Can AI be "explained"? Some considerations on XAI and its goals

Alvise de' Faveri Tron

November 8, 2020

Contents

1 Introduction **1**
1.1 Artificial Intelligence and its shortcomings 1
1.2 XAI: a new frontier of AI research 2

2 The Explainability Problem **3**
2.1 A simple instance: Neural Networks 3
2.2 Solution Approaches 3

3 What is missing **3**
3.1 Why is there an explainability problem in the first place? 4
3.2 Explainable *to whom*? 4
3.3 Explainable *for which purpose*? 4
3.4 A multidimensional solution space 4
3.5 Can they be measured separately? 4

4 Conclusion **5**

1 Introduction

1.1 Artificial Intelligence and its shortcomings

It is a fact that Artificial Intelligence has become an increasingly mature and pervasive technology nowadays. From being a matter of pure scientific research, with vague goals and great expectations, it has grown to be one of the hottest technological trends right now, both in the academic and industry world, with a massive and constantly expanding industry revolving around it and a growing relevance in the public debate. Today, using some kind of AI solution to solve previously hard or infeasible problems is a trend in many fields, most notably some that weren't previously interested at all in technology, e.g. law and medicine .

cit

cit

cit

The reason why AI is spreading so aggressively as a technology is that it really outperforms previous solutions: . examples

However, its encounter with the real world has also shown us what are the limits of this technology, and even if its fast growth trend can convince someone that these limitations are just "technical" and can be solved easily with more power, there are some intrinsic facts that we are observing in many different AI technologies.

The first one is that, when it breaks, it breaks hard: a misprediction made by an AI system, especially if it has to accomplish a highly complex or impacting task, can cast a shade on the correctness of the whole model itself, on the data it has been trained on or on its design. There's rarely such thing as "fixing one line of code" on deep neural networks that have been trained on millions of data points.

The second one is that these kind of errors are generally difficult to predict in advance: an AI algorithm can perform very well on a high number of inputs, but have a weak point that is only discovered way after the AI has been deployed. There is no general method to know in advance where an AI might fail, we just know that until now it has been pretty accurate.

The third one is that at the moment we have very little introspection tools when examining an AI system, especially when we are talking about neural networks: the same people that design and train the algorithm have generally little knowledge about what model the network is going to produce at the end, and when it does the only way of verifying its correctness is black box testing, for which the input space is generally huge.

All these points, together with market and newspapers hype, have encouraged the AI industry to tackle that which seems the hugest obstacle for AI being adopted everywhere: the problem of understanding an AI model and "opening" the black box.

1.2 XAI: a new frontier of AI research

Explainable AI is a concept that was recently formalized in a call for research made by the DARPA, the same agency where the word "Artificial Intelligence" was born in the first place. It is meant to be describe a new set of Artificial Intelligence systems which are designed to be easier to understand by humans. In particular, the goals of XAI is making artificial intelligence more: cit

- Debuggable
- Predictable
- Trustful

immagine presa dal paper del DARPA

This effort requires a variety of expertise, from Computer Science to Cognitive Psychology, and there is still a lot of work to do.

2 The Explainability Problem

2.1 A simple instance: Neural Networks

Of all the approaches that have been tried during its 50-years history, one has recently emerged as capable of solving many huge, complex and unrelated problems all together: Machine Learning. This particular form of AI is aimed at building a model of a problem from observing many examples of it, and letting the algorithm figure out the best way of connecting inputs and outputs, shaping itself based on what it "learns" from the data. More specifically, one of the most studied and heavily developed approaches right now is Neural Networks, which encodes the information that has been learned as weights in the connections between basic units, called neurons.

perchè sto spiegando cos'è una rete neurale?

A conceptual example of what a neural network is is shown in figure ...

rappresentazione grafica

However, this is just a *conceptual* representation. A more accurate representation of what is going on is a series of computations in the form of:

rappresentazione matematica

Which in itself is also an abstraction, since we don't see here how the hardware is effectively calculating and solving these equations.

On the surface, this seems like a technical detail, but this is already the core of the AI Explainability problem: looking at these figures we are trying to get an idea of the *architecture* of the algorithm, but we would have a really hard time if we wanted to correlate what we see with the decisions that are being made by the algorithm.

2.2 Solution Approaches

1. visualization
2. simplification
3. reverse engineering/fuzzing (exciting inputs with black box testing)
4. prototypes
5. differential for classifiers (what do I have to change to change the outcome)

Ma come valutare le soluzioni? Basta la complessità?

3 What is missing

The problem is that there is a lack of a formal definition of how an explanation should be measured. This is not a trivial point, since it is very difficult to quantitatively measure the goodness of any explanation. Many papers refer to complexity but this is not enough.

3.1 Why is there an explainability problem in the first place?

- causality vs correlation
- previous categories
- having a goal -¿ the decision means that I have to do something in practice, and that thing has an ethical an practical impact

3.2 Explainable *to whom*?

The users of an AI systems are:

- end user
- developer
- operator
- judge

3.3 Explainable *for which purpose*?

The main purpose for XAI are:

- debugging of the internals
- human-in-the-loop
- validation and certification
- appeal decisions

3.4 A multidimensional solution space

- complexity
- clearness
- informativeness
- fidelity

3.5 Can they be measured separately?

- fidelity vs complexity
- fidelity vs clearness
- quality vs performance

4 Conclusion

In conclusion, the main problem of XAI is that there is no single definition of what an explanation is, it depends on the purpose and on the user of the AI system.

For this reason, these should be considered different problems, at least the debugging problem vs the right of explanation problem: they are not correlated and saying that one solves the other poses some threats on the quality of the result itself.

"I always thought something was fundamentally wrong with the universe" [1]

References

[1] D. Adams. *The Hitchhiker's Guide to the Galaxy.* San Val, 1995.