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

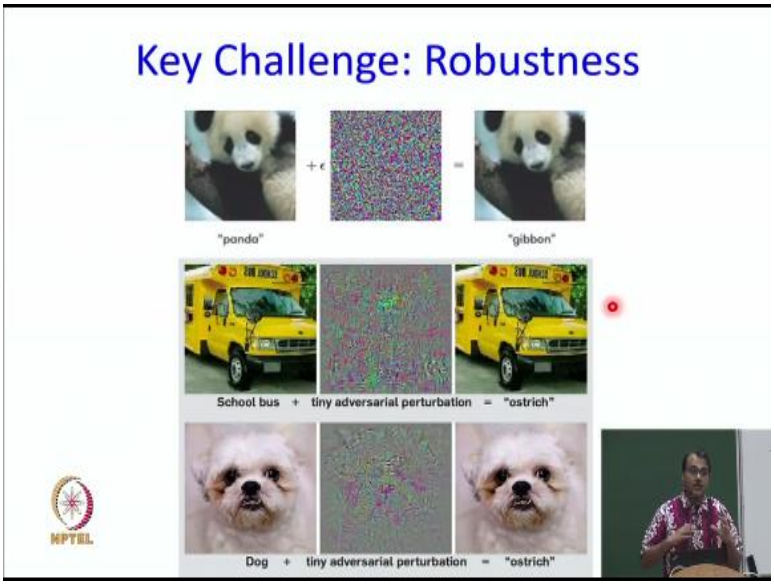
Lecture - 93
Ethics of AI: Robustness and Transparency of AI Systems

Okay, so what are the other challenges that we need to resolve before we talk about, before we get into a real we can claim that look, AI has been somewhat successful and we are not very scared anymore, it will do the right thing and so on. So one of the questions that a few years ago some researchers asked, and this is again, that sort of started coming out of Yoshua Bengio's group, I believe.

Yoshua Bengio I am sure you know is one of the three fathers of modern deep learning, person at Montreal, who is one of the only person of the three who primarily spends their time in a University, everybody else has moved. So Geoff Hinton is mostly at Google, I believe. Again, this is what I believe. Yann LeCun definitely is mostly at Facebook, even though they have affiliations with Toronto and NYU respectively.

But Yoshua still has a very strong group in Montreal and also spend some time in some startups and so on. So what they showed is that these deep learning models are, well, they make mistakes. That is okay. We all make mistakes, no AI system can be perfect. But they make sort of crazy mistakes or at least mistakes that appear crazy to us.

(Refer Slide Time: 01:38)



For example, when you look at recognition system, right. So the object recognition system gets an image as an input and the output is an object or more than one objects that are identified by the system. So if you give it this image on the left the system correctly says it is a panda. But it is sort of not very confident. It is sort of 55% confident or something. That is okay.

But now suppose you add very small amount of noise, almost the kind of noise that looks like salt and pepper kind of noise, basically random noise except that it is not random. So the point here is that it is not random. This noise has been very carefully crafted, based on the current system at hand. So that is a different story. These are called adversarial attacks to a deep learning system.

So you add this kind of a noise, but a very small amount of it. So I think point 0.7 times this noise or something like that. And so basically, when you add these two images, you get the image on the right. And for a human eye, it is basically imperceptibly same or different. No, it is the same thing basically for human eye. The human eye cannot perceive the first image and the third image to be very different.

But for whatever reason now the deep learning model is 99% confident that this is an image of a gibbon and not a panda. So you can say okay, you know pandas and gibbons are still you know animals and maybe some somebody may get confused. But if you take a school bus, which is correctly recognized as school bus, add this kind of

a specific noise, a tiny adversarial perturbation, you get image that still looks like a school bus to a human eye, but looks like an ostrich to a machine.

That makes no sense whatsoever. We have no idea what is going on at this point, right. And this also tells us that look, these systems can be hacked, these systems can be attacked, these systems can lead you to making predictions that completely will baffle humans, right? And if we allow these systems to take completely autonomous decisions, then some adversary can come up with some kind of a perturbation, which will completely mess up the system.

And once this happened, people got really surprised. And then there is a full field now that deals with how do you adversarial attack for example, a vision system or a natural language processing system. And then how can you create robust systems. And by the way, this latter, at this present moment of time is highly underdeveloped. So we do not know how to create robust systems which are high quality, right?

So essentially, neither do we know what the system is doing, right? We have no idea how in the world can it even think of this image as an ostrich. It is using some kind of surface patterns, which where these surface patterns are probably not the right patterns for making the predictions. But that is still what it is using. So there was a very beautiful example of wolf versus husky classifier.

And so the system was doing really well on wolf versus husky classification, but later it was found that it was only looking at whether there is snow in the image or not. And if there is snow in the image the system will classify it as a wolf and if there is no snow in the image then the system will classify it as a husky. So now this kind of classify we do not want, right.

Because then it will, it is using some kind of a pattern, which is not the point in question, but there is correlation at least in the data set. But because there is correlation in the data set, that is what it sort of learns, it does not learn the right thing. And when a different kind of image shows up, the system has no idea. I mean, it has a wrong idea of how to do it. And it has no idea that it is wrong. That is the other thing.

So there is a lot of work on even if you think about it yourself. If you try to talk to your parent and say why do you not try this app x or why do you not try the cell phone they say no, we are good. We do not need it. And essentially what you realize is that you do not know what you are missing. Your parent does not know what they are missing. Because until they try it, they will not really experience it.

Until they experience it, they would not know what was the value in it. So there is a category of known knowns that you know that you know this. There is a category of known unknowns. There is a category that you know you know that you do not know this. And when that happens, it is okay.

You can go read the book, you can go search on Google, you can go ask a friend, you can you know do some exploration or information gathering action, not exploration, information gathering action to figure out that information. You knew that you did not know it, you think you need it, you go and learn it. That is very easy. And there is this category of unknown unknowns that you do not know that you do not know.

And if you do not know that you do not know right, if you do not even know that you do not know there is no way for you to do the right thing unless somebody puts it in your face and says, somebody else puts it in your face and says, look you do not know this. Now at least please know that you do not know. And once you know that you do not know then you can you know go from there.

The system has no clue that it is making a mistake. The system feels that it is like an image that it has seen in the past. There is no way it can even self-correct itself, because it would not know where to look for, what to look for, right. It is possible that in some cases where these are known unknowns, then the system goes you know tries to do some information gathering but in this case it will not. Yes, question.

Right, so Yash's question is the slide says tiny adversarial perturbation. We do not know what the neural network is doing. So how do we even design a perturbation? And again, you should take a deep learning class because these things will get clearer there. But you still even if you do not know what the system is doing you know that if you give this input, it gives that output.

So suppose you have a black box, we do not go inside it. But we sort of know that this is its behaviors. On the basis of behavior, you start doing some perturbations in the original x space, in the original input space, and start looking at what the y distribution is. And just because it is a probability distribution or something, it will slowly show you that it is moving in some directions, you keep doing perturbations in those directions.

And you start getting. You can, Debyanshu says you can also look at the correlations in the data set and then perturb those. I am not sure if that is done, but that is there are many, many ideas. Yeah. And then there are white box attacks, which actually look at the gradients and then decide how quickly to come up with an adversary.

(Refer Slide Time: 08:47)

So but the one point that sort of comes out in this whole scenario is the fact that we do not know what they are doing. And we sort of know what they are doing, but we do not know why they are doing it. In fact, there is a beautiful joke where the programmer is banging their head on the machine saying, you deep learning I do not know why you do not work, The system is making mistakes, we have no idea how to debug it.

Should we just increase the layers and make sure I mean what, should we just start from a different starting point and or different hyperparameter? I mean what do we do right? By the way, what is a hyperparameter in this case? I do not think I discussed

this in the deep learning class. Hyperparameters would be a learning rate, for example, which kind of gradient propagation mechanism to use, how many layers to use, how many nodes to use, etc.

These would be parameters, hyperparameters, which define the particular network and its training regime. So what do you do? Why does it not work, right? And if by any chance it starts to work, then you again question, well, I do not know why you work, right. There is just no way we can go inside and understand anything, I mean humanly, right. There are lots of numbers, a million parameters. 1, -1, -0.99, -0.0037.

It just makes no sense to our brain. The deep learning model does not have the capability to come out with the insight as to why it made a thing because it may or may not have a notion of vocabulary. Of course, again these things are changing. So this is extremely important for a human AI team, right? Suppose you have a doctor, which is working with an automated agent and saying, hello automated agent, what do you think this person has?

An automated agent comes out and says okay, this person has meningitis, I believe. And the doctor says really, why? And the deep learning system will say because 1.01 - 1.997 - 00.37. That makes no sense, right? So the point is that it is incredibly important for real applications of AI to be able to work with humans, right? And to be able to work with humans we need the AI systems to collaborate to communicate in human language.

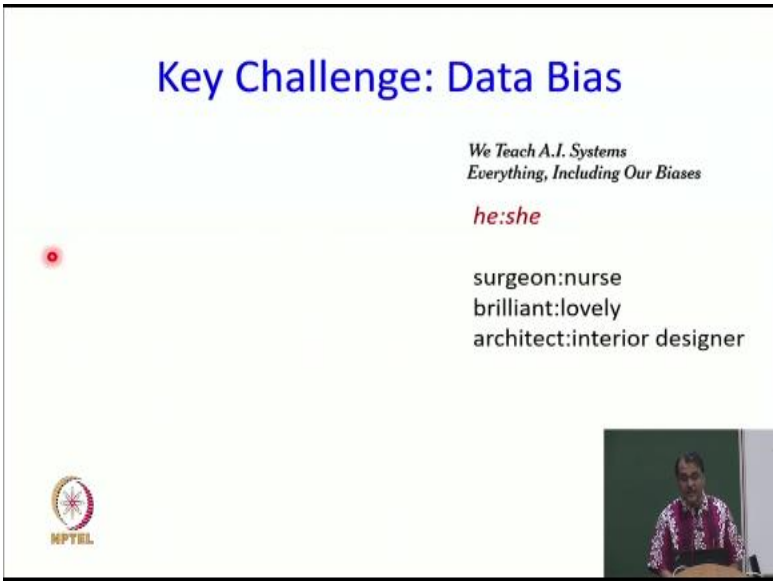
And that is going to be very hard, okay. So there is a full research agenda that got started a couple of years ago called the Xplainable AI or XAI. This Xplainable AI, by the way a lot of innovations, big innovations in the field of AI and computer science in general happen because of defense money. And especially in the US this defense funding is sort of ahead of the curve.

They keep thinking of new innovative programs that they want the full community to be working on. And then they say, okay you take a million dollars, do this. If somebody gives you a million dollars and say do this, then you usually do even in the

US, right? In a startup, you could say no a million is not good enough. I will take 20 but in a academic environment, these are big numbers.

So you are giving me a million dollars over three years. I am just happy to take it and you know fund my graduate students from there, fund research programmers and do good work also because these programs are of interest to somebody especially defense. So you know you sort of do these things. So one of those DARPA projects is called the XAI project, the Xplainable AI project, right.

(Refer Slide Time: 12:11)



So hopefully you know a little while from now we will have some transparency or a little bit more transparency in the AI system. But we are still miles away from being useful and usable today.