Explaining what learned models predict: In which cases can we trust machine learning models and when is caution required?

Athanasios D. Charisoudis
Undergraduate Electrical and Computer Engineering Student, Aristotle University of Thessaloniki
Student Member, IEEE

Introduction

Machine Learning (ML) models boast such predictive capabilities that hearing a novel or updated model that broke yet another benchmark really comes as no surprise. This is largely due to the surge of research the field enjoys, accompanied by constant advances in computer hardware. However, there remains a crucial question to be answered: To what extent can we trust these models? While being firmly convinced that machine intelligence constitutes the predominant direction in which solutions to several problems shall be searched, in this essay I will present my ideas on why blindly trusting ML models would not account for wise choice, yet.

During recent years, ML models such as Deep Neural Networks (DNN) relentlessly evolve, being found nowadays at the heart of numerous applications, ranging from Computer Vision and object detection up to Natural Language Processing and Understanding - in key areas like medicine and justice. In some cases, such as image classification, machines have exhibited performance better or on par with humans [6], [9] (at least in the absence of distortions [16] and low-SNR signals [17]). But as these models evolve, the more complex they become, making the task of understanding them and verifying their outputs more and more difficult. This, in turn, makes these models less trustworthy, posing a hurdle to their widespread adoption.

Definition and Importance of Explainability in Machine Learning Models

The process of explaining what learned models predict is itself an intricate one - it even lacks a concrete definition [14]. Nevertheless, there are increasingly more efforts in understanding what models - especially the ones trained in a supervised manner - predict and how they learn. This is also depicted in the Google Trends plot of the keyword "explainable ai" (figure 1) and is grounded on the strong correlation between a model's "transparency" (i.e. user explainability [7]) and its trustworthiness when applied to real-world scenarios and tasks [4], [15]. In what follows and according to [14], model explainability (or interpretability which is used interchangeably in this essay) is defined with respect to two principal desired properties, namely *Transparency* and *Post-hoc Interpretability*. Although

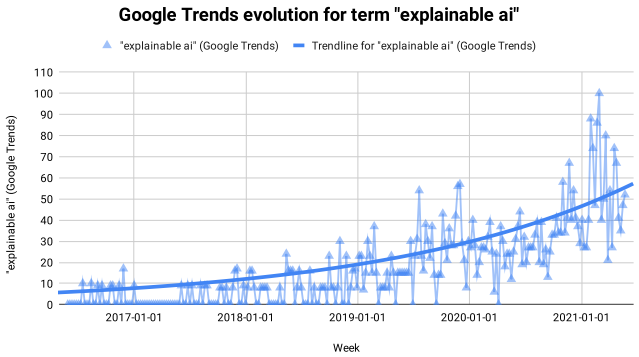


Figure 1: Google Trends plot for searches relating to the term "explainable ai". The thick trendline reveals the constantly-increasing interest of the research community around the Explainability and Interpretability of ML models.
Source: Google Trends, Google Inc. (https://trends.google.com)

the former ensures an intuitive explanation of the model's output and operations, it is so constraining considering the size and complexity of today's models, that the focus is almost exclusively on the latter to interpret machine/deep-learning models resembling black-boxes.

Before proceeding in reviewing some of the prevailing techniques in interpreting black-box models (i.e. Post-hoc interpretability), it would be worth noting some cases where trained models can be considered transparent and therefore trustworthy by design. Models that fall in this category are principally simple machine-learning models, such as linear regressors (e.g. the constrained least-squares or lasso regressor as in [3]) and simple decision trees that have been long-thought as more interpretable compared to deep neural networks. Classification trees [1] as another example are not only readily interpretable (since labels can be recursively assigned to all intermediate nodes, and therefore an explanation can be produced for every output by following the path and echoing labels) but may also help in inferring causal relationships in observational variables [18].

Post-hoc Interpretability of Trained Models

Rocketing of computing performance that modern parallelization hardware has allowed, has made machine learning models - particularly deep learning models - resemble

black boxes. Under no circumstances, however, should we categorize these models as not interpretable by default because doing so is like admitting that a great portion of today's frontiers in machine learning success and applicability should not be considered reliable. Consequently, many methods have been proposed in assessing model interpretability after the training phase (i.e. *post-hoc*) and without much digging into the lower-level mechanics of the model. Two major such groups of techniques are presented in the succeeding paragraphs, wherein the use of DNN as template models is assumed since the representations learned by such models are often incomprehensible to humans.

Explanations along with Model Outputs

Using surrogate models to explain predictions of deep networks has been a prevalent interpretability technique at the early stages of their adoption, remaining until today an important way to validate models and enhance their reliability. Initial endeavors exploited the explainability power of decision trees. In one such work, Craven et al. proposed at 1996 TREPAN [2], a method that uses a trained neural network as a black-box *for extracting comprehensible, symbolic representations* of how the model makes predictions. A more recent example of explainability with surrogate models was the joint training of a reinforcement learner and a Recurrent Neural Network (RNN) proposed by Kerning et al. [13]. In their approach, the learner model was trained to minimize an objective function while the RNN was used to map the model's state into a textual explanation describing the followed strategy.

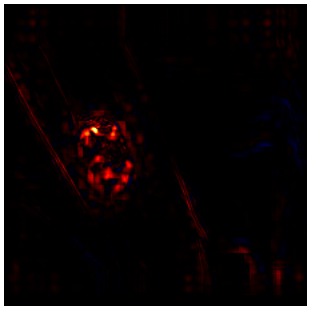
Probably a milestone in black-box model interpretability was a general technique proposed by Bach et al. for explaining predictions of Convolutional Neural Networks (CNN), known as Layer-wise Relevance Propagation (LRP) [10], [11], where the graph structure of DNN was exploited. In particular, after the output for a given image was computed they back-propagated the prediction scores to find which parts of the input image chiefly affected those scores. As can be seen in the figure 1, this technique provides valuable insights on how the predictions on a per-sample basis are made, thus helping in model validation and trust establishment.

Feature Visualizations of CNN

Another set of techniques to generate Post-hoc interpretations is focused on searching the input space to retrieve samples that maximize the activations of certain neurons or layers of DNN. Trained CNN, for instance, are known to be feature extractors [12], and therefore maximizing activation of learned convolutional filters is equivalent to finding features in observations that are recognized by those filters. Probably the most popular method for visualizing what image classifiers have learned is Activation Maximization [5], [8].

Activation Maximization fixes the weights of the entire network and tries to update the input image (i.e its pixel values) such that activation at a particular layer of the CNN is maximized. The basic idea is that by knowing what input maximally activates a layer, we may interpret and extract information on what this layer is trying to capture and consequently explain the model's output. This technique is better illustrated in the figure 3, where a specific neuron of a CNN classifier was found to by maximally active when an image of the digit "2" is given as input.

(a) Sample Input to a CNN image classifier



(b) Heatmap from relevance propagation of "ladybug" score

Figure 2: LRP demo on image classification. The right image, resulting from LRP, contributes significantly in understanding how the model reached its prediction, making it more easily verifiable and more user-friendly.
Source: Explainable AI Demos, Fraunhofer Institute for Telecommunications (https://lrpserver.hhi.fraunhofer.de)

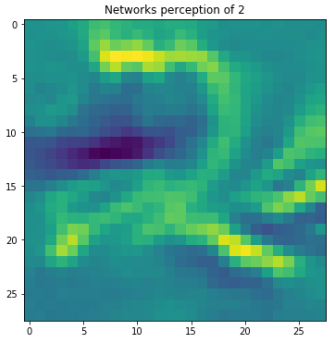


Figure 3: Visualization of the input image that maximizes activation of a particular neuron of a MNIST digits CNN classifier. It can be clearly seen that this particular neuron gets mostly activated when an image of the digit "2" enters the network. Activation-maximization here is employed in order to understand how the CNN classifier makes predictions.
Source: "Class activation maps: Visualizing neural network decision-making", Medium post by Anil Chandra Naidu Matcha, 2019

CONCLUSION

As more and more ML systems are being deployed in real-world applications and processes, the need for more transparent - or ideally self-explained - models is becoming apparent. And while the aforementioned methods for black-box interpretability have helped in extracting information on how models learn and predict there are still several limitations that prevent us from being confident enough that learned models decide in the same way a human does and that algorithmic errors or bias in ML have perished. All in all, I am inclined to believe that research in explainable

AI has to progress at the same pace as the one in the ML itself, so as for trustful and reliable models to be developed and to supersede humans in critical decision processes.

References

[1] D. H. Moore II, "Classification and regression trees, by leo breiman, jerome h. friedman, richard a. olshen, and charles j. stone. brooks/cole publishing, monterey, 1984,358 pages, $27.95," *Cytometry*, vol. 8, no. 5, pp. 534–535, 1987. DOI: https://doi.org/10.1002/cyto.990080516. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/cyto.990080516. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/cyto.990080516.

[2] M. Craven and J. Shavlik, "Extracting tree-structured representations of trained networks," in *Advances in Neural Information Processing Systems*, D. Touretzky, M. C. Mozer, and M. Hasselmo, Eds., vol. 8, MIT Press, 1996. [Online]. Available: https://proceedings.neurips.cc/paper/1995/file/45f31d16b1058d586fc3be7207b58053-Paper.pdf.

[3] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 58, no. 1, pp. 267–288, 1996, ISSN: 00359246. [Online]. Available: http://www.jstor.org/stable/2346178.

[4] G. Ridgeway, D. Madigan, T. Richardson, and J. O'Kane, "Interpretable boosted naive bayes classification," in *Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining*, ser. KDD'98, New York, NY: AAAI Press, 1998, pp. 101–104.

[5] D. Erhan, Y. Bengio, A. Courville, and P. Vincent, "Visualizing higher-layer features of a deep network," *Technical Report, Univeristé de Montréal*, Jan. 2009.

[6] D. Cireşan, U. Meier, and J. Schmidhuber, "Multi-column Deep Neural Networks for Image Classification," *arXiv e-prints*, arXiv:1202.2745, arXiv:1202.2745, Feb. 2012. arXiv: 1202.2745 [cs.CV].

[7] Y. Lou, R. Caruana, J. Gehrke, and G. Hooker, "Accurate intelligible models with pairwise interactions," in *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD '13, Chicago, Illinois, USA: Association for Computing Machinery, 2013, pp. 623–631, ISBN: 9781450321747. DOI: 10.1145/2487575.2487579. [Online]. Available: https://doi.org/10.1145/2487575.2487579.

[8] K. Simonyan, A. Vedaldi, and A. Zisserman, "Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps," *arXiv e-prints*, arXiv:1312.6034, arXiv:1312.6034, Dec. 2013. arXiv: 1312.6034 [cs.CV].

[9] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," *arXiv e-prints*, arXiv:1512.03385, arXiv:1512.03385, Dec. 2015. arXiv: 1512.03385 [cs.CV].

[10] S. Lapuschkin, A. Binder, G. Montavon, F. Klauschen, K.-R. Müller, and W. Samek, "On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation," *PLoS ONE*, vol. 10, e0130140, Jul. 2015. DOI: 10.1371/journal.pone.0130140.

[11] A. Binder, G. Montavon, S. Bach, K.-R. Müller, and W. Samek, "Layer-wise Relevance Propagation for Neural Networks with Local Renormalization Layers," *arXiv e-prints*, arXiv:1604.00825, arXiv:1604.00825, Apr. 2016. arXiv: 1604.00825 [cs.CV].

[12] Y. Chen, H. Jiang, C. Li, X. Jia, and P. Ghamisi, "Deep feature extraction and classification of hyperspectral images based on convolutional neural networks," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 54, no. 10, pp. 6232–6251, 2016. DOI: 10.1109/TGRS.2016.2584107.

[13] S. Krening, B. Harrison, K. Feigh, C. Isbell, M. Riedl, and A. Thomaz, "Learning from explanations using sentiment and advice in rl," *IEEE Transactions on Cognitive and Developmental Systems*, vol. PP, pp. 1–1, Nov. 2016. DOI: 10.1109/TCDS.2016.2628365.

[14] Z. C. Lipton, "The Mythos of Model Interpretability," *arXiv e-prints*, arXiv:1606.03490, arXiv:1606.03490, Jun. 2016. arXiv: 1606.03490 [cs.LG].

[15] M. Tulio Ribeiro, S. Singh, and C. Guestrin, "Why Should I Trust You? Explaining the Predictions of Any Classifier," *arXiv e-prints*, arXiv:1602.04938, arXiv:1602.04938, Feb. 2016. arXiv: 1602.04938 [cs.LG].

[16] S. Dodge and L. Karam, "A Study and Comparison of Human and Deep Learning Recognition Performance Under Visual Distortions," *arXiv e-prints*, arXiv:1705.02498, arXiv:1705.02498, May 2017. arXiv: 1705.02498 [cs.CV].

[17] R. Geirhos, D. H. J. Janssen, H. H. Schütt, J. Rauber, M. Bethge, and F. A. Wichmann, "Comparing deep neural networks against humans: object recognition when the signal gets weaker," *arXiv e-prints*, arXiv:1706.06969, arXiv:1706.06969, Jun. 2017. arXiv: 1706.06969 [cs.CV].

[18] S. M. Piryonesi and T. E. El-Diraby, "Data analytics in asset management: Cost-effective prediction of the pavement condition index," *Journal of Infrastructure Systems*, vol. 26, no. 1, p. 04 019 036, 2020. DOI: 10.1061/(ASCE)IS.1943-555X.0000512. eprint: https://ascelibrary.org/doi/pdf/10.1061/%28ASCE%29IS.1943-555X.0000512. [Online]. Available: https://ascelibrary.org/doi/abs/10.1061/%5C%28ASCE%5C%29IS.1943-555X.0000512.

Acronyms

AI Artificial Intelligence.
CNN Convolutional Neural Networks.
DNN Deep Neural Networks.
LRP Layer-wise Relevance Propagation.
ML Machine Learning.
MNIST Modified National Institute of Standards and Technology.
RNN Recurrent Neural Networks.
SNR Signal-to-Noise Ratio.

Athanasios D. Charisoudis Born in Thessaloniki, July 2nd, 1994. After primary and secondary education, I entered Electrical and Computer Engineering (E.C.E.) School at the Democritus University of Thrace. Two and a half years later, my position was transferred to the respective school at the Aristotle University of Thessaloniki, based on academic and financial criteria. Currently, I am in the last year of my undergraduate studies, finishing my diploma thesis, and close to acquiring the E.C.E. diploma (5-year school equivalent to combined BSc + MSc), with an expected final GPA of 8.9/10.0. My goals are to enter a competitive Master's and/or a Ph.D. program focused on Deep Learning and in particular Generative Modeling for Computer Vision applications.