Common Pitfalls When Explaining AI and Why Mechanistic Explanation Is a Hard Problem

Daniel C. Elton

Abstract Recently researchers have started using explainability techniques for several different applications—to help foresee how a model might operate in the field, to persuade others to trust a model, and to assist with debugging errors. A large number of explainability techniques have been published with very little empirical testing to see how useful they actually are for each of these use cases. We discuss several pitfalls one can encounter when trying to utilize explainability techniques. We then discuss how recent work on the double descent phenomena and non-robust features indicate that mechanistic explanation of deep neural networks will be very challenging for most real-world applications. In some cases, one may be able to use an easily interpretable model, but for many applications deep neural networks will be more accurate. In light of this, we suggest more focus should be given to implementing out-of-distribution detection methods to detect when a model is extrapolating and thus is likely to fail. These methods can be used in lieu of explainability techniques for increasing trust and debugging errors.

Keywords Interpretability · Explainability · Artificial intelligence · XAI · Out-of-distribution detection · Deep learning · Machine learning

1 Introduction

There is growing interest in developing methods to explain the inner functioning of deep neural networks. In this paper, we survey some of the pitfalls that are easily encountered when trying to "explain an explanation", some of which are not well appreciated in our experience in medical AI and other applied areas of applied AI. We first distinguish several motivations for interpretation in medical imaging. We argue that mechanistic interpretation (i.e., elucidating the underlying mechanism,

D. C. Elton (✉)
Radiology and Imaging Sciences, National Institutes of Health Clinical Center, Bethesda, MD 20892, USA
e-mail: daniel.elton@nih.gov

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2022

X.-S. Yang et al. (eds.), *Proceedings of Sixth International Congress on Information and Communication Technology*, Lecture Notes in Networks and Systems 235,
https://doi.org/10.1007/978-981-16-2377-6_38

similar to how we explain human decision-making) is potentially the most useful for increasing trust in AI. However, the recent discovery of double descent indicates that deep neural networks such as the convolutional neural networks work through the brute force local interpolation of data points. This is in contrast to popular narratives that deep neural networks work by extracting a few high level rules that we might explain in a few sentences [1]. As we will discuss, it follows that such networks are intrinsically hard to interpret and also have very little chance at being able to extrapolate outside their training distribution.

2 Motivations for Explanation

Different practitioners understand the term "intepretability" in different ways, leading to a lack of clarity on the matter [2]. Here we take the terms "explanation" and "interpretation" to be synonymous. We note several reasons we might be interested in explanations:

- **Prediction**—it can be useful to be able to predict whether a model can generalize or extrapolate to different conditions. For instance, will the model still be able to function if there is increased scanner noise or cropping?
- **Persuasion**—often we want to convince clinicians or other stakeholders that they can trust an AI system. Providing explanations can increase user's trust in a model, even if the explanations are not correct [3].
- **Debugging**—often we'd like to "open the black box" and understand why a model fails in particular cases. This can allow for iterative refinement.

While many fine distinctions can be made between different types of explanations, we find two high level definitions to be useful. We define a **descriptive explanations** as an account of model function which is descriptively accurate and relevant to the end user [2]. By "descriptively accurate", we mean that the explanation can reproduce the input–output mappings of the model to some degree. This type of explanation is typically boiled into natural language statement relevant to the domain of application. This type of explanation can be distinguished from a **mechanistic explanation**, which captures, at least approximately, the actual data manipulations occurring in the model. Only mechanistic explanation accurately predicts what the network will do when new data is presented.

Whether an explanation is descriptive or mechanistic can be distinguished by seeing if the explanation allows the user to predict the model's behavior when new inputs are presented from outside the model's training distribution or under counterfactual testing where parts of the image are removed. Most present day methods do not work very well in this regard, and thus fall under the category of descriptive explanation. Recently, Hase et al. have performed tests with a large pool of human subjects to see if different explainability techniques help users predict a neural network's behavior [4]. Out of the methods they compared, the only one that helped users for image data was the "This looks like That" approach [5], a method which was designed with

explainability in mind [4]. This is not surprising because the "Rashomon Effect", [6] which says that for any set of noisy data, there are a multitude of models of equivalent accuracy but which differ significantly in their internal mechanism. To give an example of why mechanistic explanation may be useful, consider task of pancreas segmentation in CT images, which is challenging due to the variable shape and low contrast of the pancreas relative to background tissues. A robust way of finding it would be to locate a higher-contrast and easier to locate organ first, such as the liver. The pancreas has a relative position to the liver which is fairly consistent. A somewhat more brittle method would be trace the lower intestine to the duedenum, to which the pancreas is attached. A very brittle method would be to look for the pancreas in the center of the image.

3 Pitfalls

Recently, there has been a "cottage industry" of research showing problems with saliency-based methods [7, 8] and related "heatmapping" methods such as layer-wise relevance propagation [9]. At a high level, saliency maps may show where a model is not looking, but not what is doing. Saliency maps for different classification outputs may look similar, making it hard to distinguish why the network chose the output it did [6]. Many outputs of saliency-based methods are very similar to the output of an edge detector. Confirming this, it was found that, even if most of the later layers of a neural network are randomized, saliency maps do not change much [8]. Additionally, if labels or features are scrambled and the model is retrained, the outputs do not change [8] Indeed, it has been found that features "important" by the explanations are actually no more important than randomly specified features [10].

Recently, Olah et al. hypothesized that deep neural networks are explainable in the mechanistic sense, given enough careful study of what features each node represents and the connections between them [11]. Olah's et al.'s techniques are based on activation maximization, where a neuron or group of neurons is related to an interpretable feature (or more rarely, a combination of features). We see two issues with this type of approach. The first is that pure activation maximization leads to images which look like noise to the end user, so many "regularization" constraints have to be applied (in particular, Olah uses constraints developed for the highly publicized "deep dream" visualization). We are skeptical of this procedure due to its artificial nature. The method is tailored to provide the end user a pretty picture rather than remaining true to visualization the mechanism of the network. The "noise" which naive activation maximization shows may actually be "non-robust" features (see [12]). The second problem is that, if you take a linear combination of units from a given layer instead of a single unit (or more precisely perform a random rotation / change in basis), and maximize that instead, you end up with similar types of explanations for what each unit is sensitive to [13]. While Olah has discussed this issue in one of his previous works from 2017, in our view, it has not been adequately addressed in his most recent work [11].

There are many sources of variance which are often not taken into account when performing explanations. In several different contexts, it has been shown that the output of explainability methods often varies between test cases [14, 15]. When interpretations are shown, many should be given and they should be randomly chosen rather than "cherry-picked". If possible, a holistic analysis should be done, averaging interpretability results from many cases. Non-rigid registration to a reference case may be useful here in certain contexts such as medical imaging. Another source of variance is that in some cases visualizations for different train-test folds appears different, even when the resulting models are of equivalent accuracy [14, 16]. One way of mitigating this issues is to average results over a few different visualization methods [17]. Another source of variation that effects interpretation is the hyperparameter settings used. For instance, changing the LIME hyperparameters slightly has been shown to significantly change the output of the visualization in some cases [18].

An alternative method of explanation is to train a "post-hoc" model which is simpler and more interpretable to reproduce the output of the hard to interpret model. For instance, a decision tree can be trained to reproduce the output of a CNN. This procedure is the same as model distillation, where a large model is distilled on a smaller one by running the large one on a large unlabeled dataset and training the smaller model to reproduce the output of the larger one. Lillicrap & Kording show that this technique has limits however, and distilled models for image classification with equivalent accuracy are still quite large, with millions of parameters [19]. Thus, distilling further to a small interpretable model will incur a large decrease in accuracy, and therefore, won't be properly reproducing the input–output behavior of the original model.

Several recent works add an "explanation branch" to "explain" the output of a different branch of the network (the "prediction branch") [20, 21]. For the case of diagnosing lung nodules in chest CT, an example is illustrated in Fig. 1. The explanation branch in this case was trained to predict several attribute scores which clinicians consider important for lung nodule diagnosis. By seeing which attributes were predicted, the idea is that this constitutes an "explanation" of how the network arrived at its prediction. The issue with this sort of approach is that, it is not clear how the two branches are related—in principle they could be computed independently. In a recent work, we described a possible solution, which is to use a measure of mutual information overlap to make sure the outputs of the explanation branch and the prediction branch are related [22].

4 Why Mechanistic Interpretation Is Difficult

It has been noted for several years that the most successful deep learning models have millions of parameters and appear to be vastly underdetermined, yet they still generalize. More recently, it has been shown that the bias-variance trade-off breaks down in large enough networks [23]. Belkin et al. call this the "double descent phenomena" [23]. In the regime where deep neural networks operate, they are able

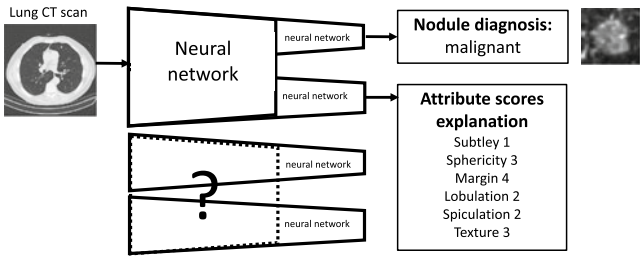

Fig. 1 A network for diagnosing lung nodules on a CT scan is based on a previously published work [20, 21]. The model contains an explanation branch in addition to the diagnosis branch, but it is not clear how the computations underlying each branch's output are related

to interpolate each data point in a "direct" way which does not exhibit the undershoot or overshoot which is typical of overfitting [1]. An illuminating example of direct fitting is given by Hasson et al. showing direct fitting of a parabolic function with noise [1]. The computations involved are clearly local—similar to nearest neighbors type computations—and the global trend ($y \propto x^2$) is not extracted. Because of this, there is clearly no hope for extrapolation. One the other hand, the model is flexible enough to fit any data. These observations call into question the popular idea that deep neural networks work by extracting high level features that are of particular interest, such as the whiskers of a cat. In actuality, it seems they are interpolating between a very large number of small "non-robust" features, some of which are particular to the training data [12]. It is tempting to tell "just-so" stories on how a deep neural network is functioning using explainability methods. These stories can mislead from what models are actually doing internally.

5 What Can Be Done?

The Rashomon effect (discussed above) suggests that in many cases, interpretable models may exist which have equivalent accuracy [6]. This seems to be especially the case with tabular data, where variants of linear regression often perform just as good as deep neural networks. However, in many other domains, deep neural networks are currently dominant. If the world is messy and complex, then neural networks trained on real-world data will also be messy and complex. Still, there are some models developed for images with interpretability in mind—one example is the "This Looks Like That" approach developed by Rudin et al. [4], which references parts of training examples.

Deep learning systems are notoriously bad at extrapolation, often failing spectacularly when small distributional shifts occur. For instance, a deep learning system for diagnosing retinopathy developed by Google's Verily Life Sciences which reached human-level diagnostic ability in the lab performed very poorly in field trials in Thai-

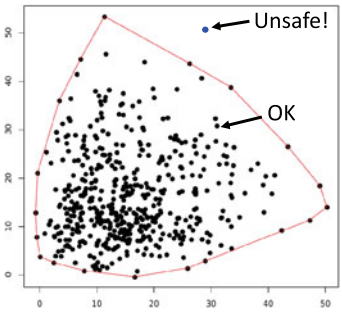

Fig. 2 Illustration of a simple approach to out-of-distribution detection

land due to poor lighting conditions, lower resolution images, and images which had been stitched together [24]. As another example, a deep learning system developed by *DeepMind* to play *Atari* games was shown to fail if minor changes are made, such as moving the paddle 3% higher in the game of Breakout [25]. In light of the fact that deep neural nets work through a sort of brute force fitting, this brittleness is not very surprising. If deep neural nets work by local interpolations over a massive number of data points, this implies an inability to extrapolate [1]. Explainability techniques can help understand how a model works within the dataset but they cannot help understand what happens when the network is asked to perform outside the training distribution. Fortunately, many techniques have been developed which can provide a "warning light" if a network is being asked to extrapolate. These techniques go under several different names—"applicability domain analysis", "out-of-distribution detection", "change point detection", and "outlier detection" [26]. A full analysis and comparison of the many different techniques that have been developed is outside the scope of this paper. A simple illustration of how many such methods work is shown in Fig. 2. In Fig. 2, the "domain of interpolation" is delineated by projecting the data into 2 dimensions and then looking at the convex hull of the data points. If the input/latent vector of a test data point (projected into 2D) falls outside the convex hull, then the model is extrapolating and a warning should be given. Typically a dataset will form one or more clusters when projected into a low dimensional space. It has been observed empirically that the average error for test data samples depends on how close to the center of the training data clusters the test point lies [27].

6 Conclusion

In this work, we discussed some of the motivations for explanation in deep learning systems and distinguished descriptive explanations from mechanistic ones. We believe mechanistic explanations to be the most important for increasing trust and ensuring robustness to distributional shift. However, recent work on double descent [22] and adversarial examples [12] indicate that mechanistic explanation is difficult, since deep neural networks operate by brute force interpolation over large datasets rather than by simple heuristics with high level features. We discussed a few

possible solutions to the issues raised—using explicitly interpretable models, adding an explanation branch, and implementing out-of-distribution detection methods.

Funding and Disclaimer No funding sources were used in the creation of this work. The author wrote this article in his own personal capacity. The opinions expressed in this article are the author's own and do not reflect the view of the National Institutes of Health, the Department of Health and Human Services, or the United States government.

References

1. Hasson U, Nastase SA, Goldstein A (2020) Direct fit to nature: an evolutionary perspective on biological and artificial neural networks. Neuron 105(3):416–434
2. Murdoch WJ, Singh C, Kumbier K, Abbasi-Asl R, Yu B (2019) Definitions, methods, and applications in interpretable machine learning. Proc Natl Acad Sci 116(44):22071–22080
3. Bansal G, Wu T, Zhou J, Fok R, Nushi B, Kamar E, Ribeiro MT, Weld DS (2020) Does the whole exceed its parts? the effect of AI explanations on complementary team performance. arXiv:2006.14779
4. Hase P, Bansal M (2010) Evaluating explainable AI: Which algorithmic explanations help users predict model behavior? arXiv:2005.01831
5. Chen C, Li O, Tao D, Barnett A, Rudin C, Su J (2019) This looks like that: Deep learning for interpretable image recognition. In: Wallach HM, Larochelle H, Beygelzimer A, d'Alché-Buc F, Fox EB, Garnett R (eds) Advances in neural information processing systems 32: annual conference on neural information processing systems 2019, NeurIPS 2019, 8–14 Dec 2019. Canada, Vancouver, BC, pp 8928–8939
6. Rudin C (2019) Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nat Mach Intell 1(5):206–215
7. Yeh CK, Hsieh CY, Suggala AS, Inouye DI, Ravikumar P (2019) On the (in)fidelity and sensitivity for explanations. arXiv:1901.09392
8. Adebayo J, Gilmer J, Muelly M, Goodfellow I, Hardt M, Kim B (2018) Sanity checks for saliency maps. In: Proceedings of the 32nd international conference on neural information processing systems, NIPS 18. Curran Associates Inc., Red Hook, NY, USA, 95259536p
9. Lie C (2019) Relevance in the eye of the beholder: diagnosing classifications based on visualised layerwise relevance propagation. Master's thesis, Lund Unversity, Sweden
10. Hooker S, Erhan D, Kindermans P, Kim B (2019) A benchmark for interpretability methods in deep neural networks. In: Wallach HM, Larochelle H, Beygelzimer A, d'Alché-Buc F, Fox EB, Garnett R (eds) Advances in neural information processing systems 32: NeurIPS 2019, 8–14 Dec 2019. Canada, Vancouver, BC, pp 9734–9745
11. Olah C, Cammarata N, Schubert L, Goh G, Petrov M, Carter S (2020) Zoom in: an introduction to circuits. Distill 5(3)
12. Ilyas A, Santurkar S, Tsipras D, Engstrom L, Tran B, Madry A (2019) Adversarial examples are not bugs, they are features. In: Wallach HM, Larochelle H, Beygelzimer A, d'Alché-Buc F, Fox EB, Garnett R (eds) Advances in neural information processing systems 32: annual conference on neural information processing systems 2019, NeurIPS 2019, 8–14 Dec 2019. Canada, Vancouver, BC, pp 125–136
13. Szegedy C, Zaremba W, Sutskever I, Bruna J, Erhan D, Goodfellow IJ, Fergus R (2014) Intriguing properties of neural networks. In: Bengio Y, LeCun Y (eds) 2nd International conference on learning representations, ICLR 2014, Banff, AB, Canada, 14–16 Apr 2014
14. Eitel F, Ritter K (2019) Testing the robustness of attribution methods for convolutional neural networks in MRI-based Alzheimer's disease classification. In: Interpretability of machine intelligence in medical image computing and multimodal learning for clinical decision support. Springer International Publishing, pp 3–11

15. Barnes BC, Elton DC, Boukouvalas Z, Taylor DE, Mattson WD, Fuge MD, Chung PW (2018) Machine learning of energetic material properties. arXiv:1807.06156
16. Sutre ET, Colliot O, Dormont D, Burgos N (2020) Visualization approach to assess the robustness of neural networks for medical image classification. In: Proceedings of the SPIE: medical imaging
17. Rieke J, Eitel F, Weygandt M, Haynes JD, Ritter K (2018) Visualizing convolutional networks for MRI-based diagnosis of Alzheimer's disease. In: Understanding and interpreting machine learning in medical image computing applications. Springer International Publishing, pp 24–31
18. Alvarez-Melis D, Jaakkola TS (2018) Towards robust interpretability with self-explaining neural networks. In: Proceedings of the 32nd international conference on neural information processing systems NIPS 18. Curran Associates Inc., Red Hook, NY, USA, 77867795p
19. Lillicrap TP, Kording KP (2019) What does it mean to understand a neural network? arXiv:1907.06374
20. Shen S, Han SX, Aberle DR, Bui AA, Hsu W (2019) An interpretable deep hierarchical semantic convolutional neural network for lung nodule malignancy classification. Expert Syst Appl 128:84–95
21. LaLonde R, Torigian D, Bagci U (2020) Encoding visual attributes in capsules for explainable medical diagnoses. In: Medical image computing and computer assisted intervention—MICCAI 2020. Springer International Publishing, pp 294–304
22. Elton DC (2020) Self-explaining AI as an alternative to interpretable AI. In: Artificial general intelligence. Springer International Publishing, pp 95–106
23. Belkin M, Hsu D, Ma S, Mandal S (2019) Reconciling modern machine-learning practice and the classical bias–variance trade-off. Proc Natl Acad Sci 116(32):15849–15854
24. Beede E, Baylor E, Hersch F, Iurchenko A, Wilcox L, Ruamviboonsuk P, Vardoulakis LM (2020) A human-centered evaluation of a deep learning system deployed in clinics for the detection of diabetic retinopathy. In: Proceedings of the 2020 CHI conference on human factors in computing systems. ACM
25. Kansky K, Silver T, Mély DA, Eldawy M, Lázaro-Gredilla M, Lou X, Dorfman N, Sidor S, Phoenix DS, George D (2017) Schema networks: Zero-shot transfer with a generative causal model of intuitive physics. In: Precup D, Teh YW (eds) Proceedings of the 34th international conference on machine learning, ICML 2017, Sydney, NSW, Australia, 6–11 Aug 2017 (Proceedings of machine learning research, vol 70, pp 1809–1818). PMLR (2017)
26. Hendrycks D, Mazeika M, Dietterich TG (2019) Deep anomaly detection with outlier exposure. In: 7th International conference on learning representations, ICLR 2019, New Orleans, LA, USA, 6–9 May 2019
27. Philipsen MP, Moeslund TB (2020) Prediction confidence from neighbors. arXiv:2003.14047