Intelligent Agents
Assignment 2.1 Interpretability

Erik Norlin, 19970807-9299

March 1, 2023

1. Interpretable models vs. black-box models

The up-rise of black box models in recent years have exploded and have become a heated trend in the world of technology. These models have shown fantastic performance in prediction and classification tasks that has blown the world away, and is currently and successfully being integrated into society, especially with the latest ChatGPT. What many people do not know however is that there lies potential great danger in these black box models because of their complex structure. In this text we will discuss what interpretability and explainability is, the importance of interpretability in machine learning (ML), and potential dangers with these black box ML models.

Before black boxes even was a thing ML models were only ”interpretable” models. Interpretable ML models are models that are composed of primitives that are humanly interpretable, meaning that the technicality of how the model computes its' predictions can be understood by humans. Interpretable models are therefore transparent ML models that can give insight into their computations when making predictions or classifications because it could be desired to have these kind of results backed up by clear explanations of how predictions and classifications are derived.

Black box models on the other hand, is on the other end of the spectrum of ML models. A black box model is a function that is too complicated for humans to fully understand. Black boxes are usually ML models that consists, most often, of deep neural networks (DNNs). These models are so complex that it is very difficult, if not impossible, for humans to understand what it is actually going on inside it and how it derives its' results, hence the name black box. Its' output is a result of a DNN that takes an input and uses its' parameters to transform it into, most often, some sort of classification or prediction, where the network's parameters have been shaped by data the model has been trained on. The output of a black box is based on probabilities of how the input matches the probability distribution of the model, so the input is processed by the very complex structure of the DNN. Because of this complexity, it becomes therefore very difficult to judge the validity of the output that the model gives, if what the model spits out is true or not. However the case, DNN black box models are very powerful and accurate models if they are trained on huge data sets. Black box models are on the uprise everywhere and widely used within research and are getting more and more complex every day. Black box models turn out to be incredibly good when it comes to computer vision and reproducing statistical properties of language, which the recent trend of ChatGPT, a statistical language model (SLM) consisting of a black box, shows fantastically. Black boxes can also be great as a starting point for decision making, but one should be wary of taking the result of a black box as an end goal [6], [7]. It turns out that black boxes besides their impressiveness are not always factually correct, tend to contain unintended bias and that they lack common sense, which we will come back to later.

The main difference between interpretable models and black box models are transperency of the systems. Interpretable models can be transparent about how they work and how they derive their computations. Not that black box models cannot be transparent, but due to their incredibly complex structure it becomes meaningless for a human to try to understand it.

Even though black box models are definitely more complex than interpretable models does not always mean that black boxes are inherently better models for prediction performance. It does not mean that interpretable models are simple either. Interpretable models can be complex but interpretable by humans. It is a myth that there is a trade-off between intepretability and accuracy when it comes to prediction, especially sorted and preprocessed data, then there is little advantage with a black box compared to an interpretable model. There is a hype among some that complex technology is automatically better technology which is just not a general truth, sometimes more interpretability gives better accuracy [6].

It shows that there are both advantages and disadvantages with interpretable models as well as with black box models. It is not the point in this text to demonize black boxes, but rather to heighten awareness of potential dangers of these models as black boxes are having an enormous hype at the moment. It is therefore of utter importance to see through the hype so that we do not proceed forward blindly and hypnotized by the trend.

2. Interpretability vs. explainability

Interpretability and exlainability are nowadays used interchangeably but they are not the same thing [6]. So what does these terms even mean and how are they different? "Interpretability is the degree to which a human can understand the cause of a decision made by an algorithm" [4]. In other words, interpretability means transparency of an algorithm. An interpretable model can for instance give a precise description of how the model came to its' results in a literal technical way that a human can understand. This gives room for understanding of how the model operates and is highly beneficial when trouble shooting the system because one can, for example, precisely locate an error and gain insight about the model. This itself allows an interpretable model to give clear explanations of how it derives its' results, which can be favourable in high-stake stake decision making. For example, if someone would want to use a ML model to predict if a person is going commit a crime it can be good to know how the model came to its' conclusion so that one can judge the validity of the models outcome. A black box however does not and cannot explain how the model came to its' results in a technical way that humans can understand. Black boxes can however offer "explanations", but those are based on the data that the model is trained on, not on the computation by the system itself. These explanations are often done through another model that is designed to explain the black box. This should raise a warning sign, because if another complex model is needed to explain what the original model computes, then the explanation cannot be true with certainty because if the explanation was fully true one would not need a second model to explain it in the first place [6].

Even though black boxes can offer explanations of their output, it is not certain that it is always right. For example, if an explanation is correct 90% of the time, how can one know when the model is wrong? If a black box model outputs "New York" to the question "What is the largest city in USA", the black boxes explanation could go something like "Because New York has largest population of all cities in the United States". This is not an insight into the model, but an insight about the world based on the data that it has been trained on (notice that data can also be factually incorrect), which is not the same thing as transparency. If one do not know how the algorithm works and when the algorithm is factually correct or incorrect, then one cannot know when it fails [4]. This leaves a void in the user that is aware of this, and leaves one to

question why to use a black box when creating solid ground to stand on for decision making.

One could argue that this is risk taking that we take almost everywhere with technology in society since there is always a risk of failure in technology. Now, 90/10 accuracy/risk ratio might for many be too risky to consider trusting a black box model, but as the technology of black boxes progresses and become better and better, this risk ratio might become smaller and smaller to the point that some might consider it justifiable to trust a black box, even in high steak decision making. This sort of risk taking however, is not the same as with other technology because with other technology we still get to know how it works and can therefore troubleshoot to find errors if they occur. This is something that is not possible with black box models. The black box becomes more of a black hole where it becomes impossible to retrieve information about how the black box goes wrong when it fails, leading to zero understanding and knowledge about potential fallacies in the model.

In the quest of trying to discover ways of making neural networks more interpretable, attempts have been made to compress neural networks in hope to gain interpretability and maintain accuracy. This however has turned out to fail because if the real world is highly complex, then a neural network that is trained on real world data will hence also become complex. Currently, there are no known ways of making neural networks more interpretable. Why black box models are even used at all is due to simpler models with the task at hand either fail or are too expensive to simulate [2], [5]. It turns out that there are value for black models, but when is it appropriate to use them?

Another arguement against interpretability is that if black boxes can do better than humans, then why not use these models? Humans can also make mistakes! Let's take an example of why black boxes become problematic to use in high stake decision making. When a clinician makes a mistake, trouble shooting for the reason of mistake can be emphasized. A surgeon cannot justify a special operation for a patient just because a black box predicted a disease that needs surgeon procedure. A comeback to this is why it is not justifiable when a black box model predicts it but only when a radiologist makes the same finding. The reason for this is interpretability. A surgeon can confirm the special findings in a patient from a radiologist based on the radiologists ground of proofs and reasoning. This is not possible with black box models, no matter how accurate these models are, because their methods for computing their result cannot be interpreted, so interpretability becomes favourable in high stake decision making [4].

3. Dangers of black models

Due to the lack of human interpretability in black boxes, there lie great consequences of danger when using these models, which has already been observed in society. Because of inaccurate predictions by black boxes, there are people that have been incorrectly denied parole, poor bail decisions have lead to release of dangerous criminals, black box models stating that polluted air was safe to breathe when it was not in reality [6]. In one study where researchers were using OpenAI's GPT-3 as a conversational agent to simulate a person having depression, it advised this fake patient to kill themselves [1].

Another take on what was previously discussed, that some argue against interpretability because black boxes can do better than humans. This is partially true, though not to the extent one might think. Since black box models results rely on specific information of the input it is given and how well it matches the probability distribution of the model, it fails to generalize and thus rely on brittle, fragile methods for computation (prediction, classification etc.). Using brittle evaluation methods for high stake decision making can lead to catastrophic outcome if some-

thing goes slightly wrong. An example of how black boxes are brittle can be understood when looking at the way radiologists identifies the L1 vertebra (a specific back bone). The L1 vertebra is identified by a radiologist by scanning from the top down and locating the last vertebra that ribs attach to. A black boxes method to do this could be, for example, "to locate the bright object just above the kidney". This method is not robust because there can be variation in the data for whatever the reason might be, so the black box can fail when a large enough difference occur. A radiologist can however identify the L1 vertebra even when greater variation in the data occurs because the radiologist has a deeper understanding and insight about the autonomy of the human body. What is further interesting about black boxes is that different initializations of the same model and data can end up with different methods for the same task and end up with the same accuracy, but all being equally brittle [2].

Another example of a black box relying on brittle methods, but this time potentially catastrophic, is a self-driven car that completely fails to recognize a stop sign if the sign has a sticker on it and classifies the stop sign as a sign saying "45 speed limit" [3]. This concretely and perfectly demonstrates the fragility of black boxes, and how they can potentially fail catastrophically, especially in high steak decision making. This goes hand in hand with the most fundamental flaw of black boxes; these models fail to generalize and therefore do not inherit common sense. One could argue against this by saying that the lack of ability to generalize is too much to expect technology to inherit, since common sense is a very complex phenomena. Do we ever expect other technology to be as intelligent as humans? Also, interpretable models does not have common sense either, so why should we trust interpretable models? This is true, but interpretable models are at least interpretable, and the point is not to necessarily fully rely on interpretable models, but to take their output and computation in consideration in decision making. A doctor making a diagnosis does not have to fully rely on the interpretable model. The doctor can consider parts of its' computation (due to its' interpretability) and consider their own reasoning and understanding in combination with the output of the model. Deeper understanding and reasoning should always be considered in high stake decision making, so trusting black box models blindly is perhaps not a good idea. With these examples mentioned, it becomes clear yet again that interpretability in ML models becomes of high importance in the case of high stake decision making .

In order to somewhat tackle the issue of black boxes not being able generalize, black boxes would have to become even larger models, trained on even more data to become more robust to noise of the input, such as the stop sign with a sticker. However, there lies danger in training models on huge data sets. The data that black boxes are trained on are so large that it becomes almost impossible to quality check the data so that it does not contain unintended biases or other inappropriate content [7]. It is therefore not a sustainable solution to make the black box models larger to compensate for its' lack of ability to generalize.

With potential dangers that come with the rise of black box models, it is important to be conscious about these dangers and to be aware of the lack of interpretability of these kinds of ML models. Reporting accuracy of black box models should not be the end goal of artificial intelligence research. Interpretability should be a goal to strive for when designing ML models [4]. Whade and Virgolin (2021) proposes a methodology for interpretability that they have designed to follow when designing ML models. Their methodology is called "The five I's" and briefly, The Five I's are five pillars, or rules to follow when designing ML models to make them more interpretable and transparent, and thus contributing to moving towards a more transparent era of machine learning and artificial intelligence [7].

4. Conclusion

It is truly great that technology has come such a long way that we have models that are as impressive as black box models are, and these models sure have earned their place in certain areas, especially statistical language models as many of us know. Black box models can be exceptionally useful in some cases, but not always. Sometimes interpretable models are equally as accurate and far less complex than black boxes which allows for interpretability, which should be more desirable in high stake decision making. An interpretable model should always be considered first in high stake decision making if an ML model is considered necessary because this allows for deeper understanding of how the output is derived. If it does not fit the job, one could use a black box model as a starting point, however, not as as an ending point due to its' lack of interpretability. With great technology comes great responsibility. We must therefore take accountability and be aware of the potential dangers that black boxes out-lie in front of us, and do our best to make sure that we strive towards a place in society where ML models are as transparent as possible, especially when it comes to high stake decision making. It is not certain that we are moving towards a direction of more interpretability, but we should make our best to do so [2]. Because one thing is certain and that is that black box models are here to stay and they will continue to become more and more complex. Very fast.

References

[1] Ryan Daws. *Medical chatbot using OpenAI's GPT-3 told a fake patient to kill themselves.* Nov. 2021. URL: `https://www.artificialintelligence-news.com/2020/10/28/medical-chatbot-openai-gpt3-patient-kill-themselves/`.

[2] Daniel C Elton. "Self-explaining AI as an alternative to interpretable AI". In: *Artificial General Intelligence: 13th International Conference, AGI 2020, St. Petersburg, Russia, September 16–19, 2020, Proceedings 13*. Springer. 2020, pp. 95–106.

[3] Kevin Eykholt et al. "Robust physical-world attacks on deep learning visual classification". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018, pp. 1625–1634.

[4] Adarsh Ghosh and Devasenathipathy Kandasamy. "Interpretable artificial intelligence: why and when". In: *American Journal of Roentgenology* 214.5 (2020), pp. 1137–1138.

[5] Timothy P Lillicrap and Konrad P Kording. "What does it mean to understand a neural network?" In: *arXiv preprint arXiv:1907.06374* (2019).

[6] Cynthia Rudin. "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead". In: *Nature machine intelligence* 1.5 (2019), pp. 206–215.

[7] Mattias Wahde and Marco Virgolin. "The five Is: Key principles for interpretable and safe conversational AI". In: *2021 The 4th International Conference on Computational Intelligence and Intelligent Systems*. 2021, pp. 50–54.