

Deep Learning
Prof. Prabir Kumar Biswas
Department of Electronics And Electrical Communication Engineering
Indian Institute of Technology, Kharagpur

Lecture – 01
Introduction

Hello. Welcome to the NPTEL certification course on Deep Learning. So, in today we are going to introduce the content of this course and we are going to talk about that what all we will be covering in this lecture series on Deep Learning.

(Refer Slide Time: 00:49)

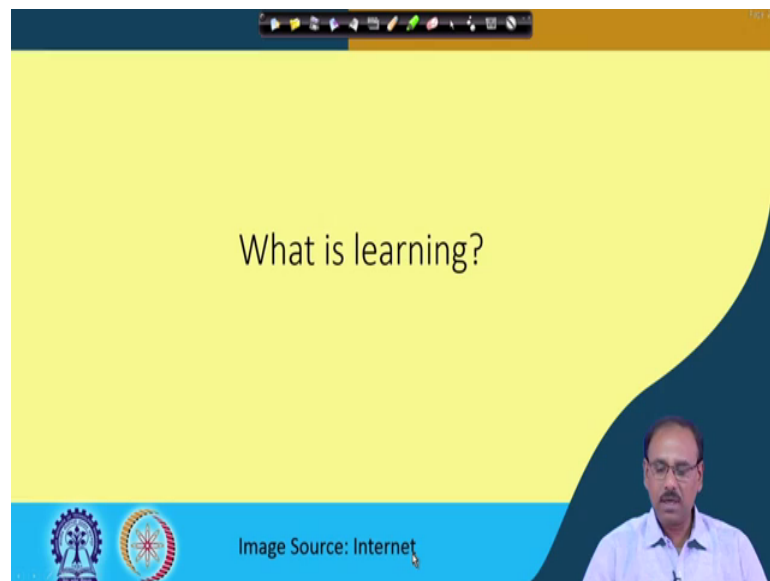


So, the topics that I will covered is; obviously, the first one is an introduction to deep learning, then we will talk about that when we learn something how do we learn, in the sense that we learned certain properties or certain features or certain descriptors using which we recognize an event or we recognize an object. So, we are going to have a brief introduction to what are descriptors or what are feature vectors.

Following that we will talk about what is machine learning and what is deep learning. Machine learning is the traditional approach and how we are deferring when we are going for deep learning, what is the difference between machine learning and deep learning. Then coming to deep learning again we will talk about two different models of deep learning, one of them is in the discriminative model and the other one is the generative model.

Then we will see that what are the challenges of deep learning applications when you want to have any application of the deep learning techniques, what are the challenges that we face and how we try to mitigate those challenges and then we will briefly talk about that what is the power of deep learning techniques or what we can do using deep learning methods.

(Refer Slide Time: 02:19)



So, first let us try to see that what is learning, or then we will come to what is machine learning.

(Refer Slide Time: 02:29)



So, to talk about what is learning I will show you these two pictures or you take any picture for that matter or even a word say father. So, coming to these two pictures, I will simply ask you can you recognize these two pictures. The answer will be obviously, yes.

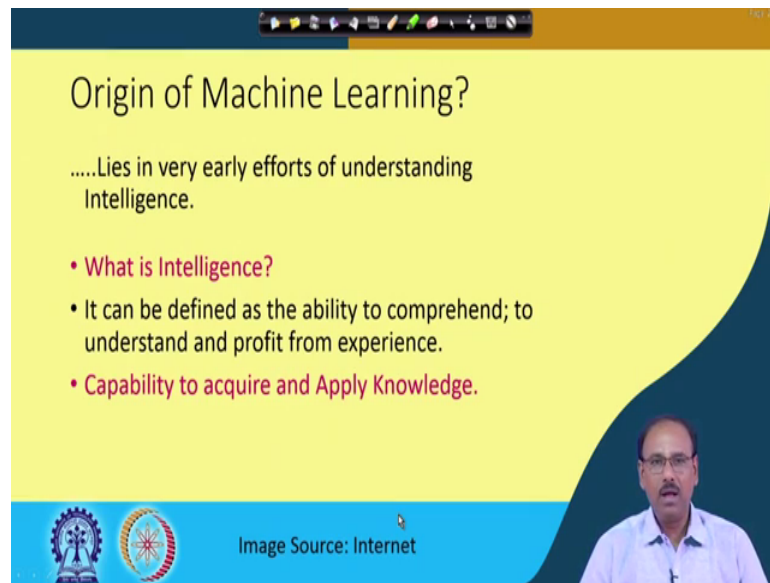
All of you hopefully will say that the picture on the left is a cave painting in Ajanta caves and picture on the right is the picture of parliament building in New Delhi, ok. So, we have been able to recognize these two pictures or will be able to recognize thousands of such pictures which are shown to us or we will be able to understand thousands of words which we will hear or thousands of sentences we will hear.

But the question is we have been able to understand we or we have been able to recognize these two pictures that is fine, but the question is how do you recognize it. So, you find that when we looked at any of these two pictures; obviously, how do I recognize that the picture on the left is Ajanta is a painting from Ajanta caves, the reason is either my parents, my friends, they have shown me that these are the paintings from Ajanta cave. If not, maybe I have visited Ajanta cave and there I have seen this painting or maybe I have seen these paintings in text books, in history books, right. In most of the history books such paintings, so images of such paintings are abundant.

So, I have while seeing those pictures, I have unknowingly, unintentionally try to capture certain properties or certain descriptors from this picture. And using those properties or those pictures I have built a model which is embedded in my brain. So, next time whenever this painting is shown to me, I try to find out similar descriptors or similar features and I try to match those features with the model that I have been that is embedded in my brain; that means, this picture is associated with an a priori knowledge that I have, right. So, using that a priori knowledge I try to recognize or I can recognize this picture very easily.

But suppose the picture on the right that has never been shown to me, I have never seen parliament building or have never seen an image of parliament building, nobody has shown me, nobody has told me. So, if this picture on the, right is shown to me I will probably say, ok, this is some building because I know the buildings look like this, but I will not be able to say that this is parliament building because that association I do not have. So, how do you recognize or how do you get the description or the features of any event or any object or any picture we will try to see through this course.

(Refer Slide Time: 05:49)



Origin of Machine Learning?

.....Lies in very early efforts of understanding Intelligence.

- **What is Intelligence?**
- It can be defined as the ability to comprehend; to understand and profit from experience.
- **Capability to acquire and Apply Knowledge.**

Image Source: Internet

So, again come back to learning, let us see what is the origin of machine learning. The origin of machine learning or the origin of learning as such is nothing new this lies in very early efforts in trying to understand what is intelligence. So, an intelligence can be defined as that it is the ability to comprehend to understand and profit from experience.

So, you find that this definition is very very crisp by definition. And these three words comprehend, understand and profit from experience these three lies at the heart of what is machine learning or what is deep learning. Or in other words we can also say that the intelligence is the capability to acquire and apply knowledge. So, we are experiencing various events every day; each and every day we are watching new objects every day and through that we are acquiring knowledge. And then what we are trying to do is we are applying the knowledge that we have acquired through experience and that is what is intelligence.

(Refer Slide Time: 07:13)

The slide has a yellow background with a dark blue curved shape on the right side. At the top, there is a navigation bar with various icons. The main text reads 'Learning?' followed by '2300 Years ago....'. Below this, there is a bulleted list. To the right of the text is a small black and white portrait of Plato. At the bottom left, there are two circular logos. At the bottom center, it says 'Image Source: Internet'. In the bottom right corner, there is a small video inset showing a man with glasses and a white shirt.

Learning?

2300 Years ago....

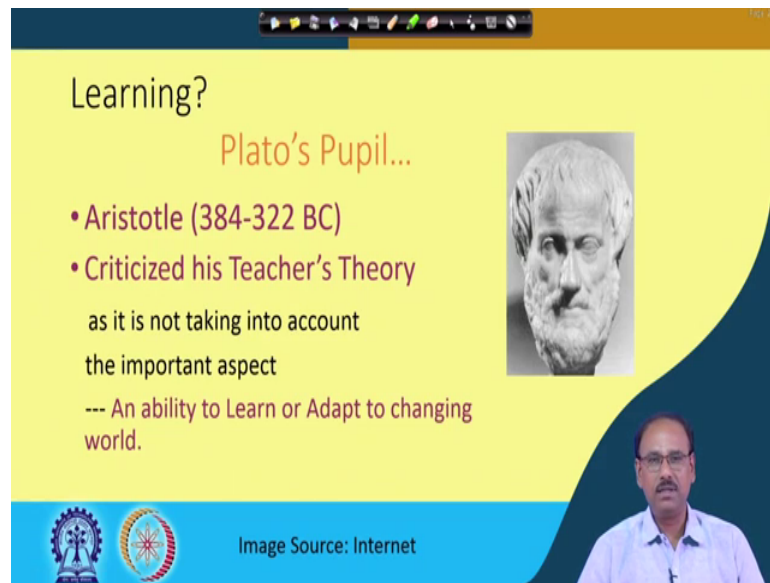
- Plato (427-347 BC)
- The concept of Abstract Ideas are known to us a priori, through a Mystic connection with world.
- He concluded that ability to think is found in *a priori* knowledge of the concepts.

Image Source: Internet

So, coming back it is almost 2300 years ago or even more than, that the Great Philosopher Plato during the who was there during the period 427 to 347 BC. He brought the concept that the abstract ideas are known to us a priori through a mystic connection with the world. So, you note what Plato said that it is a mystic connection with the world so, the abstract ideas are known to us after. And of course, this makes sense because otherwise how is it possible that a newly born baby can easily recognize his or her mother, right. So, definitely it is something mystic connection with the world.

And Plato concluded that ability to think is found in a priori knowledge of the concepts, right. So, everything is as per Plato everything is a priori. So, does it mean that we do not learn anything new; so, that was something which was missing in what Plato said more than 2300 years ago. But soon after it was actually Plato's pupil, Plato student who brought in the concept of learning.

(Refer Slide Time: 08:43)



Learning?

Plato's Pupil...

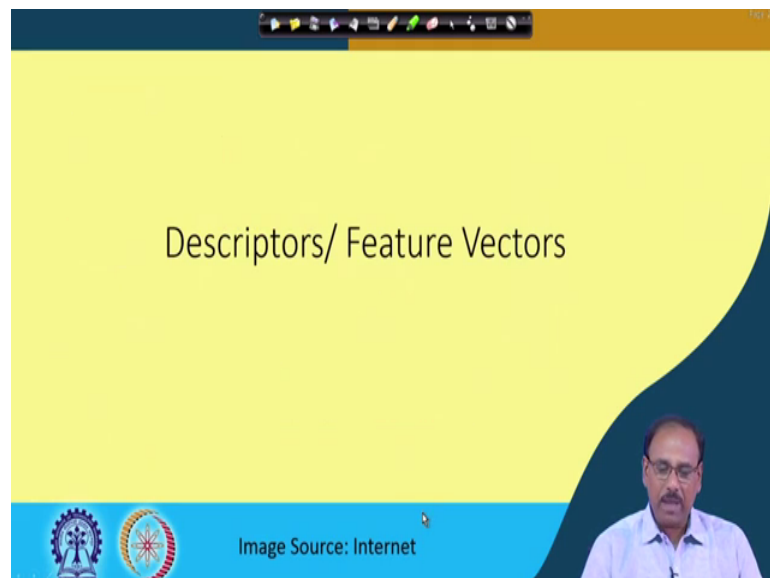
- Aristotle (384-322 BC)
- Criticized his Teacher's Theory
 - as it is not taking into account the important aspect
 - An ability to Learn or Adapt to changing world.

Image Source: Internet

The slide features a yellow background with a dark blue curved shape on the right. At the top, there is a navigation bar with various icons. Below the title, there is a list of bullet points. To the right of the text is a small, square, grayscale image of a man's face, likely Aristotle. At the bottom left, there are two circular logos. At the bottom right, there is a small inset image of a man in a white shirt and glasses, presumably the speaker.

He pointed out; Aristotle pointed out that in Plato's concept a very important aspect was missing that is ability to learn or adapt to changing world. So, all of you know that every day as we experience new and new events, as we see new and new things we are learning continuously. So, learning is a never ending process, every day we are learning something new, ok. So, that is what has been introduced by Plato; that is what has been introduced by Aristotle who was Plato's student.

(Refer Slide Time: 09:41)



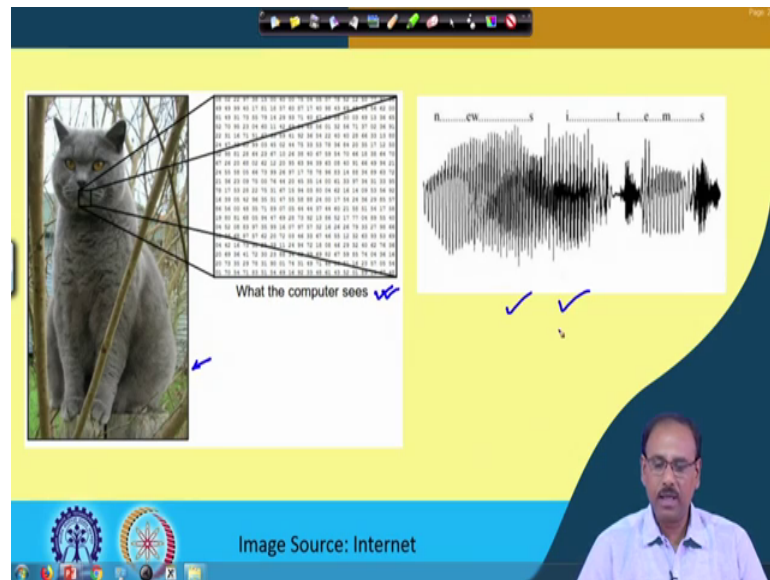
Descriptors/ Feature Vectors

Image Source: Internet

The slide has a yellow background with a dark blue curved shape on the right. At the top, there is a navigation bar with various icons. The title 'Descriptors/ Feature Vectors' is centered on the slide. At the bottom left, there are two circular logos. At the bottom right, there is a small inset image of a man in a white shirt and glasses, presumably the speaker.

So, coming back again to the machine learning or when we try to recognize something why when we learned some object or some animal or some word how do we learn it. Let us come to this.

(Refer Slide Time: 09:57)



So, either we see an object or an image of an object or maybe we listen to some words, to some sentences every day. So, when we talk about the machine learning or deep learning, a very very important application of this deep learning or machine learning is being able to understand or recognize objects that we see or the images that we see or being able to understand or comprehend the sentences that we will hear.

See, when I am someone says go to school I do not I know that what does that sentence mean go to school, so that means, we understand it, ok. So, coming to the machine, when the machine is to learn and the machine is to deliver then; obviously, these sentences or these images has to be converted into a form which the machine will understand.

So, come to a picture which is shown on the left apparently it is a picture of a cat, all these pictures are actually represented by two-dimensional arrays of numerical values and mostly those are integer values. So, those of you who know about the digital images, you know that we talk about pixels, we talk about megapixels 18 megapixel, 40 megapixel, 50 megapixel and so on which are actually the power of the camera that we have.

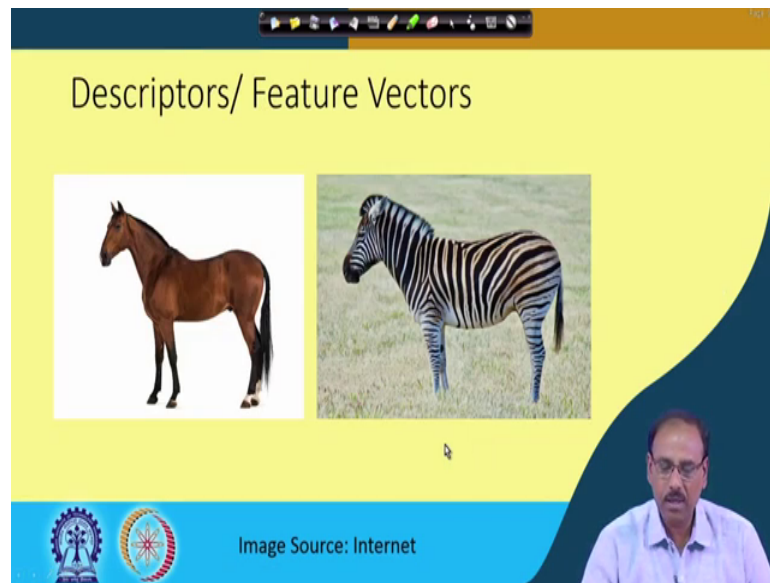
So, these pixels are nothing, but an element in a two-dimensional array or an element in matrix. So, any image is represented as a two-dimensional matrix which is as shown over here. So, this is a two-dimensional matrix which is part of the image which is shown on the left. And every element in this matrix is normally an integer value and you know that these integer values are 8 bit integer values; that means, every element in the image every element in this matrix can assume a value from 0 to 255.

And that is true for a black and white image or a gray level image which does not have any color information. But if I have a color image the color image is represented in two planes, the red plane, green plane and blue plane and each of these planes can be considered as a grayscale image and every pixel in such grayscale images are again 8 bit quantized. That means coming to a color image every pixel in a color image we will have three components, the red component, blue component and green component and each of these components can take values from 0 to 255; that means, every pixel in a colored image is represented by 24 bits, 8 bit per color component. So, that is how an image is represented in a computer.

Similarly, when it comes to word recognition or voice recognition or speech recognition then the speech signals asked utter to be represented digitally. So, what is shown on the right hand side is what you get from the output of a microphone, you know that microphone converts an acoustic signal which is a voice signal or any other sound input into an electrical form and then into an electrical waveform. So, this is in the snapshot of and waveform which is output of a microphone and this will get when you utter a pair of words say news items, then the microphone output will be something like this.

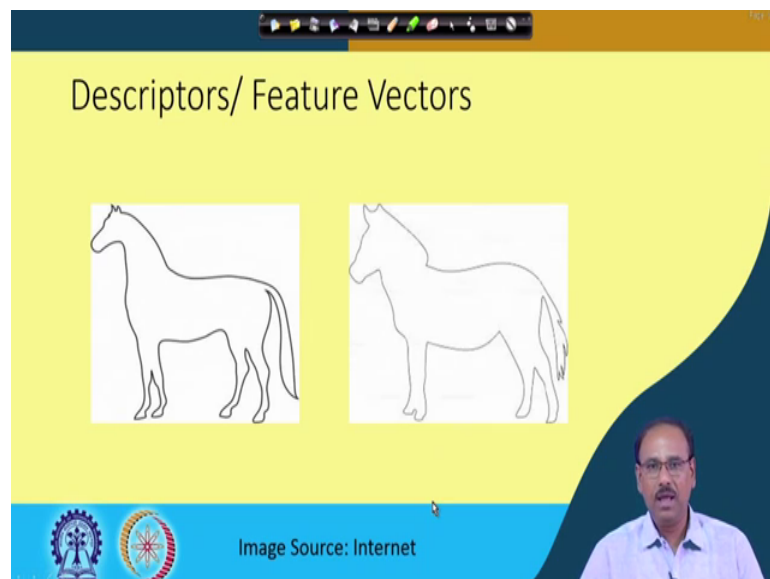
And now, if we sample these outputs and it digitize each of the samples what we get is a time series, a sequence of samples in time, ok. So, this is how I voice signal or a speech signal or a sound signal can be represented in the computer. So, once I have these representations next is how do we process these informations, so that this processed information can be used by computer for understanding or for recognizing.

(Refer Slide Time: 14:29)



So, for that what we need are the descriptors or the feature vectors. So, here I am showing you two different pictures, one is the picture of a horse the other one is the picture of a zebra. Now, given those two pictures you will find that both of them has got two types of properties, one is the shape property otherwise the region property.

(Refer Slide Time: 15:09)



So, what is the shape property? Let us come to the next picture. If I simply take the boundary of a horse or the boundary of a zebra what we get is shape of these two animals. And here you find that the shapes of these two animals are almost similar. So,

by looking at the shape only possibly I will not be able to say which one is horse or which one is zebra, right. But maybe that using this shape information, I will be able to say that, this is either a horse or a zebra, it is not a bird, ok. So, this is a kind of information that I will be possibly be generated; I will be able to generate given the shape information that this might be a horse or a zebra, but definitely this is not a bird.

But to distinguish between horse and zebra what I need is the additional information, that is what is the property of the region bounded within this boundary. So, if you look at on the left where I have a horse or on the right or I have a zebra, in case of a horse the color of the bounded region and the texture of the bounded region is totally different from the color and texture of the bounded region corresponding to a zebra. So, in case of a zebra I have black and white stripes which usually I do not have in case of a horse.

So, if I want to distinguish between a horse and a zebra or given an image, if one I have to recognize that which is the image of a horse or which is the image of a zebra, then possibly I will make use of this information both the shape information as well as the region information or the region properties which will help me to identify that which figure is the figure of a horse or which figure is the figure of a zebra. So, we need both these kind of informations, the shape information and the boundary information.

There are different shape informations possible. So, given a shape I will have multiple informations because I may have to distinguish one shape from other. So, I have to distinguish a rectangle from a circle. So, the shape information of a rectangle and the shape information of a circle will not be same they will be different. And this discrimination can be done using multiple number of properties. It may not be possible that I will have a single property. This has to be done using multiple number of properties.

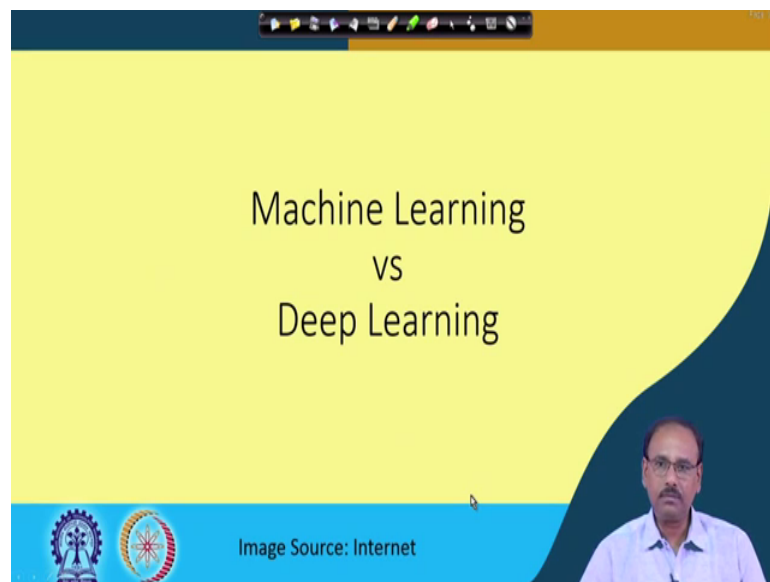
Similarly, whenever go for color again from the color I can extract multiple number of properties. Similarly, when I go for texture for texture, I can extract multiple number of properties. And each of these properties if I can represent in the form of numerical values then you concatenate all these properties together, right.

Say from shape if I extract 5 properties, then the shape information can be represented by of by a vector having 5 elements. Similarly, color from the color if I extract 10 properties or 10 descriptors describing a color all those 10 descriptors putting together gives me a

vector having 10 number of elements. Similarly, for texture again I can have 10 number of elements. So, if I concatenate all of them together, I gets a total 25 elements set. That means, a bounded shape, a bounded region is described by 25 features or by feature vector having 25 elements.

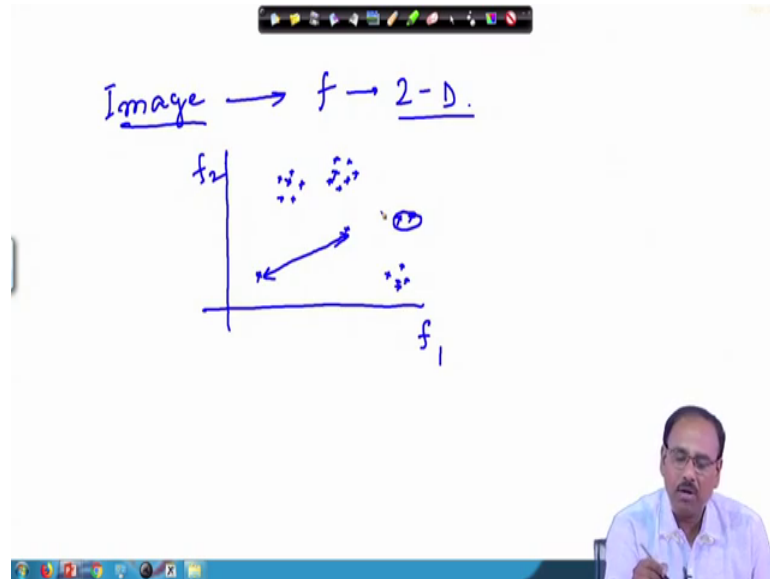
Now, using these feature vectors I go and go for identification or classification of the objects or classification of the bounded region, whether the bounded region is corresponds to a horse or it corresponds to a zebra or it corresponds to a bird, if it is a bird, what kind of bird and all these detailed classification can be made once I compute these descriptors or features.

(Refer Slide Time: 19:29)



So, the first level of understanding or learning is to find out the features. So, once I have this features then let us talk about what is machine learning, and when I talk about the machine learning then I will come to deep learning. So, for that let me try to find out let me see that what is machine learning, right.

(Refer Slide Time: 20:07)



So, what you have obtained so far, given an object or given an image. So, given an image or an object the image is converted into a set of vectors into a set of features are the features collected together is forming a feature vector f . So, you will find that if I have say 25 such features or have a feature vector having 25 elements, then this image is represented by a vector or by a point in a 25 dimensional feature space.

Now, for simplicity let me assume that this f is actually a two-dimensional feature vector for simplicity; obviously, a shape cannot be represented by or any image cannot be represented by two-dimensional feature vector accurately. But for simplicity I am doing this.

So, if I have this two-dimensional feature vector, let me put this feature vector feature components as component f_1 and component f_2 . So, f_1, f_2 gives you a feature vector. Now, if I have images of horse, they will form point distribution somewhere over here, so in my two-dimensional space. Similarly, if I have images of zebras they are very similar, right, so they can have a set of point cloud somewhere over here. But if I have images of say apples and I compute the same similar feature vector feature similar features f_1 and f_2 for apples, they may form a set of points somewhere over here.

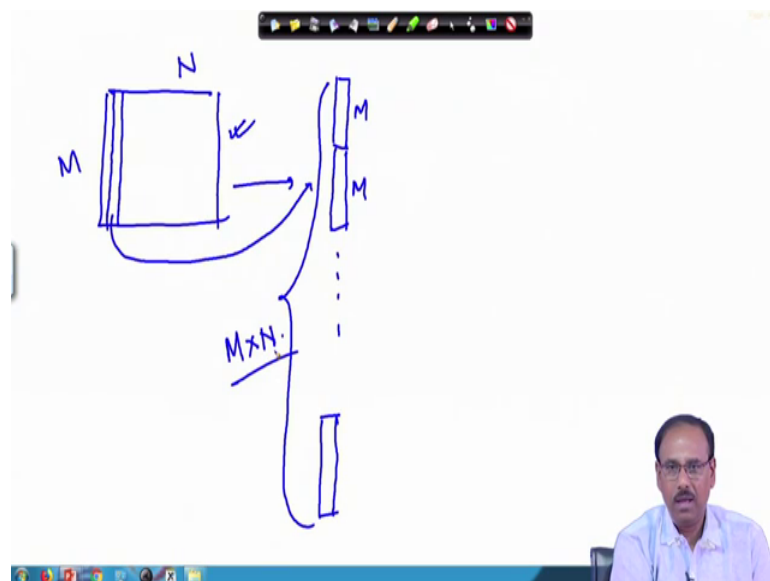
Now, why this distribution? The distribution is because every apple is not identical, every horse is not identical, every zebra is not identical. So, for all those different

pictures when I compute f_1 and f_2 , they will not always give me unique values, but there will be a variation and because of this I get distribution of this feature vectors.

Now, what is the advantage of having such a kind of distribution? The advantage is, advantage of having such a kind of feature vectors or descriptors. The advantages the moment I have a feature vector; that means, my image is now represented by point in the feature space. And given two points in the feature space, if I find that the distance between the two points is very large, I can immediately say that these two images are not similar they are different. But given two points like this where the distance is very small, I can immediately say that these two images or these two objects are similar. So, that is what you get, the advantage that you get when I represent these as feature vectors.

Now, given an image I can have different types of feature vector says that as I said, the shape, color, texture and many other features.

(Refer Slide Time: 23:25)



Given an image I can directly compute a vector from that. Some image is typically of size M by N , which has got M number of rows and N number of columns. One way of vectorization is you take every column from this image, so I get one column having N number of elements, then I take the second column having again N number of M number of elements and concatenate with this.

So, first here I have the first column that is first M number of elements, here I have the second column second M number of elements and so on, and that will be continued, and concatenate a concatenate with this the last column. So, you find that this entire image of size M by N is now converted to one-dimensional vector, what this vector has M into N number of elements. So, this whole image is now represented by a single vector.

Now, coming to what is the difference between our conventional machine learning and the deep learning. In case of conventional machine learning techniques, what used to be that you decide that what are the features or what are the descriptors that are that are to be extracted from the input signal, whether shape features, color features, texture features and what are those features.

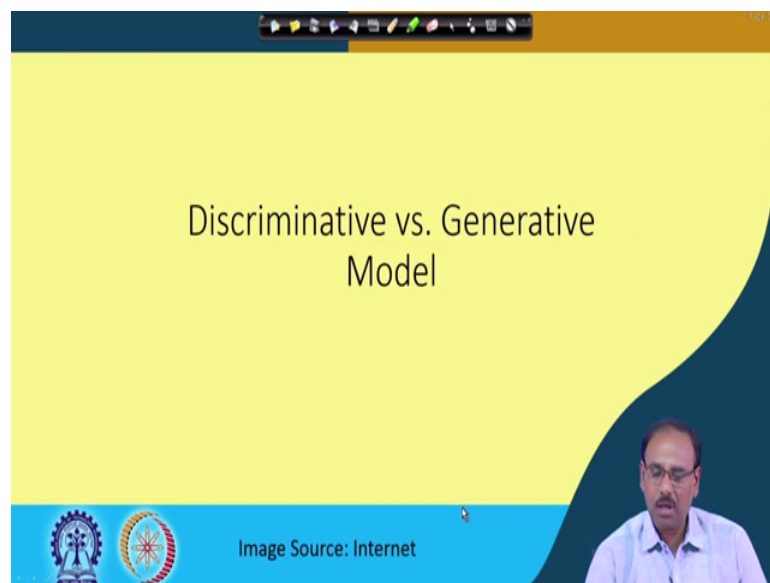
So, accordingly you have a pre-processing technique and this pre-processing technique gives you the feature vectors and these feature vectors are inputted to your machine learning algorithms. And for that we can have different types of machine learning algorithms, statistical machine learning like based rules, based classification rules. I can have linear or non-linear classifier, I can have support vector machines. We will talk about each of these in our later part, later lectures.

So, I can have different types of classifiers. Even I can have neural networks as classifiers. So, this feature vectors are inputted to those classification algorithms or machine learning algorithms. And for training the machine learning algorithm or pertaining the classifier what you do is you feed a large number of feature vectors taken from known objects; that means, the feature vectors I know that from which object or from which class that featured vector has been computed. And using this knowledge you will try to train your classifiers whether it is neural network, it is Bayes classifier, it is a support vector machine or whatever it is and this is something which is very similar to what we do.

Coming to our previous the first example, when you are shown the image of a cave painting Ajanta cave painting, it was also told that it was Ajanta cave painting. So, I knew what that object is or what that image is. So, in the same manner you train your machine learning algorithms, but the feature vectors are pre-computed by some pre-processing modules.

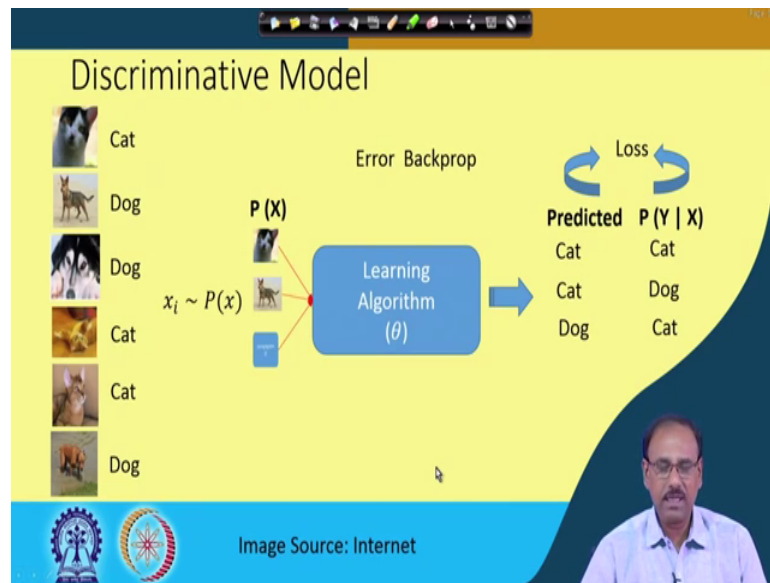
In deep learning what we do is no I will not have a person model, I will directly input the raw signal to our machine learning algorithm. So, the machine learning algorithm will not only learn the classes it will also learn which feature to look for, and usually this all the machine learning deep learning algorithms that we will be talking about they are all neural networks. So, effectively whatever you are doing is you are adding additional layers in your neural network to learn the features as well. So, typically that is the difference between your traditional machine learning and the deep learning algorithms.

(Refer Slide Time: 27:27)



Now, there are two types of deep learning algorithms, one is discriminative learning, one is generative learning. In case of discriminative learning what you try to do is once you are shown a set of images or set of objects you are try to discriminate among those objects; that means, given the image of a dog, I try to say that it is a dog it is not a cat, right.

(Refer Slide Time: 27:53)



So, basically as I said that every image or every object if I represent as a point to distribution given by $P(X)$ in case of discriminative learning or discriminating model given this distribution x , $P(X)$ and given the set of classes say Y , I try to find out what is the posteriori a posteriori probability P of Y given X ; that is given an input what is the probability that it belongs to certain class. That is what is discriminative model.

(Refer Slide Time: 28:25)

The slide is titled "Generative Model". In the center, there is a quote in pink text: "What I can not create, I do not understand." Below the quote is the attribution "- Richard Feynman". At the bottom, there are two red bullet points: "Collect a large amount of data in some domain" and "Train a model to generate data like it." At the bottom of the slide, there are two logos on the left and a small video inset of a man in the bottom right corner.

In case of generative model, it is motivated by what was told by Richard Feynman that what I cannot create I do not understand. That is, it is not only that you will be able to

classify or discriminate, it you should also be able to recreate what you learnt and that is what is your generative model.

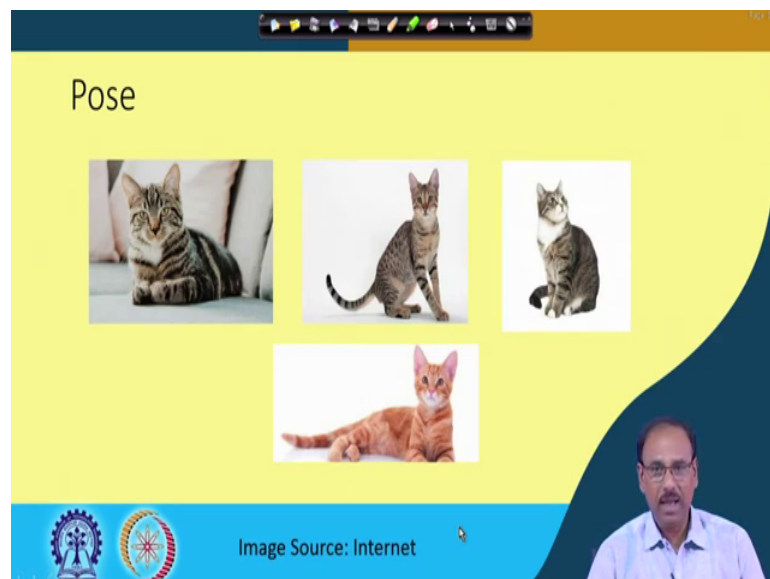
So, how do you do this? You collect a large amount of data in the same domain and then train a model to generate a data like it. We will also talk about this in our future lectures.

(Refer Slide Time: 29:07)



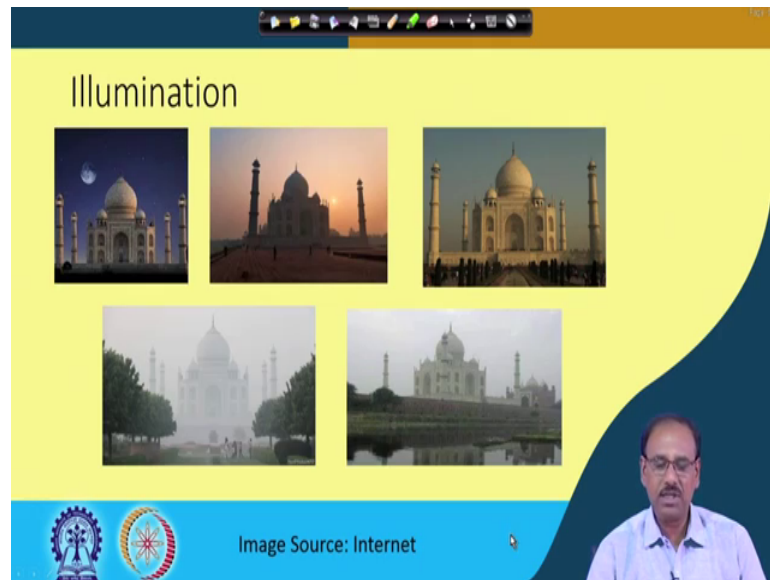
Then what are the challenges in deep learning or the challenges in machine learning? When you look at an object you can look at object from various angles and from various angles or various view in angles, it will appear to be different.

(Refer Slide Time: 29:27)



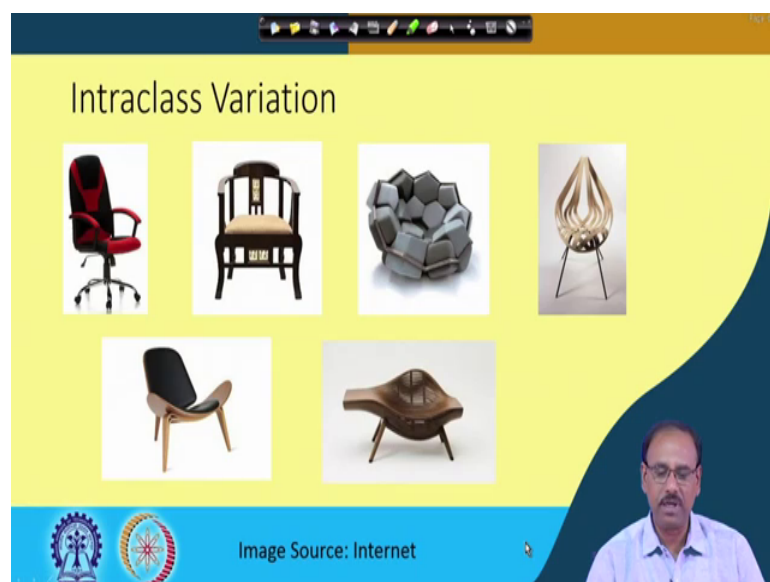
So, same object can have different views and from those different views we have to identify the object or you have to classify the objects that is one of the challenges. They can, it may be available in different poses.

(Refer Slide Time: 29:33)



There may be illumination variation.

(Refer Slide Time: 29:35)



There may be interclass variation. Say for example, here, who will say that this is a chair, though all these are chair classes So, I can also have intraclass variation, that is another

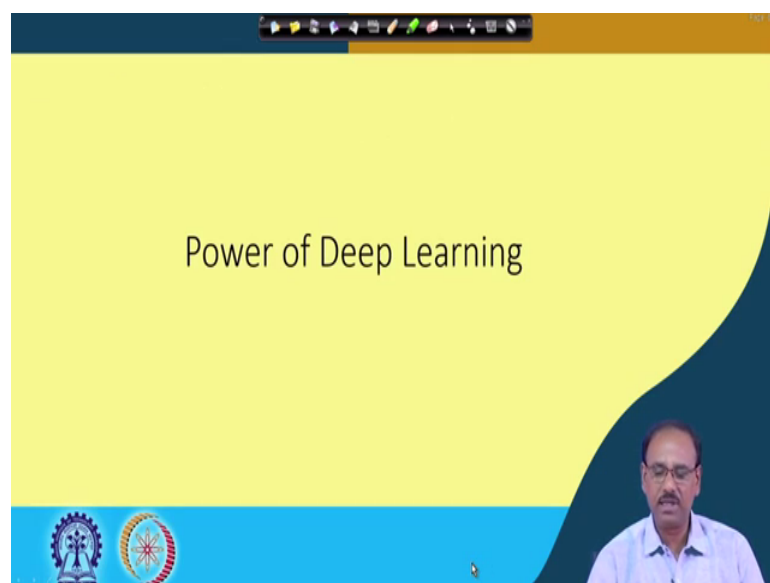
challenge in this deep learning algorithms, but because our deep learning algorithm has to work even in presence of intraclass variation.

(Refer Slide Time: 30:03)



There may be distortions and occlusions, may be part of the object is visible the majority of the object is not visible. It might be present in a distorted fashion, and so on. So, those are all challenges of the deep learning algorithm.

(Refer Slide Time: 30:13)



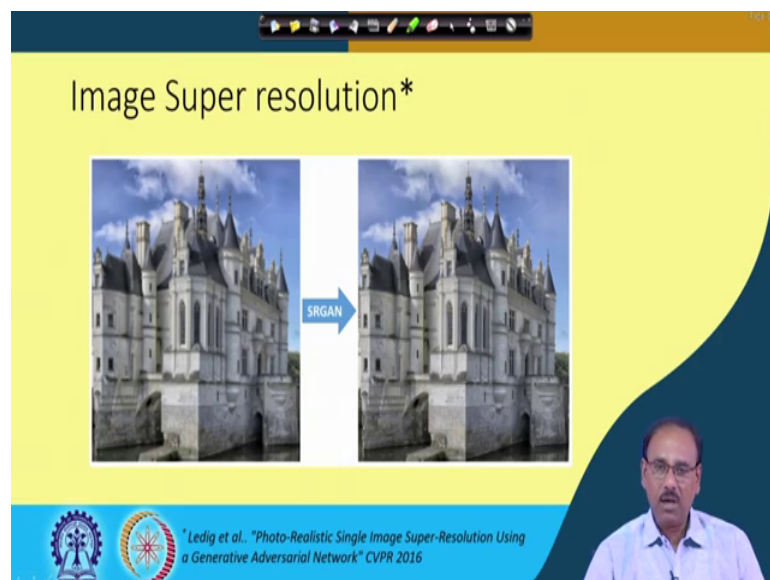
And coming to the power of deep learning techniques what we can achieve is we can even able to synthesize high resolution images and this is what is done through generative model.

(Refer Slide Time: 30:23)



From a low resolution image, we can create super resolution image.

(Refer Slide Time: 30:35)



Given any sketch of image, any sketch of an object, I can generate, I can synthesize a photograph of that object. Given symmetric segmented output I can have a real looking image.

(Refer Slide Time: 30:41)

The slide displays several examples of image-to-image translation. At the top, it is titled "Image to Image Translation*". Below the title, there are six pairs of images, each labeled "input" and "output".

- Labels to Street Scene:** An input image with colored regions (purple, green, blue) representing different object classes is translated into a realistic street scene.
- Aerial to Map:** An aerial photograph of a city is translated into a stylized street map.
- Labels to Facade:** An input image with colored regions representing different parts of a building facade is translated into a realistic facade image.
- Day to Night:** A daytime street scene is translated into a nighttime scene.
- BW to Color:** A black and white image of a flower is translated into a color image.
- Edges to Photo:** An edge-detection image of a handbag is translated into a realistic photo of the handbag.

At the bottom of the slide, there are logos for IIT Bombay and IIT Madras, and a citation: "Isola, Phillip, Jun-Yan Zhu, Tinghui Zhou, and Alexei A. Efros. 'Image-to-image translation with conditional adversarial networks.' CVPR, 2017". A small video inset of a speaker is visible in the bottom right corner.

So, all these are possible using the recent modern deep learning techniques.

(Refer Slide Time: 30:57)

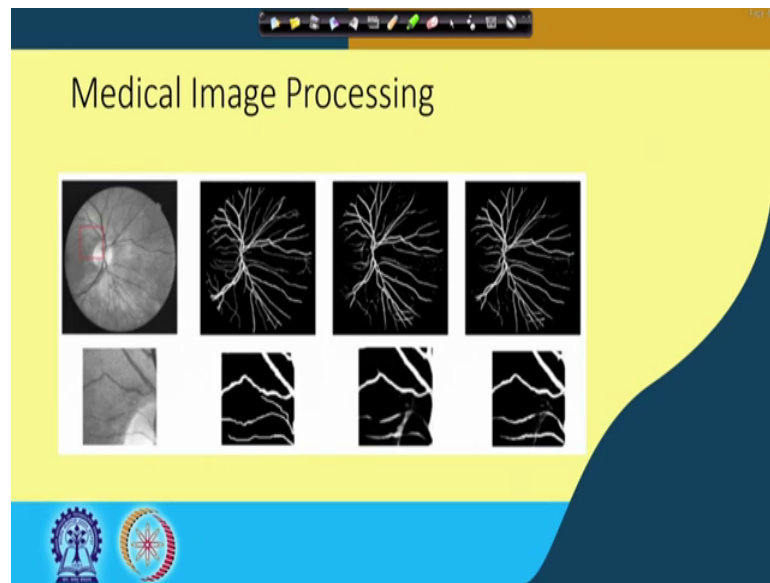
The slide displays examples of video-to-video translation. At the top, it is titled "Video to Video Translation*". Below the title, there is a grid of four video frames.

- Input Labels:** A video frame with colored regions representing different object classes.
- Style 1:** A realistic street scene video frame.
- Style 2:** A different realistic street scene video frame.
- Style 3:** A third realistic street scene video frame.

At the bottom of the slide, there are logos for IIT Bombay and IIT Madras, and a citation: "Wang, Ting-Chun, Ming-Yu Liu, Jun-Yan Zhu, Guilin Liu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. 'Video-to-video synthesis.' NeurIPS, 2018". A small video inset of a speaker is visible in the bottom right corner.

Similarly, we can also compose videos with different styles.

(Refer Slide Time: 31:13)



So, all these are possible using modern day planning techniques and there were multiple applications like in medical image processing, in object recognition, in speech recognition and so on. So, we will talk about all these different aspects of these deep learning techniques in subsequent lectures of this course. I hope we will be able to learn the content of these deep learnings and we will be able to apply it in the real-life problems.

Thank you.