Does Explainable Machine Learning Uncover the Black Box in Vision Applications?

Manish Narwaria*

Abstract

Machine learning (ML) in general and deep learning (DL) in particular has become an extremely popular tool in several vision applications (like object detection, super resolution, segmentation, object tracking etc.). Almost in parallel, the issue of explainability in ML (i.e. the ability to explain/elaborate the way a trained ML model arrived at its decision) in vision has also received fairly significant attention from various quarters. However, we argue that the current philosophy behind explainable ML suffers from certain limitations, and the resulting explanations may not meaningfully uncover black box ML models. To elaborate our assertion, we first raise a few fundamental questions which have not been adequately discussed in the corresponding literature. We also provide perspectives on how explainablity in ML can benefit by relying on more rigorous principles in the related areas.

I. The Need to Revisit Current Explainable ML Philosophy

The powerful modeling capabilities of ML/DL has fueled a transition from white box to black box modeling, in research, industry and education [1]. Consequently, despite the success of ML/DL in several applications, the issue of explainability (i.e. ability to understand how an ML model arrives at a decision/prediction) remains one of the bottlenecks in adopting ML/DL models to a wider canvas of applications. Accordingly, it is fairly well accepted that ML/DL models should be amenable to a level of scrutiny that goes beyond simple audit. Hence, explainable ML remains an active research area [2], [3] in both industry and academia. However, there appears a lack of discussion on whether the current philosophy behind explainable ML can help to achieve the eventual goal of uncovering black box models. This is especially required in vision applications as meaningful visualization/explanation can greatly facilitate more transparent deployment of ML/DL models in practice. In that context, we raise

*Department of Electrical Engineering, Indian Institute of Technology Jodhpur, NH 62, Surpura Bypass Rd, Karwar, Rajasthan 342037, India. e-mail: narwaria@iitj.ac.in

three important questions which in our opinion are fundamental to a rigorous discussion on explainable ML, and have not been adequately raised or addressed in the current literature. We also attempt to provide perspectives and some insights into these questions. Hence, the goal is not to criticize existing efforts on explainable ML. Rather it is to build upon them, and in the process raise awareness about possible issues whose mitigation may help to develop more understandable ML/DL models.

A. Explainable ML: A Prehoc Necessity or A Posthoc Evil?

A commonly accepted argument for the need of explainability in ML is based on how serious the consequences might be if the trained model makes an error. For instance, consider medical image analysis where an ML/DL algorithm erroneously classifies an image as "normal" when in fact there was say a tumor (or some other serious medical abnormality) present. Obviously, such an error can have serious medical implications. Likewise, errors made by an ML algorithm in the context of autonomous vehicles might also undeniably lead to disastrous consequences. On the other hand, errors made by ML/DL are considered less serious, from a practical view point, in several other vision applications such as object recognition in photos/videos on the web, visual content retrieval, gaming, vision based recommender systems, audiovisual communication etc. Such a dichotomous but seemingly logical philosophy has unfortunately lead to a proliferation of black box models in the first place. As a result, explainability has been possibly relegated to merely a *posthoc* analysis of the trained ML model rather than being be viewed as a fundamental and inherent concept in ML/DL algorithm design. It is therefore reasonable to ask if we should revisit the current philosophy by making explainability a pre-design necessity and not merely a post design baggage.

B. Do visual explanations generate new knowledge?

DL in particular has gained popularity since it achieves better prediction performance than traditional non ML and ML approaches, especially in many vision applications. This can be largely attributed to the fact that feature extraction in DL is typically implicit (i.e. does not require handcrafted features) and data dependent. In contrast, traditional approaches rely heavily on feature engineering (explicit modeling) and possible use of apriori domain knowledge. It is therefore natural to ask: what additional information was extracted and exploited by the DL model that was missed in the more traditional approach for solving the same problem? Stated differently, can the explanations obtained from trained ML/DL models bridge the seeming gap in terms of additional knowledge that DL supposedly exploits which lead to better performances in vision applications? In the context of this question, we note that several existing efforts attempt to "explain" how DL works in applications like image classification/recognition etc. by

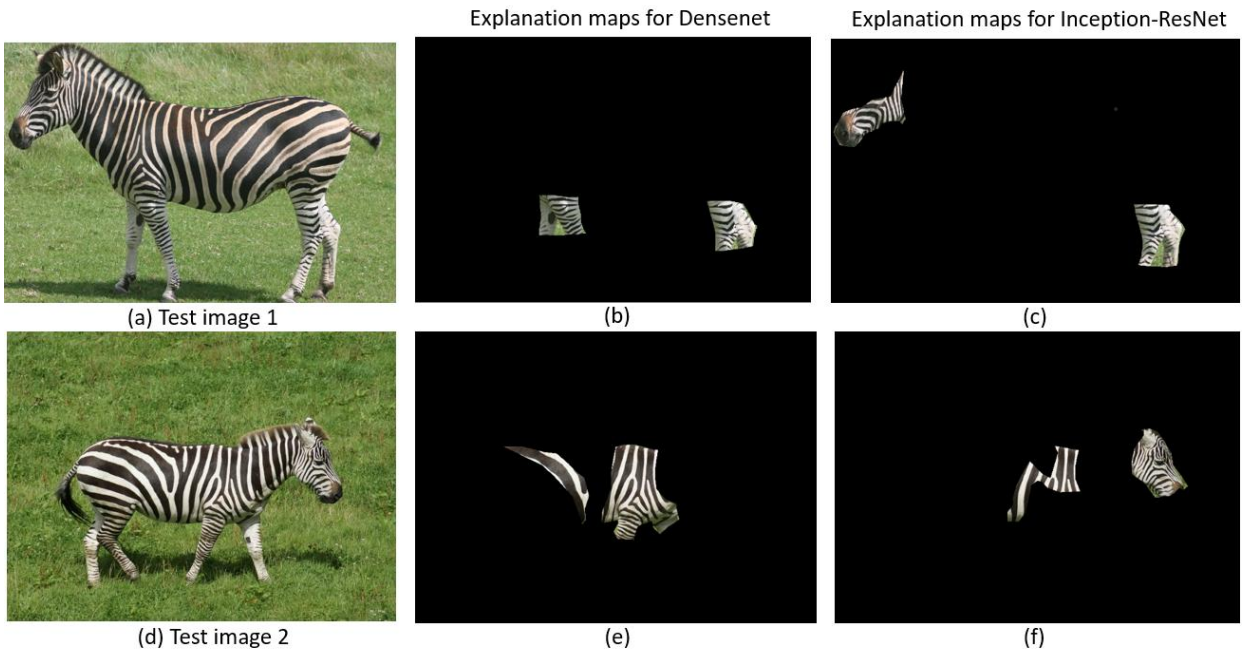

Fig. 1: Visual explanations generated for two test images. The explanations were generated by LIME method [4] for pre-trained Densenet and Inception-Resnet models. Only super pixels that correspond to top two features as determined by LIME are shown for easier visualization.

analyzing important regions (in the form of superpixels [4], salient regions [5] or similar ideas like layer-wise relevance propagation LRP [6] etc.). Hence, the main idea is to find regions in a test image (video) on which the trained DL network focuses on (or has relevance) for the purpose of classification/recognition. This information presumably explains how the ML/DL algorithm arrives at a decision. To analyze this in more details, it is convenient to consider the example in Figure 1 where we used two test images in which the main object of interest is a zebra. We then employed two pre trained models namely Densenet [7] and Inception-Resnet [8] to recognize the object in the test images. We found that both the models (pre-trained on nearly a million images from Imagnet [9]) correctly recognized the object namely 'zebra' (with a probability above 0.9) in both the test images. However, from an explainable ML view point, it will be obviously more interesting to understand visual signal information that the two DL models used to make the correct decision.

To that end, we employed LIME [4] to get further insights into the two black boxes. LIME identifies regions in the image (in the form of superpixels) which were more important for the ML/DL algorithm to arrive at the decision. In Figure 1, the second and third image in each row shows the two most important regions (visualized as explanation maps) that the respective DL models focused on to correctly classify

the image as ‘zebra’. While the 4 explanation maps in Figure 1 (b), (c), (e) and (f) correspond to some parts of the object (zebra), two key questions remain unanswered. First, do these specific image parts stand out (i.e. are different) in comparison to other image regions and if so how? Second, are these parts amenable to interpretation in terms of simpler signals or patterns similar to basis functions (or basic building blocks) for the object under consideration (zebra)?

C. Can visual explanations constitute a many-to-one mapping?

Another important issue in explainable ML is that of a many-to-one mapping. It means that a unique object (class label) can be potentially mapped to more than one explanations. For example, ‘zebra’ object in Figure 1 has 4 different explanations despite the fact that both the DL models recognized the object class correctly. In an analogous manner, we can conclude that a given DL model might not always rely on certain unique signal information (features) to correctly identify the same object. For instance, we observe from Figure 1 (b) that Densenet uses “legs” part of the zebra for first test image while it relies on middle body part (Figure 1 (e)) in case of the second test image. Likewise, while Inception-Resnet employs “legs and face” (Figure 1 (c)) region of the zebra for first test image but focuses on “ face and middle part of body” (Figure 1 (f)) in the second test image. That is, the two DL models under consideration correctly identify the same object but rely on very different visual signal information in the two test images in Figure 1. Moreover, the dimensionality of the mapped space (i.e. the number of explanations for the same object of interest) can increase quickly. Suppose we have m images of an object of interest whose class (label) is being predicted by say n ML models, and we employ p explanation methods. Obviously, unless we have unique explanations, we are potentially looking at $m \times n \times p$ explanations for the same object. This leads to the following question: how logical is it that ML models correctly recognize the same object present in more than one test image but characterize it differently (i.e. use very different explanations for each test image to arrive at the correct prediction)?

II. Explainable ML: A Possible Way Forward

We now attempt to provide some perspectives in the light of the questions raised. Before doing that, two points are however worth emphasizing. First, we note that the questions raised in the previous section are not influenced by our choice of specific test images in Figure 1 or the 2 DL models (Densenet and Inception-Resnet) or even the explanation method (LIME). Rather, the genesis of these questions lies primarily in the current philosophy behind explainable ML. Second, the said questions are not necessarily independent of each other. Thus, it is possible that focusing on inherent ML model explainability right from the start may reveal useful insights about knowledge generation and uniqueness of explanations.

A. Explainability as a first principle

With regard to the question raised in section I-A, it should be reasonable to conclude that explainability ought be an integral part of ML/DL algorithm design process itself and not an after thought. Thus, irrespective of the use-case scenario, an ML algorithm should be explainable/interpretable. Of course, the type of interpretability (i.e. the dimension) required can vary depending on the application (similar to how one views various tools in say signal processing). This could for instance involve learning of data-dependent but more intuitive and interpretable basis functions or representations. Likewise, ideas and design philosophy from signal processing, information theory and related disciplines can be borrowed. For example, fundamental ideas of frequency have been exploited to improve the interpretability of DL [10]. Other examples include the use of classical but interpretable concepts (such as filtering) in Digital Signal Processing (DSP) [11], in vision [12] or from physics [13], [14] etc. Similarly, exploitation of apriori knowledge such as logical rules and knowledge graphs [15] could lead to possibilities of enhancing our understanding of black box ML systems. Thus, it could be argued that irrespective of the approach eventually used, the design philosophy should focus on ML models that are inherently explainable. A *posthoc* analysis could of course always be employed as an add-on for further refinement or seeking application specific insights.

B. Possibility of Knowledge Generation and Transfer

Supervised ML/DL approach depends heavily on data to derive mapping from a visual signal to the task of interest (eg. recognition of objects from images). Because the training process potentially uses a large amount of labeled data (millions of images for instance in Imagnet [9]), it may not be unreasonable to expect that explanations of black box ML/DL models should reveal some new insights. However, it is likely that the current methods for explainablity, that tend to rely exclusively on posthoc analysis, are inadequate for this purpose. For instance, it is unclear if further analysis of the explanations (such as those in Figure 1) say in terms of explicit shape, texture, orientation or color will reveal any new insights about the object under consideration. Thus, the answers to the questions discussed in section I-B are probably in the negative. But perhaps more pertinently, these questions are meant to emphasize the importance of explanations that are meaningfully quantifiable and uniquely interpretable at least in the context of specific application (or a set of related applications). Such functionality may eventually open the possibility of generating potentially new and explicit knowledge which is open to scrutiny and amenable to knowledge transfer. A relevant example in this regard would be that of vision science which includes calibrated psycho visual experiments and computational modeling [16]. The findings of such experiments are not only fundamental but have had tremendous impact on conceptualization and meaningful advancement of

applications. This includes next generation video technologies like HFR [17], HDR [18], visual saliency [19], modern video compression [20], to list a few. On similar lines, a bottom-up and knowledge centric explainable ML design philosophy may be practically more useful and scalable.

C. Uniqueness of explanations

The keen reader will probably agree that the aspect of many-to-one mapping highlighted in section I-C should in general be problematic in the context of explainable ML. Specifically, it implies that:

1) An ML/DL algorithm may not be using unique visual information to correctly identify the same object present in more than one test image.
2) The method used to explain/interpret black box models may not be capturing the correct features (information) that is being used by the ML model to make predictions.
3) Two or more ML models with same prediction accuracies might be using very different explanations.

The first point indicates lack of consistency and reliability of the ML/DL model under consideration. The second refers to possible deficiencies in the method itself that was used to explain the ML model. The third point is interesting since it refers to non equivalence of two or more ML models having the same prediction performance. That is, it would be more logical that a *better* ML model should tend to use similar description (explanation) of the same object in more than one test image.

Thus, consistency of explanations in addition to prediction accuracy should be a more reasonable performance benchmark metric. In an ideal case, one could hope that explanations for same object are same/similar across more than one test images. This may provide an opportunity to simply characterize those explanations and generate *signatures* for different objects. In such case, deployment of ML models in practice would be greatly simplified since prediction could be done using logical rules and/or simpler models which in turn were derived from more complex DL/ML models trained on large sets of labeled data. Hence, uniqueness of explanations would increase ML model reliability and also open possibilities of new knowledge generation.

III. Concluding Remarks

There is no denying that ML models ought to be more transparent from the view point of signal information exploited to make predictions. We have, however, reasoned that the current philosophy behind explainable ML may not meaningfully uncover the black box. To that end, we highlighted three specific but not entirely independent aspects that probably deserve more attention. These include:

i. A generally accepted but probably unnecessary dichotomy wherein the need for explainability depends on use-case scenario. As a result, explainability might not be treated as an inherent concept but merely an add-on in the context of ML algorithm design.

ii. An apparent lack of clarity on what the explanations imply and how they could help to improve our existing knowledge about specific tasks (say object recognition). The expectation about new knowledge generation should not be unreasonable especially when a tremendous effort goes into large scale data collection and labeling, and eventually training the model on modern computing devices.

iii. The problem of many-to-one mapping where one class label (object) might be mapped to more than one explanations. Analogously, the same ML algorithm might use different features to identify the same object.

We also provided perspectives on the identified issues and possible ways of mitigating them. In particular, we emphasized that explainability should be accorded high priority in ML algorithm design process and not left merely as a *posthoc* exercise. This could, for instance, be facilitated by relying on philosophy of fundamental principles in related domains. As already pointed out, such approach has already been employed [10]-[15], and could prove to be the gateway to meaningfully more transparent ML models.

REFERENCES

[1] M. Narwaria, "The transition from white box to black box: Challenges and opportunities in signal processing education," *IEEE Signal Processing Magazine*, vol. 38, no. 3, pp. 163–173, 2021.

[2] N. Xie, G. Ras, M. van Gerven, and D. Doran, "Explainable deep learning: A field guide for the uninitiated," *arXiv*, 2020.

[3] V. Buhrmester, D. Mnch, and M. Arens, "Analysis of explainers of black box deep neural networks for computer vision: A survey," *arXiv*, 2019.

[4] M. T. Ribeiro, S. Singh, and C. Guestrin, ""why should i trust you?": Explaining the predictions of any classifier," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD '16. New York, NY, USA: Association for Computing Machinery, 2016, p. 11351144. [Online]. Available: https://doi.org/10.1145/2939672.2939778

[5] V. Petsiuk, A. Das, and K. Saenko, "RISE: randomized input sampling for explanation of black-box models," in *British Machine Vision Conference 2018, BMVC 2018, Newcastle, UK, September 3-6, 2018*. BMVA Press, 2018, p. 151. [Online]. Available: http://bmvc2018.org/contents/papers/1064.pdf

[6] S. Bach, 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, no. 7, p. e0130140, 07 2015.

[7] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 2261–2269.

[8] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi, "Inception-v4, inception-resnet and the impact of residual connections on learning," in *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, ser. AAAI'17. AAAI Press, 2017, p. 42784284.

[9] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 248–255.

[10] A. M. Tseng, A. Shrikumar, and A. Kundaje, "Fourier-transform-based attribution priors improve the interpretability and stability of deep learning models for genomics," in *Proceedings of the 34th International Conference on Neural Information Processing Systems*, ser. NIPS'20. Red Hook, NY, USA: Curran Associates Inc., 2020.

[11] J. Engel, L. Hantrakul, C. Gu, and A. Roberts, "DDSP: Differentiable digital signal processing," *CoRR*, 2020. [Online]. Available: http://arxiv.org/abs/1907.07374

[12] R. Chellappa, "The changing fortunes of pattern recognition and computer vision," *Image and Vision Computing*, vol. 55, pp. 3 – 5, 2016, recognizing future hot topics and hard problems in biometrics research. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S026288561630066X

[13] N. Thuerey, P. Holl, M. Mueller, P. Schnell, F. Trost, and K. Um, *Physics-based Deep Learning*. WWW, 2021. [Online]. Available: https://physicsbaseddeeplearning.org

[14] M. Kellman, "Physics-based learning for large-scale computational imaging," Ph.D. dissertation, EECS Department, University of California, Berkeley, Aug 2020. [Online]. Available: http://www2.eecs.berkeley.edu/Pubs/TechRpts/2020/EECS-2020-167.html

[15] K. Beckh, S. Müller, M. Jakobs, V. Toborek, H. Tan, R. Fischer, P. Welke, S. Houben, and L. von Rüden, "Explainable machine learning with prior knowledge: An overview," *CoRR*, vol. abs/2105.10172, 2021. [Online]. Available: https://arxiv.org/abs/2105.10172

[16] D. Marr, *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. USA: Henry Holt and Co., Inc., 1982.

[17] A. B. Watson, "High frame rates and human vision: A view through the window of visibility," *SMPTE Motion Imaging Journal*, vol. 122, no. 2, pp. 18–32, 2013.

[18] A. Chalmers and K. Debattista, "Hdr video past, present and future: A perspective," *Signal Processing: Image Communication*, vol. 54, pp. 49–55, 2017. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S092359651730019X

[19] M. Carrasco, "Visual attention: The past 25 years," *Vision Research*, vol. 51, no. 13, pp. 1484–1525, 2011, vision Research 50th Anniversary Issue: Part 2. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0042698911001544

[20] T. Guionnet, M. Raulet, and T. Burnichon, "Forward-looking content aware encoding for next generation uhd, hdr, wcg, and hfr," *SMPTE Motion Imaging Journal*, vol. 129, no. 7, pp. 26–32, 2020.