MARCH 2021

Key Concepts in AI Safety: Interpretability in Machine Learning

CSET Issue Brief

AUTHORS

Tim. G. J. Rudner

Helen Toner

This paper is the third installment in a series on "AI safety," an area of machine learning research that aims to identify causes of unintended behavior in machine learning systems and develop tools to ensure these systems work safely and reliably. The first paper in the series, "Key Concepts in AI Safety: An Overview," described three categories of AI safety issues: problems of robustness, assurance, and specification. This paper introduces interpretability as a means to enable assurance in modern machine learning systems.

Introduction

Interpretability, also often referred to as *explainability*, in artificial intelligence (AI) refers to the study of how to understand the decisions of machine learning systems, and how to design systems whose decisions are easily understood, or *interpretable*. This way, human operators can ensure a system works as intended and receive an explanation for unexpected behaviors.

Modern machine learning systems are becoming prevalent in automated decision making, spanning a variety of applications in both the private and public spheres. As this trend continues, machine learning systems are being deployed with increasingly limited human supervision, including in areas where their decisions may have significant impacts on people's lives. Such areas include automated credit scoring, medical diagnoses, hiring, and autonomous driving, among many others.[1] At the same time, machine learning systems are also becoming more complex, making it difficult to analyze and understand how they reach conclusions. This increase in complexity—and the lack of interpretability that comes with it—poses a fundamental challenge for using machine learning systems in high-stakes settings.

Furthermore, many of our laws and institutions are premised on the right to request an explanation for a decision, especially if the decision leads to negative consequences.[2] From a job candidate suing for discrimination in a hiring process, to a bank customer inquiring about the reason for receiving a low credit limit, to a

soldier explaining their actions before a court-martial, we assume that there is a process for assessing how a decision was made and whether it was in line with standards we have set. This assumption may not hold true if the decisionmaker in question is a machine learning system which is unable to provide such an explanation. In order for modern machine learning systems to safely integrate into existing institutions in high-stakes settings, they must be interpretable by human operators.

Why Are Modern Machine Learning Systems Not Interpretable?

Many modern machine learning systems use statistical models called deep neural networks which are able to represent a wide range of complex associations and patterns. To understand why decisions by deep neural networks are hard to interpret, consider two types of systems that *are* interpretable.

One example is an earlier generation of AI systems which use human-specified rules instead of relying on data. As a simplified example, the autopilot system on an airplane uses a set of *if-this-then-that* rules to keep the plane on course—if the nose drops, lift it; if the plane banks left, roll a little right, and so on. While the rules in real systems are more complicated, they nonetheless allow humans to look back on system behavior and recognize which *this* triggered which *that*.

A second example is a linear model, a simple kind of machine learning model. Like all machine learning systems, a linear model uses number values called *parameters* to represent the relationship between inputs and outputs. For example, one could create a model to predict someone's salary from their age and years of schooling, two *explanatory variables*. In a linear model, the main parameters would be one number value to be multiplied by the explanatory variable "age," and one number to be multiplied by the other explanatory variable "years of schooling." Determining what value those two parameters should take is the *learning* part of machine learning. For a linear model, good parameter values can be found by simple calculations that may take a computer less than a

second to perform. More importantly, because each parameter in a linear model is directly associated with one explanatory variable, understanding how the model works is simple. If, say, the parameter that gets multiplied by "age" is much higher than the parameter for "years of schooling," then the model is predicting age is a more important determinant of salary.

Deep neural networks are different. By design, they have far more parameters than linear models, and each parameter is connected in complex ways with inputs and other parameters, rather than directly linking explanatory variables to the outcome that the model seeks to predict. This complexity is a double-edged sword. On one hand, models can represent highly complex associations and patterns, allowing them to solve problems previously considered out of reach for computers, including image recognition, autonomous driving, and playing the game of Go. On the other hand, unlike older or simpler computer systems, the internal functioning of each model is very difficult to understand.

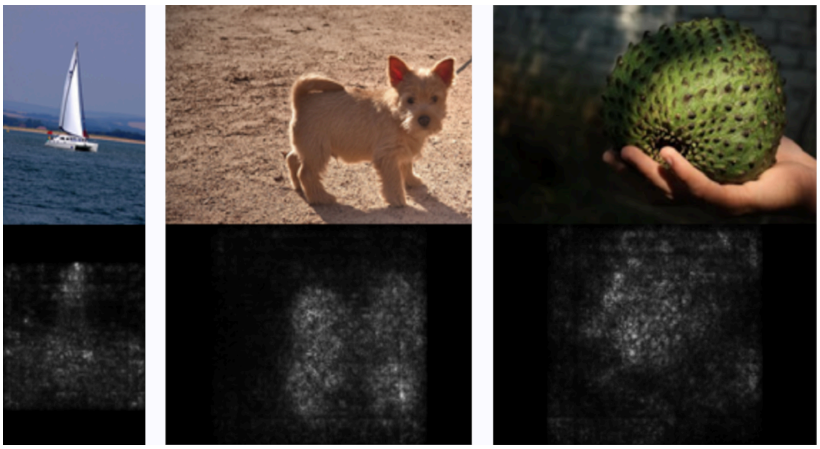
At this point it is worth noting why the term "black box," often used in this context, is not quite right to describe why deep neural networks are hard to understand. Machine learning researchers understand perfectly well how the mathematical operations underlying these systems work, and it is easy to look at the parameter values that make up the model. The problem lies in understanding how these millions (or even billions) of number values connect to the concepts we care about, such as *why* a machine learning model may erroneously classify a cat as a dog.

In other words, interpreting deep neural networks requires both the ability to understand *which* high-level features in the data—such as a certain part of an image or a specific sequence of words—affect a model's predictions and *why* a model associates certain high-level features with a corresponding prediction—that is, how deep neural networks "reason" about data.

How to Make Modern Machine Learning Systems More Interpretable

Researchers are pursuing a range of different approaches to improving the interpretability of modern machine learning systems. A fundamental challenge for this work is that clear, well-defined concepts have yet to be developed around what it would mean for different types of systems to be interpretable. So far, interpretability research seeks to build tools making it somewhat more possible for a human operator to understand a system's outputs and inner workings.

Figure 1. Examples of images and their corresponding saliency maps indicating which parts of the images contribute most to how a machine learning system trained on a large set of images would classify them.



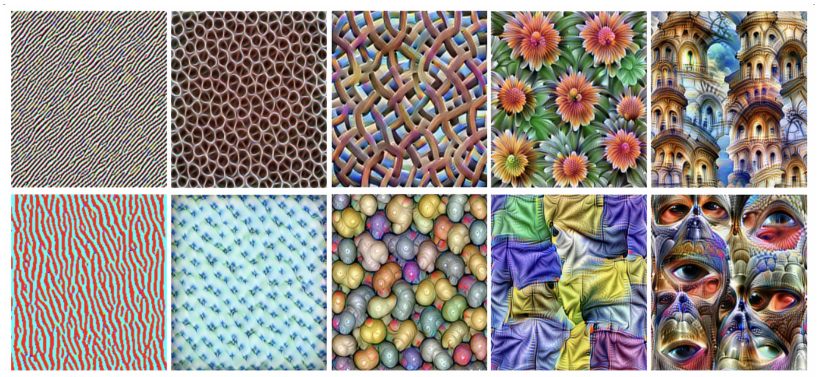
Source: Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman, "Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps," *arXiv [cs.CV]* (April 19, 2014), preprint, https://arxiv.org/abs/1312.6034.

Saliency maps are one popular set of tools for making modern machine learning systems used for computer-vision applications more interpretable. Broadly speaking, saliency maps visualize which areas of an image led to a model's classification of the same image.[3] For example, we might investigate why a deep learning model has learned to identify images of cats and dogs from a large dataset of cat and dog images labeled as such. If we wish to understand why the model classified an image of a German Shepherd as a dog, a saliency map may highlight the parts of the image containing features present in dogs, but not in cats (for

example, a large muzzle). In this way, a saliency map communicates to a human operator which part of the image prompted the machine learning system to classify the image as it did.

Another popular method for making deep neural networks more interpretable is to visualize how different components of the model relate to high-level concepts that may affect the model's predictions—concepts such as textures and objects in image classifiers, grammar and tone in language models, or short-term vs. long-term planning in sequential decision-making models.[4]

Figure 2. A visualization showing examples of what different layers of an image classification network "see." The left-most column, depicting an early layer, is picking up lines; middle layers are detecting textures and patterns of increasing complexity; later layers, shown on the right, are looking for objects.



Source: Chris Olah, Alexander Mordvintsev, and Ludwig Schubert, "Feature Visualization," *Distill*, November 7, 2017, https://distill.pub/2017/feature-visualization.

Unfortunately, existing methods to make modern machine learning systems interpretable fall short; they typically only provide one angle from which to view the system instead of taking a holistic view. To fully understand how a machine learning system works, we must understand how the data and learning algorithm affected training, whether training could have resulted in a different model under modified training conditions, and how all of these factors ultimately affect the system's predictions. At this time, our understanding of these questions is still very limited and the insights obtained from existing methods are fallible, require human supervision, and only apply to a small subset of application areas.

For example, the saliency maps in Figure 1 do shed some light on how the model in question works. One can see, for instance, that the model focuses on the dog to classify it as such, but identifies the sailboat in part by looking at the ocean. Saliency maps, however, do not help a human observer understand what might have led to different outcomes. Likewise, Figure 2 shows a method for understanding what the different parts of an image classifier detect. Unfortunately, looking at this type of visualization does not help a human operator evaluate whether the system is likely to be accurate, fair, or reliable.

The lack of a common, well-defined vocabulary for interpretable machine learning systems further exacerbates these shortcomings.[5] Key concepts, such as trustworthiness, reliability, transparency, or verifiability, are often used loosely or interchangeably rather than referring to standardized or generally accepted technical definitions of such terms, making it difficult to measure progress and to accurately and reliably communicate research results to the public.

Outlook

Interpretability will allow us to understand potential failure modes for machine learning systems, enable regulatory compliance and audits, and reduce the risk of models with algorithmic or data-induced bias being deployed. Rendering modern machine learning systems like deep neural networks interpretable will help us ensure that any such system deployed in safety-critical settings works as intended. Unfortunately, while an active and ongoing area of research, existing approaches to achieving interpretability do not yet provide satisfactory solutions. It remains unclear when—or even if—we will be able to deploy truly interpretable deep learning systems. In the meantime, our best option may be to stick with simpler, more inherently interpretable models whenever possible.[6]

Authors

Tim G. J. Rudner is a non-resident AI/ML fellow with CSET and a PhD candidate in computer science at the University of Oxford. Helen Toner is director of strategy at CSET.

Acknowledgements

For feedback and assistance, we would like to thank Igor Mikolic-Torreira, Shelton Fitch, Alexandra Vreeman, Lynne Weil, Larry Lewis, Jack Clark, Dewey Murdick, and Michael Page.

Document Identifier: doi: 10.0.201.137/20190042

Endnotes

[1] Danny Yadron and Dan Tynan, "Tesla Driver Dies in First Fatal Crash While Using Autopilot Mode," The Guardian, June 30, 2016, https://www.theguardian.com/technology/2016/jun/30/tesla-autopilot-death-self-driving-car-elon-musk; Forough Poursabzi-Sangdeh, Daniel G. Goldstein, Jake M. Hofman, Jennifer Wortman Vaughan, Hanna Wallach, "Manipulating and Measuring Model Interpretability," arXiv [cs.AI] (November 8, 2019), arXiv, preprint, https://arxiv.org/abs/1802.07810; and Jennifer Valentino-DeVries, "How the Police Use Facial Recognition, and Where It Falls Short," The New York Times, January 12, 2020, https://www.nytimes.com/2020/01/12/technology/facial-recognition-police.html.

[2] For a more in-depth discussion of what an "explanation" is in a legal context, why explanations are necessary, and how AI explanations compare to human explanations, see Finale Doshi-Velez, Mason Kortz, Ryan Budish, Chris Bavitz, Sam Gershman, David O'Brien, Kate Scott, Stuart Schieber, James Waldo, David Weinberger, Adrian Weller, Alexandra Wood, "Accountability of AI Under the Law: The Role of Explanation," arXiv [cs.AI] (December 20, 2019), preprint, https://arxiv.org/abs/1711.01134.

[3] Simonyan, Vedaldi, and Zisserman, "Deep Inside Convolutional Networks"; Julius Adebayo et al., "Sanity Checks for Saliency Maps," arXiv [cs.CV] (October 28, 2018), arXiv, preprint, https://arxiv.org/abs/1810.03292; Ruth Fong and Andrea Vedaldi, "Interpretable Explanations of Black Boxes by Meaningful Perturbation," arXiv [cs.CV] (January 10, 2018), arXiv, preprint, https://arxiv.org/abs/1704.03296; Ruth Fong, Mandela Patrick, and Andrea Vedaldi, "Understanding Deep Networks via Extremal Perturbations and Smooth Masks," arXiv [cs.CV] (October 18, 2019), preprint, https://arxiv.org/abs/1910.08485.

[4] Jesse Vig, "A Multiscale Visualization of Attention in the Transformer Model," arXiv [cs.HC] (June 12, 2019), arXiv, preprint, https://arxiv.org/abs/1906.05714; OpenAI, "Dota 2 with Large Scale Deep Reinforcement Learning," arXiv [cs.LG] (December 13, 2019), preprint, https://arxiv.org/abs/1912.06680.

[5] Zachary C. Lipton, "The Mythos of Model Interpretability," *arXiv [cs.LG]* (March 6, 2017), arXiv, preprint, https://arxiv.org/abs/1606.03490.

[6] Cynthia Rudin, "Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead," Nature Machine Intelligence 1 (May 2019): 206–215, https://www.https://www.nature.com/articles/s42256-019-0048-x.