would like to give a word equivalent description of this image. So, what; that means, is say that this is an image and now a lot of you upload your images on Instagram on Facebook and all other social networking sites and anyway. Now the point is what you are doing by doing all of these image sharing over there is just sharing some sort of an experience over there. So that people get to know where, have you been, what have you seen, what have you learned whatever you experience, and then just sharing it across. So, that the more you share the more collective knowledge is what humankind gains over there.

So, if this is the image which is given down, and now the question is that, say that is a blind person who is not able to see things. Now you would like that blind person will still on your friend list or maybe not even there I mean this this blind person to somehow know, and contemplate on your own experience itself.

Now, for that what would happen is that the blind person has to somehow be able to understand what is there in the image, and that necessitates an action something like a image to text conversion in order to get down whatever it is annotated and present in this image so; that means, that if this image is given can I have a text equivalent of this one or a very simple thing which is called as an image captioning problem as of today.

So, what will happen is something like this that as you would see that initially a computer program it if it is a so it will be doing this sort of like divide something into some blocks. So, using these blocks over here it will know that these are the blocks, which are related to somewhat and it can understand that if it can identify these blocks independently then it can; obviously, give you a particular reasoning over what this is. So, what makes it would do is that each of these blocks it will start annotating each of these blocks.

Now, if you look into that curve on the side over there what you would see is that as it is able to understand and recognize each block itself. So, you see that your experience is gaining, and then your performance is also increasing along that one. Now going down through that one what happens is that in the next instant that it will be able to identify some more objects over there and they are those faces which you see, and eventually it goes down it identifies, what is the Great War logo, and there is a Great Wall tower.

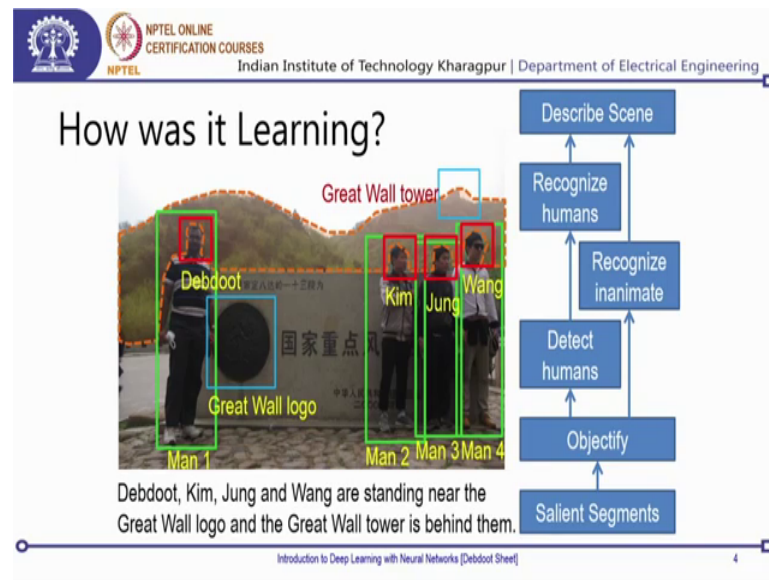
And finally, it can identify these different people if it has a corpus of all the faces over that together, and then write down a sentence equivalent of whatever it has identified over here. Now the interesting aspect which happens is if you see, that it is able to come down from an image why all of this segmentation and recognition, and identification, and then to a final sentence equivalent description through getting a lot of experience.

So, that necessarily means that it is not a one shot job. So, the whole computer over here took a lot of time itself in order to get down to that experience now, as it took all of this a good amount of time coming down over there. So, what happens is that over as it gains experience it also gains its performance, and the best performance is when you are able to get done a sentence equivalent. So, initially when you are just able to give down block over there so, it will say that the image has multiple number of fragments say some 7 7 or 8 fragments in which you can divide it out.

Now, within each of these blocks which are fragmented blocks say there are 4 of them are human beings, and there are there is a mountain and there is some sort of a logo kind of thing, but they do not match closely to the sentence. So, there is a lot of error. So, finally, when you go closely on the sentence which says as Kim, Debdoot Kim, Jung and Wang standing near the Great Wall logo at the Great Wall tower is behind them. This is exactly what this closed to the sentence and you get the best performance error coming down over here.

Now, that we know that this is what essentially we meant down, when we were saying down that it is learning something. The next objective is to understand what was it learning and how was it learning more than what it is actually how does it actually go on to learn this one. So, let us again get back over there so, as you see in the whole image you would be getting down the image first. And then the first objective over there is to break it down into some number of segments right.

(Refer Slide Time: 09:28)



So, these are the different segments to which it comes and let us say that this is breaking down an image into it is salient segments. Now once you have broken down the image into salient segments. The next part is that you are going to identify some of these segments or what is also called as an objectification task, now once you are able to objectify these blocks over there then the next part is trying to identify some of the detect whether there are humans or not, so you find out that they have a certain number of humans over there.

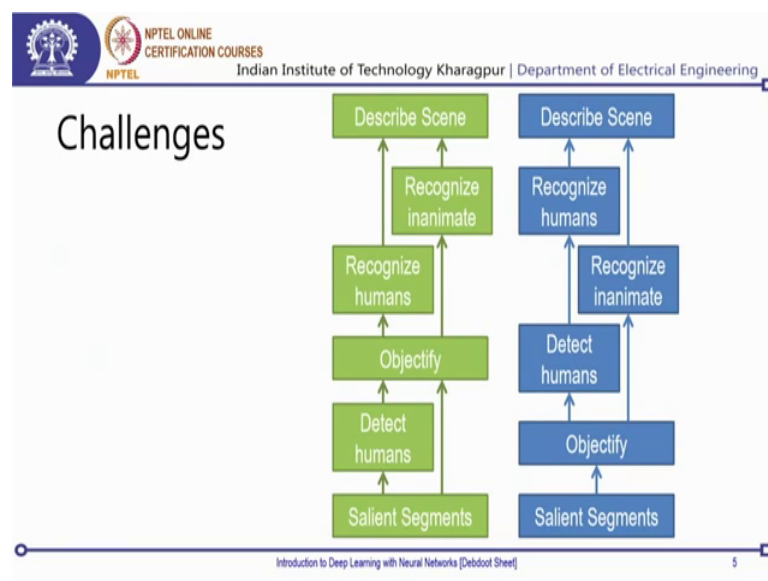
Then the machine is able to find out that there are certain number of inanimate objects as well. Now humans are in any way targets are different one story now once it knows that there are humans over there it will try to recognize humans find out who is who actually which portion is over there on the image. So, there was a human figure, but that each person is that human who is being shown over here. Now once we are able to do all of this then you can describe a whole scene coming down from that one.

And this is essentially what this machine is able to do, but the question is even bigger, the question is that we know that how it was learning was by doing something of this sort and the deeper it keeps on going. So, that is over the hierarchy as it keeps on going which is it starts with the base of the pyramid which is on the salient segment, and then it keeps on climbing up to the description of the scene. So, as it keeps on growing up that hierarchy this is where the depth of the whole learning comes into it.

Now, the aspect of deep learning says that as it is able to go down. So, it is; obviously, gaining this depth by gaining looking at more number of images getting down more and more experience, and accordingly its performance is increasing. Now the point is that you see some sort of a hierarchical nature in which this whole recognition task is based. Now is this hierarchical nature really unique or is it non-unique it is the major question which we have as of now.

So, I would give you a few seconds to actually ponder on this one whether it is unique or not do you think there can be a non-unique way of solving this problem as well, instead of this being the only possible way of solving the problem. So, that brings us to an actual pertinent challenge, as it turns out that this is not a very unique way of solving the problem and there can be multiple ways of it.

(Refer Slide Time: 11:54)



Now, let us look into a thing let us make a parallel universe in which we are able to create a different rationally different model over there. So, initially what I would do is I would make a replica of this itself. Now, let me just remove certain of these connections and reorder the blocks itself, now once I have reordered I can connect them, and if you carefully look over here the number of blocks as well as the objective of each of these blocks is still the same whereas, the order in which the blocks were connected somehow got changed over here.

Now if you carefully introspect on this one you would see that within these blocks it is still the same. So, there is not any change coming down over here as you look into over here, what happens is that you can still put down image it will break it down into salient segments in the earlier case we were objectifying first, but here it detects humans in the segments then goes on to objectify what they are and then goes on to recognize.

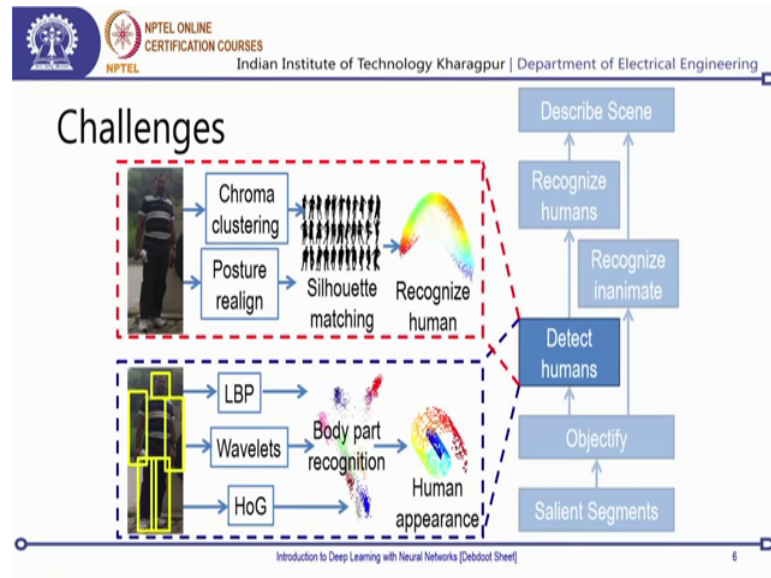
And some of you can even say that we can pull down this recognizer below the object if I then, push the object away above then, have the inanimate things over there and that is perfectly fine I mean that is also another possibility of doing it. So, as you see what happens is essentially it turns out that there is no unique way of solving this problem and that is not a very happy thing to be here.

So, as researchers for us it is a very interesting point, because we know that there can be multiple ways of solving a problem. So, it means that every time I find out one way of solving the problem I get down a paper route and there can be a PHD theses out of it know, then from a product development perspective it is really a very dicey situation. Because if you have non unique ways of solving a problem; that means, you will have to explode down each and every possibility of solving out that problem and find out, which is the best possible solution in order to achieve a solution to this problem.

This possible method to achieve a solution to that problem, and now this this is it the only way of doing and this problem, this challenge which we have over here that there can be non-unique solutions, and we can always keep on proposing yet another way of solving this problem and yet more another way of. In fact, there are certain interesting papers which do come out in conferences which called as yet another way of solving this problem.

So, in terms of an industry this is really a major issue because over there as you see these non-unique ways of solving the problems creep up, you would see that these are bigger challenges for industry itself. And now one was that this arrangement of these blocks over there that was becoming as non-unique, but then the point is that the only issue which comes out or can there be some other ways of doing it as well, so as it turns out this is not the only challenge which you face.

(Refer Slide Time: 15:09)



In fact, if we consider one of these blocks say detect humans, that can have one way in which what we can do is I get down the salient segments over there, and then say that objectified the saying that there is some sort of object and let us see whether there is a human over there or not. So, we take down all segments over there from the objectify block, and then what we do is we go around very extracting some features on this small block itself.

So, that should be enough to say whether there is so, this is a simple image classification problem human or not a human. Now I can find features like wavelets LBPs and histogram of oriented gradients these are the ones which we have already studied, and then what we do is from there using each of them we can do some sort of a body part recognition or run down one classifier which can identify, which parts over here. So, in this patch we can divide it into number of small segments and say which segment is what.

So, I can identify head leg hands over there, and this is a body part recognition. Now once I have my body part recognition what you can do is between these body parts I can draw down lines and find out what are the distance relationship between these parts, and now for a standard human there would be a range of these distance relationships the angles in which they vary and then using that these measures I can again run another

classifier and find out whether this appearance matches down that of a human or not because see there can be other bipedal animals which may not be human.



So, there can be an orangutan, there can be say even a kangaroo which are bipeds as such. Now, but then the ratio of length of their hands to legs they are very pretty different over there, then the posture because I they do not always stand upright there can be a bear which is standing upright, but then the ratios are different the posture is different, the distance between the legs, and the hands are different the angles at which these things are connected they are also pretty different. And that is what makes it easy to classify now using this whole thing I can find out that there is a human.

Apparently it turns out this is not dears the unique way of doing what I can do is I can have another way, in which I can take down this segment, and then on the segment what I can do is run down a (Refer Time: 17:17) romo clustering and a posterior line. So, as a human I can be standing upright I can lean down I can have twisted views any of these and Chroma clustering is what will help you find out which is the background and which is the foreground.

Now, once you have all of this together you can do some sort of a silhouette matching, which is like if human beings are present over there in black and whites this is what the shadows or outlines would look like now. So, a cat dog or monkey or a goat and their solutes will be very different from that one, now I can run down some sort of a distance algorithm on top of it and do a classification task which is to get down my humans over there and recognize humans.

So, as you see over here also we have two different ways and 2 non unique ways of detecting humans, and as it turns out that you can have some infinite number of non-unique ways of detecting humans itself. And this is not such a easy thing this is actually raises to a huge dilemma within practice and that dilemma does not just exist in the field of machine vision.

(Refer Slide Time: 18:24)



NPTEL ONLINE
CERTIFICATION COURSES
Indian Institute of Technology Kharagpur | Department of Electrical Engineering

Dilemma only in Machine Vision?

- **Speech Recognition and Signal Processing**
 - Dahl, et al.(2012). Context dependent pre-trained deep neural networks for large vocabulary speech recognition. *IEEE Trans. Audio, Speech, and Language Processing*, 20(1), 33-42.
- **Handwriting Recognition**
 - Bengio, et al (2006). Greedy layer-wise training of deep networks. In *NIPS 2006*.
- **Natural Language Processing**
 - Hinton, (1986). Learning distributed representations of concepts. In *Proc. 8th Conf. Cog. Sc. Society*, pp. 1-12.
- **Hierarchical and Transfer Learning**
 - Goodfellow, et al. (2011). Spike-and slab sparse coding for unsupervised feature discovery. In *NIPS 2011 Workshop on Challenges in Learning Hierarchical Models*.
- **Medical Imaging**
 - Karri and Sheet, et al (2014). Deep learnt random forests for segmentation of retinal layers in optical coherence tomography images. In *ISBI 2014*.

Introduction to Deep Learning with Neural Networks [Debdoot Sheet] 7

So, this dilemma is there in the field of speech and signal processing as well. So, where what can be a way of recognizing speech and apparently it turns out that there is no unique way of doing it.

It exists in the field of handwritten digit recognition as well. So, if you are writing down something then how can we identify it exists in the field of language natural language processing which is from your sentences, can you make inferences out of it or say today what you write down on Google you do not anymore put down some keywords to query down, you ask a full question as if you are asking.

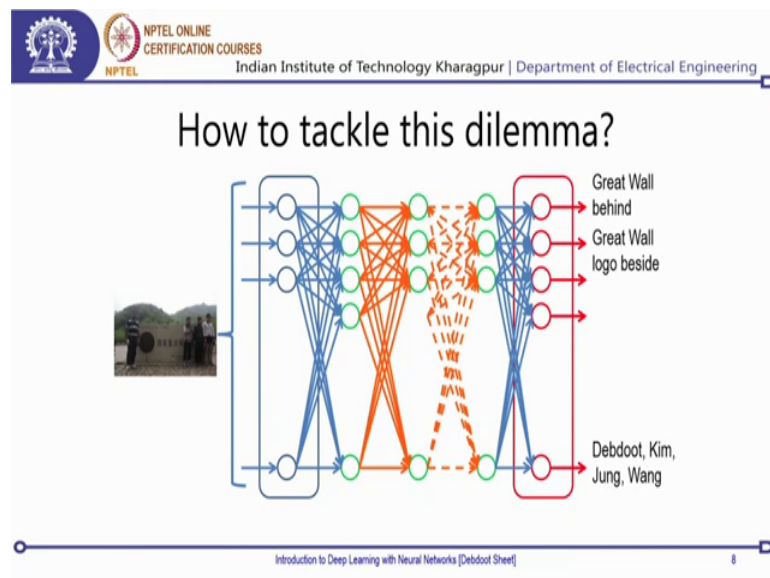
So, a human being in front of you, what is the weather today will it rain today, so it has to bring break it down into computer readable instructions. So, if there is if you ask what is the weather today, so it knows that it has to put down get todays date, and generate a query to our website on weather may be weather dot com, and then send out todays date you are location over here and ask this weather dot com to give a feedback. And then it will have to reconstruct a sentence back onto it and that that is about natural language processing.

Now, from there are interesting problems on hierarchical and transfer learning as well and so, we would eventually go down a bit later on into what this transfer learning and hierarchical learning is all about, and it does exist in the field of medical imaging and image analysis itself. So, these non-uniqueness dilemma while it is a interesting avenue

and scope for researchers was for a longer duration of time, but today if you see with the advent of deep learning over here.

So, this this avenue is sort of closed and what we come down to is let us come down to the most consistent solution available, by discovering the solution through a learning method, and eventually using all of this discovered solution can be make a analysis and can be say or what is called as the explained ability of this learnt model, says what is more of a research challenge today.

(Refer Slide Time: 20:30)



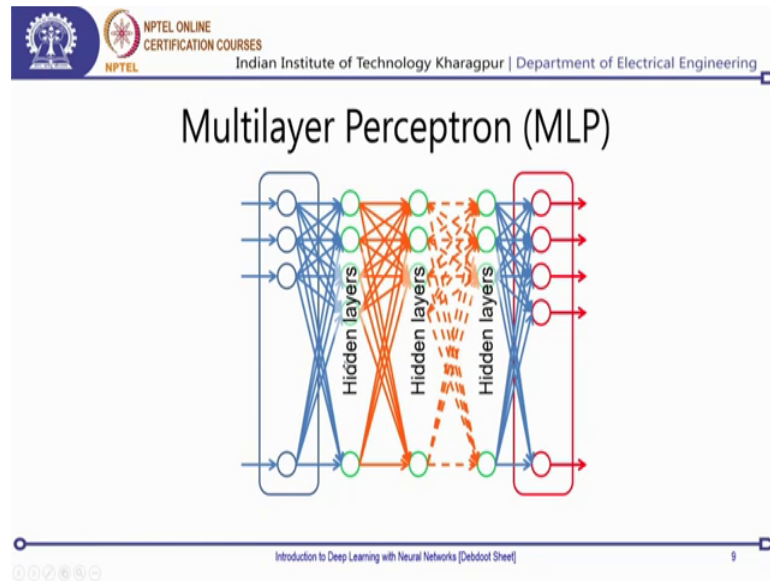
But then going down that we have these dilemmas then the objective is how do we tackle this dilemma, and for that what is done is something of that sort, so say you have this image captioning problem over there. So, what I can do is I can take an image I can organize all the pixels of the image into a vector, and then I can subsequently keep on connecting these through subsequent nodes over there. And now finally what it would do is that there is it would generate some sort of an output which would say that there is a great wall behind, and there is a great wall logo beside, and there is these there are these 4 people over there.

And if you look into this one what this does is this is some sort of a network like architecture which says that you give in all the pixels and from pixels it will translate to some alternate representations by clubbing all the pixels together into one representation

than another, and then subsequently as it goes down it follows down a hierarchy and finally, comes down to a classification or associating certain labels over there.

Now, carefully getting back this model actually very closely represents into what is called as a multi-layer perceptron.

(Refer Slide Time: 21:35)

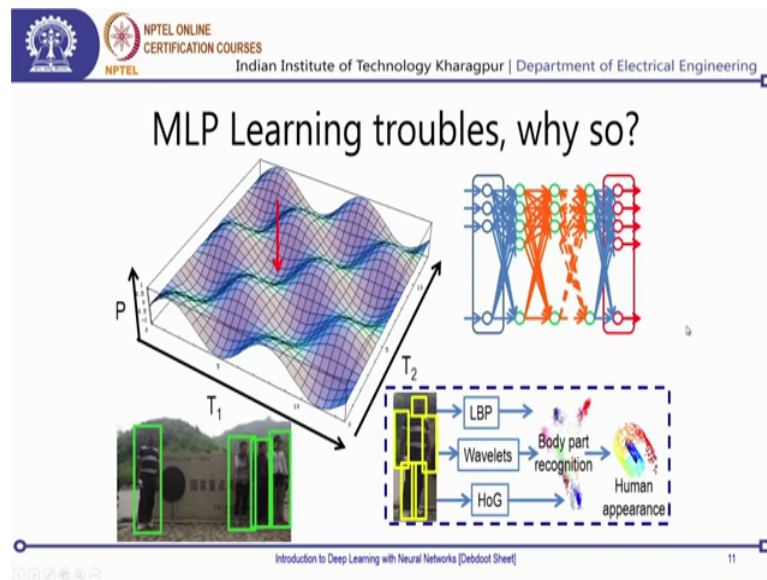


So, within a multi-layer perceptron what happens so, this is what we have learned in the earlier lecture on simple neural networks. So, we will get down into exactly what how the mathematics of multi-layer perceptron is handled, but before that within this what happens is that you get down all the input pixels over here, it will translate via a multiple number of neuron layers.

And each of this is what is called as a hidden layer the reason it is hidden is that it is not hidden from the coat, but it says because you do not see any output coming out of these layers, like it is its no target output which comes out the target output only comes out of this last layer which is also called as output layer.

So, these output layer and the input layer to which you give an input and you draw an output from is what is called as the visible layers, and inside all of these intermediate layer through which the whole operation and mathematics project is what are called as the hidden layers over there.

(Refer Slide Time: 22:29)



Now as you get a multi-layer perceptron what comes down is that you will also have to train a multi-layer perceptron. Now, in order to train a multi-layer perceptron let us have that same kind of analogy as we had done in the first week on the lecture on neural networks. So, you have your weight space and you have your P or performance or your cost function $J W$ which you had done.

Now it does have the same troubles over here as you observed over there that you can get done multiple number of minima s and maxima s while the problem well the whole objective over, there was that say you start down at some random point over there, and then you had to gradient descent and learn and come down to a minimum.

But my point is that every time you have a different combination of initial weights given down you would be near a different minimum position, and that is where the challenge is now apparently it turns out that this is not a very weird kind of a thing there is a whole reasoning, and why as to why. So, do you have that kind of a behavior over there, so what turns out is that if you look into this cost function curve, and then say take down one of these points what comes out does if you can get an explanation out of what this deep neural network was doing in a hierarchical way.

You would get down one of these kind of models which comes down which is taken image, then get it is features extracted on each of the small segments within that image and then find out a body reactor (Refer Time: 23:48). So, this essentially is where you

have it going down in a hierarchy so, you have to complete this layer of wavelets plus LBP s plus HoG, and then only you can go down to the layer of body part recognition, and once this layer of body part recognition is done only then you can go down to the layer of human record appearance module. And that is what will associate itself to one of these bits over there.

(Refer Slide Time: 24:12)

The slide, titled "MLP Learning troubles, why so?", is part of an NPTEL online certification course from IIT Kharagpur. It features a 3D surface plot with axes labeled P, T₁, and T₂, showing a complex, wavy surface with a red arrow pointing to a specific point. To the right is a diagram of a multi-layer perceptron (MLP) neural network with three layers of nodes. Below the plot is a small image of a person in a field with green bounding boxes around them, labeled T₁. To the right of this is a dashed red box containing a flowchart: "Chroma clustering" and "Posture realign" lead to "Silhouette matching", which then leads to "Recognize human". The slide footer includes the text "Introduction to Deep Learning with Neural Networks [Debdoot Sheel]" and the number "11".

Now, as it turns out if you can go down to a different initialization you will have a different model or doing it up. And so, on and so forth it turns out that for every single peak location every single different kind of a unique minima coming down over there, you will get down a different sort of a model, and this is the major reason why you really have a trouble or a major issue in order to analyze and emphasize, and explain these multi-layer perceptron's.

So, while we have done this one this lecture does come to an end over here now in the in the subsequent lecture what we will be doing is we will so here, I have just told you about the problem, and then the issues which come down with these kind of models of multiple layers. So, in the next one I will be discussing about what are the different kind of layers, there can be and subsequently we will enter eventually into the math of trying to solve out in order to get rid of this kind of problems.

So, with that wait and watch for the next a subsequent one on the history of deep learning, and learning with deep neural networks and till then bye.