arXiv:2206.04394v1 [cs.LG] 9 Jun 2022

Xplique
A Deep Learning Explainability Toolbox

Thomas Fel[1,3,4*] Lucas Hervier[2*]
David Vigouroux[2] Antonin Poche[2] Justin Plakoo[2] Remi Cadene[3] Mathieu Chalvidal[1,3]
Julien Colin[1,3] Thibaut Boissin[1,2] Louis Bethune[1] Agustin Picard[5,1] Claire Nicodeme[4]
Laurent Gardes[4] Gregory Flandin[1,2] Thomas Serre[1,3]

[1]Artificial and Natural Intelligence Toulouse Institute, Université de Toulouse, France
[2] Institut de Recherche Technologique Saint-Exupery, France
[3]Carney Institute for Brain Science, Brown University, USA
[4] Innovation & Research Division, SNCF , [5] Scalian

Abstract

*Today's most advanced machine-learning models are hardly scrutable. The key challenge for explainability methods is to help assisting researchers in opening up these black boxes — by revealing the strategy that led to a given decision, by characterizing their internal states or by studying the underlying data representation. To address this challenge, we have developed **Xplique**: a software library for explainability which includes representative explainability methods as well as associated evaluation metrics. It interfaces with one of the most popular learning libraries: Tensorflow as well as other libraries including PyTorch, scikit-learn and Theano. The code is licensed under the MIT license and is freely available at* `github.com/deel-ai/xplique`.

1. Introduction

Deep neural networks [27, 40] are widely used in many applications including medicine, transportation, security and finance, with broad societal implications [5, 23, 34]. Yet, these networks have become almost impenetrable. Furthermore, in most real-world scenarios, these systems are used to make critical decisions, often without any explanation. A growing body of research thus focuses on making those systems more trustworthy via the development of explainability methods to make their predictions more interpretable [8]. Such methods will find broad societal uses and will help to fulfill the "right to explanation" that European laws guarantee to its citizens [21]. Hence, it is important for explainability methods to be made widely available. Indeed, several libraries have already been proposed including Captum [25] for Pytorch.

CVPR 2022 Workshop on Explainable Artificial Intelligence for Computer Vision (XAI4CV)

In this work, we propose the first of such libraries – based on Tensorflow [1]. Our library includes all main explainability approaches including: (1) attribution methods (and their associated metrics), (2) feature visualization methods and (3) concept-based methods.

1.1. Attribution methods

aim to produce so-called saliency maps or more simply, heatmaps, to explain models' decisions. These maps reveal the discriminating input variables used by the system for arriving to a given decision. The score assigned to a region of an image (or a word in a sentence) reflects its importance for the prediction of the model. We have reimlpemented more than 14 representative explanation methods [2, 7, 9, 11, 13, 29, 33, 35, 38, 39, 41, 43, 44, 48–50]. We provide support for images, tabular data and time series. As one can imagine, the large number of explanation methods available has brought to the forefront a major issue: the urgent need for metrics to evaluate explanations. Indeed, inconsistencies produced across these methods have raised questions about their legitimacy [2–4, 6, 10, 12, 16, 17, 19, 20, 26, 28, 36, 42, 45, 46]. Our implementation thus also includes several common metrics associated with these attribution methods.

1.2. Feature Visualization

Even though attribution methods are sometimes useful to understand a decision, they leave aside the global study of a Deep Learning model. Several methods attempt to tackle this issue including feature visualization methods for studying the internal representations learned by a model.

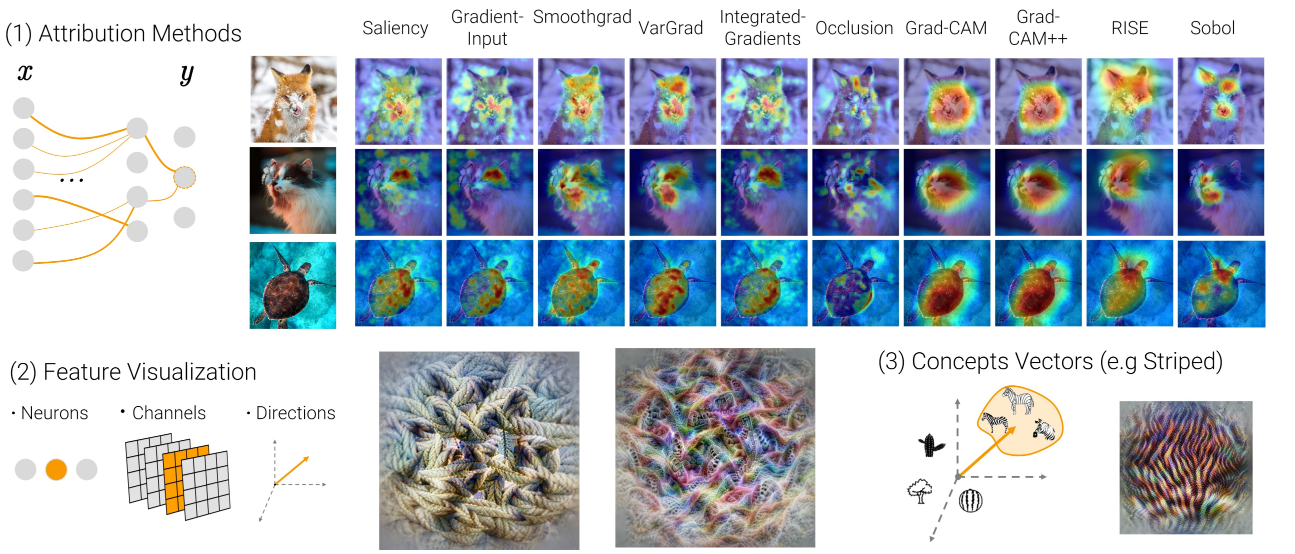

Figure 1. **Xplique modules.** The library contains 3 main modules: **(1)** an "Attribution Methods" module, **(2)** a "Feature Visualization" module and **(3)** a "Concepts" module .

The method proposed in [30–32] is a popular technique employed to explain the internal representations of a model. This method aims to find an interpretable input (or stimulus) that maximizes the response of a given neuron, a set of neurons (e.g., a channel) or a direction in an internal space of the model. Thus, the corresponding stimulus is a prototype of what the neuron responds to. We provide an API able to optimize such input by targeting a layer, a channel, a direction or combinations of these objectives. The optimization tool leverages the latest advances in the field (e.g., Fourier preconditioning, robustness to transformations).

1.3. Concept-based methods

Nevertheless, the interpretation of feature visualization methods is left to the user. Fortunately, another approach consists in letting the user derive concept vectors that are meaningful to them: Concept-based methods.

[14, 15, 18, 22, 24, 37, 47] work on high-level features interpretable by humans. This includes a method to retrieve Vectors of Activations of these human Concepts (CAV) [22]. These vectors help to make the passage between human concepts and a vector base formed by the neurons of a model at a specific layer. In addition, we have also re-implemented TCAV, which then tests how important these human vectors are to the model's decisions.

Finally, the library also allows interactions between all 3 modules such that one can leverage the feature visualization module to visualize the extracted CAV (see Fig.1) or the feature attribution module to visualize the location of the CAV on an image. A major effort has been made to facilitate the use of the software and various examples are provided as notebooks for each of the modules.

2. Acknowledgments

This work was conducted as part of the DEEL project[1]. Funding was provided by ANR-3IA Artificial and Natural Intelligence Toulouse Institute (ANR-19-PI3A-0004). Additional support provided by ONR grant N00014-19-1-2029 and NSF grant IIS-1912280. Support for computing hardware provided by Google via the TensorFlow Research Cloud (TFRC) program and by the Center for Computation and Visualization (CCV) at Brown University (NIH grant S10OD025181).

References

[1] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dandelion Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. 1

[2] Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. Sanity checks for saliency maps. In *Advances in Neural Information Processing Systems (NIPS)*, 2018. 1

[3] Marco Ancona, Enea Ceolini, Cengiz Öztireli, and Markus Gross. Towards better understanding of gradient-based at-

[1] https://www.deel.ai/

tribution methods for deep neural networks. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2018. 1

[4] Umang Bhatt, Adrian Weller, and José M. F. Moura. Evaluating and aggregating feature-based model explanations. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, 2020. 1

[5] Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*. PMLR, 2018. 1

[6] Diogo V. Carvalho, Eduardo M. Pereira, and Jaime S. Cardoso. Machine learning interpretability: A survey on methods and metrics. *Electronics*, 2019. 1

[7] Aditya Chattopadhay, Anirban Sarkar, Prantik Howlader, and Vineeth N Balasubramanian. Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2018. 1

[8] Finale Doshi-Velez and Been Kim. Towards a rigorous science of interpretable machine learning. *ArXiv e-print*, 2017. 1

[9] Thomas Fel, Remi Cadene, Mathieu Chalvidal, Matthieu Cord, David Vigouroux, and Thomas Serre. Look at the variance! efficient black-box explanations with sobol-based sensitivity analysis. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021. 1

[10] Thomas Fel, Julien Colin, Rémi Cadène, and Thomas Serre. What i cannot predict, i do not understand: A human-centered evaluation framework for explainability methods. *arXiv preprint arXiv:2112.04417*, 2021. 1

[11] Thomas Fel, Melanie Ducoffe, David Vigouroux, Remi Cadene, Mikael Capelle, Claire Nicodeme, and Thomas Serre. Don't lie to me! robust and efficient explainability with verified perturbation analysis. *arXiv preprint arXiv:2202.07728*, 2022. 1

[12] Thomas Fel and David Vigouroux. Representativity and consistency measures for deep neural network explanations. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2022. 1

[13] Ruth C. Fong and Andrea Vedaldi. Interpretable explanations of black boxes by meaningful perturbation. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017. 1

[14] Jessica Zosa Forde, Charles Lovering, George Konidaris, Ellie Pavlick, and Michael L Littman. Where, when & which concepts does alphazero learn? lessons from the game of hex. In *AAAI Workshop on Reinforcement Learning in Games*, 2022. 2

[15] Asma Ghandeharioun, Been Kim, Chun-Liang Li, Brendan Jou, Brian Eoff, and Rosalind W Picard. Dissect: Disentangled simultaneous explanations via concept traversals. *arXiv preprint arXiv:2105.15164*, 2021. 2

[16] Amirata Ghorbani, Abubakar Abid, and James Zou. Interpretation of neural networks is fragile. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2017. 1

[17] Leilani H. Gilpin, David Bau, Ben Z Yuan, Ayesha Bajwa, Michael Specter, and Lalana Kagal. Explaining explanations: An overview of interpretability of machine learning. In *Proceedings of the IEEE International Conference on data science and advanced analytics (DSAA)*, 2018. 1

[18] P Hitzler and MK Sarker. Human-centered concept explanations for neural networks. *Neuro-Symbolic Artificial Intelligence: The State of the Art*, 342:337, 2022. 2

[19] Sara Hooker, Dumitru Erhan, Pieter-Jan Kindermans, and Been Kim. A benchmark for interpretability methods in deep neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019. 1

[20] Cheng-Yu Hsieh, Chih-Kuan Yeh, Xuanqing Liu, Pradeep Ravikumar, Seungyeon Kim, Sanjiv Kumar, and Cho-Jui Hsieh. Evaluations and methods for explanation through robustness analysis. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2021. 1

[21] Margot E Kaminski. The right to explanation, explained. In *Research Handbook on Information Law and Governance*. Edward Elgar Publishing, 2021. 1

[22] Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, et al. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). In *International conference on machine learning*. Proceedings of the International Conference on Machine Learning (ICML), 2018. 2

[23] Svetlana Kiritchenko and Saif M Mohammad. Examining gender and race bias in two hundred sentiment analysis systems. *Proceedings of the 7th Joint Conference on Lexical and Computational Semantics (*SEM)*, 2018. 1

[24] Pang Wei Koh, Thao Nguyen, Yew Siang Tang, Stephen Mussmann, Emma Pierson, Been Kim, and Percy Liang. Concept bottleneck models. In *International Conference on Machine Learning*, pages 5338–5348. PMLR, 2020. 2

[25] Narine Kokhlikyan, Vivek Miglani, Miguel Martin, Edward Wang, Bilal Alsallakh, Jonathan Reynolds, Alexander Melnikov, Natalia Kliushkina, Carlos Araya, Siqi Yan, et al. Captum: A unified and generic model interpretability library for pytorch. *arXiv preprint arXiv:2009.07896*, 2020. 1

[26] Isaac Lage, Emily Chen, Jeffrey He, Menaka Narayanan, Been Kim, Sam Gershman, and Finale Doshi-Velez. An evaluation of the human-interpretability of explanation. In *Workshop on Correcting and Critiquing Trends in Machine Learning, Advances in Neural Information Processing Systems (NIPS)*, 2019. 1

[27] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 2015. 1

[28] Zhong Qiu Lin, Mohammad Javad Shafiee, Stanislav Bochkarev, Michael St Jules, Xiao Yu Wang, and Alexander Wong. Do explanations reflect decisions? a machine-centric strategy to quantify the performance of explainability algorithms. In *Advances in Neural Information Processing Systems (NIPS)*, 2019. 1

[29] Scott Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems (NIPS)*, 2017. 1

[30] Anh Nguyen, Jason Yosinski, and Jeff Clune. Multifaceted feature visualization: Uncovering the different types of features learned by each neuron in deep neural networks. *Visualization for Deep Learning workshop, Proceedings of the International Conference on Machine Learning (ICML)*, 2016. 2

[31] Anh Nguyen, Jason Yosinski, and Jeff Clune. Understanding neural networks via feature visualization: A survey. *arXiv preprint arXiv:1904.08939*, 2019. 2

[32] Chris Olah, Alexander Mordvintsev, and Ludwig Schubert. Feature visualization. *Distill*, 2017. 2

[33] Vitali Petsiuk, Abir Das, and Kate Saenko. Rise: Randomized input sampling for explanation of black-box models. In *Proceedings of the British Machine Vision Conference (BMVC)*, 2018. 1

[34] Inioluwa Deborah Raji and Joy Buolamwini. Actionable auditing: Investigating the impact of publicly naming biased performance results of commercial ai products. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pages 429–435, 2019. 1

[35] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "why should i trust you?": Explaining the predictions of any classifier. In *Knowledge Discovery and Data Mining (KDD)*, 2016. 1

[36] Laura Rieger and Lars Kai Hansen. Irof: a low resource evaluation metric for explanation methods. In *Workshop, Proceedings of the International Conference on Learning Representations (ICLR)*, 2020. 1

[37] Jessica Schrouff, Sebastien Baur, Shaobo Hou, Diana Mincu, Eric Loreaux, Ralph Blanes, James Wexler, Alan Karthikesalingam, and Been Kim. Best of both worlds: local and global explanations with human-understandable concepts. *arXiv e-prints*, pages arXiv–2106, 2021. 2

[38] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017. 1

[39] Junghoon Seo, Jeongyeol Choe, Jamyoung Koo, Seunghyeon Jeon, Beomsu Kim, and Taegyun Jeon. Noise-adding methods of saliency map as series of higher order partial derivative. In *Workshop on Human Interpretability in Machine Learning, Proceedings of the International Conference on Machine Learning (ICML)*, 2018. 1

[40] Thomas Serre. Deep learning: The good, the bad, and the ugly. *Annual review of vision science*, 2019. 1

[41] Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. Learning important features through propagating activation differences. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2017. 1

[42] Leon Sixt, Maximilian Granz, and Tim Landgraf. When explanations lie: Why many modified bp attributions fail. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2020. 1

[43] Jost Tobias Springenberg, Alexey Dosovitskiy, Thomas Brox, and Martin Riedmiller. Striving for simplicity: The all convolutional net. In *Workshop Proceedings of the International Conference on Learning Representations (ICLR)*, 2014. 1

[44] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2017. 1

[45] Richard Tomsett, Dan Harborne, Supriyo Chakraborty, Prudhvi Gurram, and Alun Preece. Sanity checks for saliency metrics. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2019. 1

[46] Chih-Kuan Yeh, Cheng-Yu Hsieh, Arun Sai Suggala, David I. Inouye, and Pradeep Ravikumar. On the (in)fidelity and sensitivity for explanations. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019. 1

[47] Chih-Kuan Yeh, Been Kim, Sercan Arik, Chun-Liang Li, Tomas Pfister, and Pradeep Ravikumar. On completeness-aware concept-based explanations in deep neural networks. *Advances in Neural Information Processing Systems*, 33:20554–20565, 2020. 2

[48] Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *Proceedings of the IEEE European Conference on Computer Vision (ECCV)*, 2014. 1

[49] Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *Proceedings of the IEEE European Conference on Computer Vision (ECCV)*, 2014. 1

[50] M. D. Zeiler, G. W. Taylor, and R. Fergus. Adaptive deconvolutional networks for mid and high level feature learning. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2011. 1