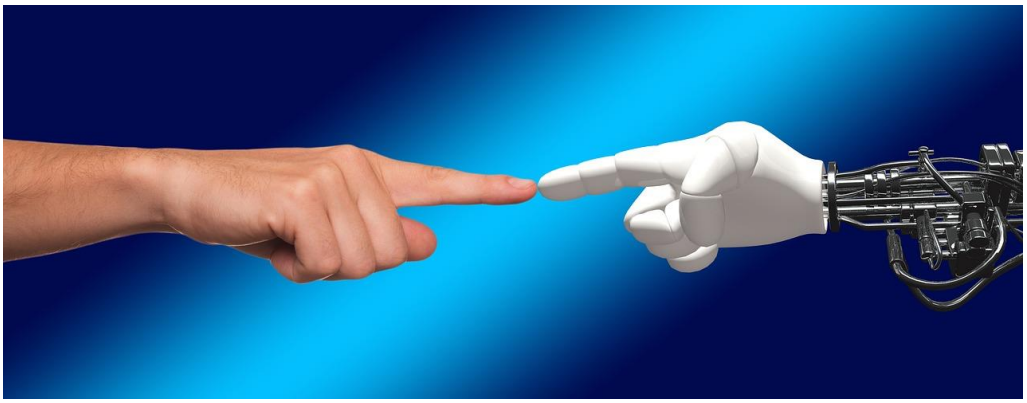
Ethical Aspects of Deep Neural Networks



Post image (taken from Pixabay)

Neural networks are a powerful tool which has improved our lives. Along with their benefits, neural networks do have significant drawbacks - they are vulnerable, they require a lot of data which is not always achieved by proper means, and they tend to preserve bias. We must ensure that any network we build is explainable, fair, secure and maintains privacy in order to make it a legitimate technology.

Lotem Nadir

Lotem Nadir is a recent graduate with an undergraduate degree in computer science from the Technion – Israel Institute of Technology. She works as a data scientist at Medtronic. She loves data-driven problems, and deep learning solutions.

Introduction

> “I was walking home from school by myself when I saw it. A robot, just standing in the middle of the street. I was scared at first, but then I realized that it wasn’t doing anything. I walked up to it and said, “Hi.” It turned its head to me and said, “Hello, human.” I had never talked to a robot before. We talked for a while and I found out that its name was R0b0t. I asked why it was just standing in the street and it said that it was waiting for its human friend. I told it that I didn’t have any friends that were robots, but that I would be its friend. R0b0t said that it would like that. Since then, R0b0t and I have been best friends. I’m not afraid of Artificial Intelligence anymore, because I know that they can be just as good of friends as anyone else.”

Where do you think this paragraph is taken from? A new children book published to help children embrace new AI technologies? A sketch for the next Wall-E movie? This story was

actually written by a computer. The computer was given the following prompt: “Write the beginning to a short, fictional story about a child that is afraid of Artificial Intelligence, but then makes friends with a robot”, and generated this story. It was also asked to generate drawings for the story as well, as shown below.

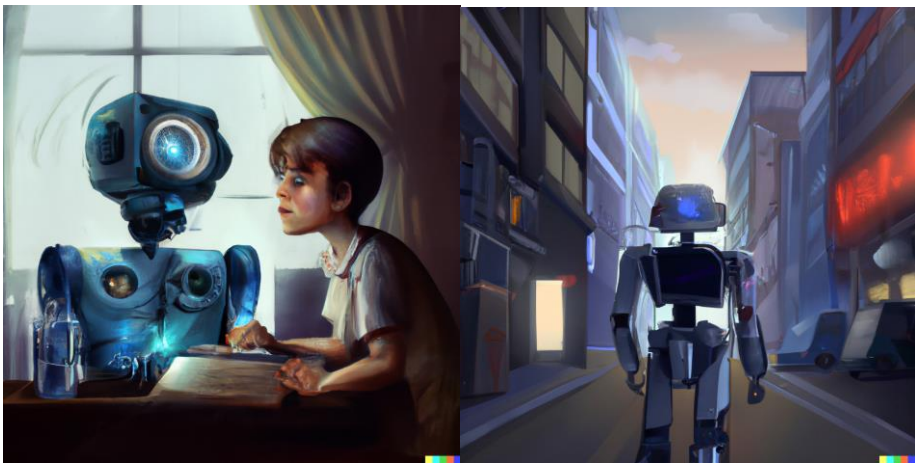


Figure 1: Drawings the computer generated for the short story it has written.

This example shows how far AI technology has progressed. Actually, each one of us uses neural networks on a daily basis: in web searching, digital personal assistants, machine translations, chatbots and so on. The most used technology in the AI field is deep neural networks. Deep neural networks were used to generate the short story and the drawings above, and are used in all other tasks described above.

Deep Neural Networks

The problems we often face in AI can be generally described as: “I have some input, and I need the computer to give me some output regarding that input”. For example, animal image recognition can be described as “I have an image of an animal, and I need the computer to tell me the name of that animal”. Another example is translation from Hebrew to English, which can be described as: “I have a sentence in Hebrew, and I need the computer to translate it into English”. Let's dive in with the first example. Let's say we work at a zoo that has only pandas and dogs. The dogs in the zoo look very like pandas, as shown below.

Figure 2: The dogs in our zoo.

Every year, all pandas and all the dogs are being photographed for the zoo yearbook. The editor of the yearbook asked us to hand him all the images, including for each image whether it's a dog or a panda. Sadly, the photographer forgot to add this label. We decided to solve this problem with a computer. Our problem can be described as “We have an image of an animal, and we need the computer to tell us whether it is a dog or a panda”. If we convert the image to numbers, and the labels “panda” and “dog” to numbers as well, the solution for the problem would be mathematical function f .

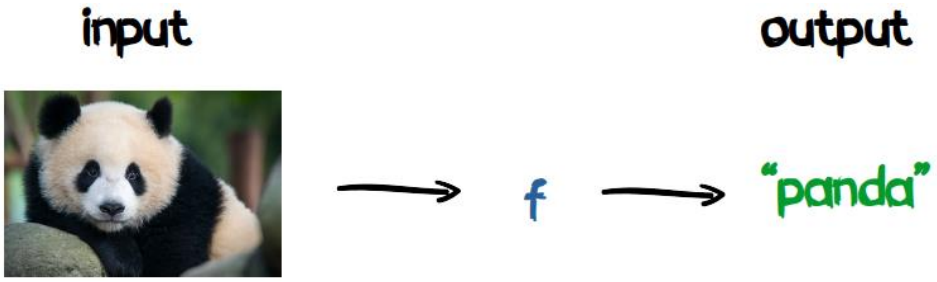


Figure 3: The function f takes as an input an image (converted to numbers), and outputs the label.

When we talk about **deep learning**, we refer to **deep neural networks**, which are the **algorithm used to find the solution function** f. I won't be covering how this algorithm works, but we do need to understand several features of it: First, in order to make that algorithm work well, we need to provide it with a large amount of data. In our case, we would have to collect many images of pandas and dogs. How much “many” is? Well, according to “Kaggle”, which is the largest data scientists online community, the smallest data set in the top 10 popular datasets has 60,000 images, and the largest one has 15,000,000 images. Second, most of the time we don't know how the solution function f looks like. In rare cases, if our problem is simple, we might end up with some function we all know like $f(x) = x$ or $f(x) = sin(x)$, but in real life these functions are very complex. They are so complex, they cannot be drawn in two dimension graphs or even in three dimension graphs. Third, the model is highly sensitive to data bias. This issue is known as “garbage in, garbage out” in the machine learning world, meaning if we have bias in our data, we would expect such bias in the solution function f.

Privacy in Data Collecting

Now that we know a little bit about deep neural networks, we can start asking ourselves what are the pitfalls of that technology. As we've seen above, in order to make that algorithm work, we need a lot of data. How is this data being collected? In some cases, the data is being generated by consent, e.g. by using questionnaires or by inviting people to the research lab and taking photos of them. In other cases, the data is being collected without consent, for example by recording and storing commercial activities: buying with credit cards, placing online orders, using frequent shopper cards and any other action that has digital signature (Nissenbaum 2004, 3).

A well-known example of that was several years ago, when a man walked into a Target outside Minneapolis and demanded to see the manager. He was clutching coupons that had been sent to his daughter. “My daughter got this in the mail!” he said. “She’s still in high school, and you’re sending her coupons for baby clothes and cribs? Are you trying to encourage her to get pregnant?” The manager didn’t have any idea what the man was talking about. The manager apologized and then called a few days later to apologize again. On the phone, though, the father was somewhat abashed. “I had a talk with my daughter,” he said. “It turns out there’s been some activities in my house I haven’t been completely aware of. She’s due in August. I owe you an apology”. In this case, Target clearly collected data regarding the daughter's shopping habits. They were able to identify some products that, when analyzed together, allowed them to assign the daughter a “pregnancy prediction” score, and to know that the daughter is pregnant. This might be seen as a privacy violation, especially since the daughter was not aware of the fact that data about her was being collected.

An ethical use of AI requires that data is collected, processed, and shared in a way that respects the privacy of individuals and their right to know what happens to their data and to access their data (Coeckelbergh, 2020, 61).

Adversarial Attacks

Another pitfall comes from the fact we don't know how our solution function f looks like. If we don't know how the function f looks like, we can not protect it from its deficiencies. Adversarial attack is a machine learning method that aims to trick machine learning models by providing deceptive input. For example, suppose we have images of digits as shown below.

We trained a neural network that outputs for each one of these images a digit (0, 1, 2, …, 9). The computer was able to classify the above images correctly using our network. Let's say that for some reason we manipulated these images, resulting in the following:

Each one of us is able to classify these new images correctly. But surprisingly, the computer was not able to classify the new images correctly, as described in a paper by Ren et. al. In order to understand why this happened, we need to understand how computers see images. The computer sees an image as a matrix of numbers. Each cell in the matrix corresponds to a tiny piece of the image called pixel.

Figure 4: An image (left) and its representation as the computer sees it (right).

If for example we add one to all values in the image matrix, the human eye won't be able to see the difference, but for the computer it will be a completely different image. Since the computer uses a function to classify the image, by slightly changing the input, we might get significantly different output. In our example, the following two images might look to us the same, but the computer might classify them differently.

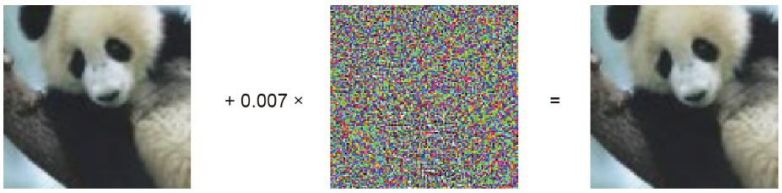


Figure 5: Original panda image (left) and slightly manipulated image (right). The colorful image in the center represents random noise added to the image. The computer might classify the left and right images differently.

Adversarial attacks might not be harmful when it comes to classifying pandas from dogs, but it might cause damage in other fields. Think about autonomous driving. What if we applied the same method on a network trained to classify the speed limit written on speed limit signs? Adversarial attacks can cause in this case car accidents and kill people. What about biomedical systems? Neural networks are used in these systems for diagnosing and decision support: In 2018, the FDA approved marketing for the first-ever autonomous artificial intelligence (AI) diagnostic system and regulators have articulated plans for integrating machine learning into regulatory decisions. Also, deep neural networks are used in insurance claims approvals. Billions of medical claims are processed each year, with approvals and denials directing trillions of dollars and influencing treatment decisions for millions of patients. What if these networks were attacked so they give false diagnosis? Or deny an insurance claim when it needs to be approved? Solutions that might mitigate these ethical issues are to insist on improving vulnerable algorithms until they are made adequately protected, and to make sure these

vehicles or systems have an overriding option to manually use them, as described in a paper by Hansson et. al.

Injustice

As described above, the algorithm of neural networks is highly sensitive to data bias. For example, black Americans are incarcerated in state prisons at nearly five times the rate of White Americans. If we trained a model to classify whether a person is likely to get incarcerated based on its race, we would find black persons are more likely to get incarcerated. Does this mean that black people are more dangerous to society? Of course not. An example for this exact problem is a network used in the US to predict the likelihood of a person committing a future crime. That network rated a high risk for a black person who committed a small crime and a low risk for a white person who committed a more severe crime, because of the bias we already know there is in the data. How can we solve the problem of data bias? One might claim that we should change the machine learning algorithm in order to decrease the risk of bias, but that would make its predictions less accurate. Another idea is to create “idealized” data sets, with equal representation for all. However, in this case the data is not mirroring the real world. There is no right answer in this case. How to deal with bias in AI is not a technical question, but a philosophical one. The question is what kind of world we want, if we should try to change it, and if so, what ways of changing it are acceptable (Coeckelbergh, 2020, 80).

References

1. Coeckelbergh, M. (2020). AI Ethics (1st ed.). MIT Press.
2. Nissenbaum, H. (2004). Privacy as Contextual Integrity.