Interpretability of Deep Learning Models: A Survey of Results

Supriyo Chakraborty*, Richard Tomsett§, Ramya Raghavendra*, Daniel Harborne†, Moustafa Alzantot‡,
Federico Cerutti†, Mani Srivastava‡, Alun Preece†, Simon Julier††, Raghuveer M. Rao¶,
Troy D. Kelley¶, Dave Braines§, Murat Sensoy‖, Christopher J. Willis**, Prudhvi Gurram¶
*IBM T. J. Watson Research Center, †Crime and Security Research Institute, Cardiff University, ‡ UCLA, § IBM UK,
¶ Army Research Lab, Adelphi, ‖ Ozyegin University, ** BAE Systems AI Labs †† University College London

***Abstract*—Deep neural networks have achieved near-human accuracy levels in various types of classification and prediction tasks including images, text, speech, and video data. However, the networks continue to be treated mostly as black-box function approximations, mapping a given input to a classification output.**

The next step in this human-machine evolutionary process ? incorporating these networks into mission critical processes such as medical diagnosis, planning and control ? requires a level of trust association with the machine. To establish this trust, neural networks should provide greater visibility and human-understandable justifications for their decisions leading to better insights about the inner workings. We call such models as interpretable deep networks.

There are a multitude of dimensions that together constitute interpretability. In addition, the interpretation itself can be provided either in terms of the low-level network parameters, or in terms of input features used by the model. In this paper, we outline some of the dimensions that are useful for model interpretability, and categorize prior work along those dimensions. In the process, we perform a gap analysis of what needs to be done to improve model interpretability.

I. Introduction

Advances in machine learning and deep learning have had a profound impact on many "low-level" tasks such as object recognition and behaviour monitoring. Recently, researchers have begun to explore how these approaches can be used in "high-level" domains such as healthcare, criminal justice system, finance, and military decision making [1]. As the importance of the decisions aided using machine learning increases, it becomes more important for users to be able to suitably weight the assistance provided by such systems. A key property is *interpretability* — users should have the ability to *understand* and *reason* about the model output. However, despite several years of research effort, progress in this area remains limited [2]. For example, multi-layer neural networks, in spite of the their tremendous success in achieving near-human accuracy levels in certain prediction and classification tasks [3], operate as *black boxes*, and offer little to no explanation/visibility into why specific features are selected over others during training, or how are the correlations in the training data represented in the choice of the features, or why a specific pathway in the network (e.g., transforming raw data to classification output) is selected over others.

While deep learning based models are motivated by neuro-scientific advancements in the understanding of the working of the human brain, a critical distinction, that has often been made between the two, is attributed to the human ability to "think" [4]. Informally, it is this ability to *think*, that allows humans to not only make a prediction, but also *justify* or *rationalize* it through a series of logically consistent and *understandable* choices leading up to the prediction. This justification, in turn, enables the decision maker to implicitly or explicitly associate a measure of *confidence* to the prediction aiding the decision making process. The counterpart to the human thought process in deep learning models is often referred to as *interpretability* [2].

In fact, as observed in [2], the notion of interpretability is not even a monolithic concept but reflects several different dimensions which are summarized below:

- **Model Transparency:** This is defined in terms of three parameters: (i) *simulatability* – whether a human can use the input data together with the model to reproduce every calculation step necessary to make the prediction. This allows the human to understand the changes in the model parameters caused by the training data; (ii) *decomposability* – whether there is an intuitive explanation for all the model parameters; and finally (iii) *algorithmic transparency* – which is essentially an ability to explain the working of the learning algorithm. For example, the choice of a hyperplane in the Support Vector Machine (SVM) can be explained in terms of the marginal points and the decision boundary. However, for a deep neural network, the non-linearities added into the features at each layers makes it difficult to explain the features being used for the output.
- **Model Functionality:** This is defined in terms of (i) *textual description* – providing a semantically meaning description of the model output. To do so, one might use a composition of models, one for prediction and another one to generate a textual explanation; (ii) *visualization* – another common means of explaining the working of a model is through visualization of the parameters. One popular approach to visualize high-dimensional distributed representations is using the t-SNE mechanism [5]; and finally (iii) *local explanation* – where instead of explaining the entire mapping of a model, local changes introduced by a specific input vector for a given

output class is computed. Gradient of the output is used to identify specific weights and the local changes that are influenced by the input vector.

In this paper, we provide a brief exposition of a coalition setting in which we want to train an interpretable deep neural network and conclude by identifying challenges that are unique to this setting and their influence on model interpretability.

II. Summary of prior art

Please refer to the extended version of this paper for a summary of prior work [6].

III. A coalition perspective

We consider the problem of model interpretation within a coalition setting in which multiple disparate parties come together to forge an ad-hoc coalition geared towards achieving a common mission. Each party owns a slice of data but has policy-based constraints that places restrictions on the information that it can share with other coalition members. The success of the mission is thus contingent upon maximum utilization of this distributed data to build a common model shared among all the parties.

As is evident from the above setting, any decision made using the common model has to be adequately justified for it to be accepted by all the coalition members. Such a justification can only be generated using an interpretable model. In addition, it is quintessential that the common model is established as fair (i.e., unbiased), accountable and transparent to the coalition members. Finally, the policy-constraints within a coalition together with the non-homogeneity between the model architectures might make it difficult to use techniques such as layer-wise relevance propagation for interpretation.

IV. Discussion and Challenges

We now discuss in detail challenges unique to coalition and possible alternative approaches to providing interpretability.

A. Fairness and Accountability

With rapid adoption of machine learning techniques there has also been a growing recognition that the same techniques also raise novel ethical, policy, and legal challenges. In particular, policymakers, regulators, and advocates have expressed fears about the potentially discriminatory impact of data-driven learning approaches, especially the dangers of inadvertently encoding bias into automated decisions. At the same time, there is an increasing alarm that the complexity of machine learning and opaqueness of data mining processes may reduce the justification for consequential decisions to "the algorithm made me do it" or "this is what the model says."

Accountability may be viewed as the ability to inspect a model in post hoc, and make it available for human or algorithmic inspection. Many important decisions historically made by humans are now being made by algorithms, whose accountability measures and legal standards are far from satisfactory [7].

B. Interpretability versus Explainability

Computational models that impart reasoning behind their decisions, often use the terms "interpretability" and "explainability" synonymously [8], [9], [10]. This is true, even when the community acknowledges the need for clear taxonomy [2]. We would like to propose a differentiation between these terms and in doing so, we are able to clarify the process of forming testable metrics within the problem space.

When talking about the explainability of a model, we suggest that this refers specifically to the type and completeness of the output given when a model is queried for reasoning behind its decision. This means that explanations of the same type can be compared using a metric without need for any further context [11]. However, explanations of different types (saliency map images [12] and text captions for example [13]) can't be compared using a metric.

Acknowledgements

This research was sponsored by the U.S. ARL and the UK Ministry of Defence under Agreement Number W911NF-16-3-0001. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the U.S. ARL, the U.S. Government, the UK Ministry of Defence or the UK Government. The U.S. and UK Governments are authorized to reproduce and distribute reprints for Government purposes notwithstanding any copy-right notation hereon.

References

[1] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016, http://www.deeplearningbook.org.

[2] Z. C. Lipton, "The mythos of model interpretability," *CoRR*, vol. abs/1606.03490, 2016.

[3] Y. Lecun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 5 2015.

[4] B. M. Lake, T. D. Ullman, J. B. Tenenbaum, and S. J. Gershman, "Building machines that learn and think like people," *CoRR*, vol. abs/1604.00289, 2016.

[5] L. v. d. Maaten and G. Hinton, "Visualizing data using t-sne," *Journal of Machine Learning Research*, vol. 9, no. Nov, pp. 2579–2605, 2008.

[6] S. Chakraborty, R. Tomsett, R. Raghavendra, D. Harborne, M. Alzantot, F. Cerutti, M. Srivastava, A. D. Preece, S. Julier, R. M. Rao, T. D. Kelley, D. Braines, M. Sensoy, C. J. Willis, and P. Gurram, "Interpretability of deep learning models: a survey of results," in *DAIS 2017*, 2017.

[7] J. A. Kroll, J. Huey, S. Barocas, E. W. Felten, J. R. Reidenberg, D. G. Robinson, and H. Yu, "Accountable algorithms," 2017.

[8] P. W. Koh and P. Liang, "Understanding black-box predictions via influence functions," *arXiv preprint arXiv:1703.04730*, 2017.

[9] M. Bojarski, P. Yeres, A. Choromanska, K. Choromanski, B. Firner, L. D. Jackel, and U. Muller, "Explaining how a deep neural network trained with end-to-end learning steers a car," *arXiv preprint arXiv:1704.07911*, 2017.

[10] S. Jha, V. Raman, A. Pinto, T. Sahai, and M. Francis, "On learning sparse boolean formulae for explaining ai decisions," 2017.

[11] W. Samek, A. Binder, G. Montavon, S. Lapuschkin, and K.-R. Müller, "Evaluating the visualization of what a deep neural network has learned," *IEEE transactions on neural networks and learning systems*, 2017.

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

[13] L. A. Hendricks, Z. Akata, M. Rohrbach, J. Donahue, B. Schiele, and T. Darrell, "Generating visual explanations," in *European Conference on Computer Vision*. Springer, 2016, pp. 3–19.