3.3 How can we make deep learning more transparent?

Deep neural networks are notoriously opaque to human inspection. It’s difficult for users to understand how they arrived at their output. As deep learning algorithms are being used for applications from facial recognition to healthcare diagnoses to autonomous vehicles, it is increasingly more important for their inputs and outputs to be transparent.

We’d like you to consider the balance between the accuracy of the models and the need for transparency. Are there situations where deep learning should never be used? Or are additional safeguards required?

* <https://www.entrepreneur.com/slideshow/289621#9>

Deep learning what is it

We all use deep learning in our day to day life. I’ll take the example of speech recognition.

We have siri, cortana, alexa for example all here to help us.

Cortana advertisement: <https://www.youtube.com/watch?v=DxrJWSi_IWo>

Sometimes they can also embarrass us:

**Cortana embarrassing Microsoft CEO:** <https://www.youtube.com/watch?v=QSFL2soY9hI>

Why did I get such different result? Deep learning algorithms use neural network which optimise an output to given set of input variables.

**Physical illustration of neural network -**

We’re using statistical methods where computer learn to classify patterns.

But not only is deep learning is a greedy process (huge amount of data required to be accurate; it is also opaque (using a number of black boxes which cannot be explained)

Problem

Neural network are obscure, unknown “selection” method for process optimisation

Should they be used everywhere – accuracy vs transparency

They have shown their invaluable value. Taking the same example of speech recognition

<https://medium.com/applied-innovation-exchange/the-good-the-bad-and-the-ugly-of-artificial-intelligence-and-machine-learning-3f7e663c317a>

The use of automatic speech recognition in a mobile device application — to analyse the audible noises a child makes upon birth. Taking a baby’s cry as the input, the machine learning system will analyse various characteristics of the cry to provide an instant diagnostic of birth asphyxia. Not only will the tool diagnose the condition, but it will do so at a dramatically reduced cost. Their solution claims to be over 95% cheaper than an existing clinical alternative — a breakthrough in cost reduction, in a world of ever increasing cost pressures.

Where not good?

In 2016, Microsoft launched its experimental AI chatbot, Tay, onto Twitter with not so pretty results. The intention was for Tay to mimic the language patterns of a millennial female using Natural Language Understanding (NLU) and adaptive algorithms in a bid to learn more about “conversational understanding” and AI design.

After just 16 hours, Tay was removed from the internet after her jovial exchange turned into an A-Z of insults, sexism and racism.

Now what?

Common biases in ML are sampling, performance, confirmation & anchoring bias. With the increasing use of deep learning in fields such as intelligent warfare and self-driving cars, the need for increasing transparency is critical.

So what can we do to remove these bias? Recent efforts have been put into:

estimating unintended consequences

assigning value or negative impact to them

pyramid flow

* The set up and interpretability: **Bring together a social scientist with a data scientist.** Two types of scientist who speak a different language and can reduce the discriminatory potential of a model.

“The people who understand the world, don’t understand the math. The people who understand the math, don’t understand the world”

* Sampling: balance in representativeness

A deep learning algorithm can only be as good as the data set which has been used to train and validate it.

* Testing and validation of the model: in as many situation as possible regardless of how likely
* Adversarial attack and defence – to prevent the use of AI algorithms by hackers