Journal of Machine Learning Research 24 (2023) 1-11 Submitted 2/22; Revised 11/22; Published 1/23

Quantus: An Explainable AI Toolkit for Responsible Evaluation of Neural Network Explanations and Beyond

Anna Hedström[1,†] ANNA.HEDSTROEM@TU-BERLIN.DE
Leander Weber[3] LEANDER.WEBER@HHI.FRAUNHOFER.DE
Dilyara Bareeva[1] DILYARA.BAREEVA@CAMPUS.TU-BERLIN.DE
Daniel Krakowczyk[4] DANIEL.KRAKOWCZYK@UNI-POTSDAM.DE
Franz Motzkus[3] FRANZ.MOTZKUS@HHI.FRAUNHOFER.DE
Wojciech Samek[2,3,5] WOJCIECH.SAMEK@HHI.FRAUNHOFER.DE
Sebastian Lapuschkin[3,†] SEBASTIAN.LAPUSCHKIN@HHI.FRAUNHOFER.DE
Marina M.-C. Höhne[1,5,†] MARINA.HOEHNE@TU-BERLIN.DE
[1] *Understandable Machine Intelligence Lab, TU Berlin, 10587 Berlin, Germany*
[2] *Department of Electrical Engineering and Computer Science, TU Berlin, 10587 Berlin, Germany*
[3] *Department of Artificial Intelligence, Fraunhofer Heinrich-Hertz-Institute, 10587 Berlin, Germany*
[4] *Department of Computer Science, University of Potsdam, 14476 Potsdam, Germany*
[5] *BIFOLD – Berlin Institute for the Foundations of Learning and Data, 10587 Berlin, Germany*
[†] *corresponding authors*

Editor: Joaquin Vanschoren

Abstract

The evaluation of explanation methods is a research topic that has not yet been explored deeply, however, since explainability is supposed to strengthen trust in artificial intelligence, it is necessary to systematically review and compare explanation methods in order to confirm their correctness. Until now, no tool with focus on XAI evaluation exists that exhaustively and speedily allows researchers to evaluate the performance of explanations of neural network predictions. To increase transparency and reproducibility in the field, we therefore built `Quantus`—a comprehensive, evaluation toolkit in Python that includes a growing, well-organised collection of evaluation metrics and tutorials for evaluating explainable methods. The toolkit has been thoroughly tested and is available under an open-source license on PyPi (or on `https://github.com/understandable-machine-intelligence-lab/Quantus/`).

Keywords: explainability, responsible AI, reproducibility, open source, Python

1. Introduction

Despite much excitement and activity in the field of eXplainable artificial intelligence (XAI) (Montavon et al., 2018; Arya et al., 2019; Lapuschkin et al., 2019; Samek et al., 2021; Bykov et al., 2021b), the evaluation of explainable methods still remains an unsolved problem (Samek et al., 2017; Adebayo et al., 2020; Holzinger et al., 2020; Yona and Greenfeld, 2021; Arras et al., 2022). Unlike in traditional machine learning (ML), the task of *explaining* generally lacks "ground-truth" data. There exists no universally accepted definition of what

a "correct" explanation is, or what properties an explanation should fulfil (Yang and Kim, 2019). Due to this lack of standardised evaluation procedures in XAI, researchers frequently conceive new ways to experimentally examine explanation methods (Bach et al., 2015; Samek et al., 2017; Adebayo et al., 2018; Yang and Kim, 2019; Kindermans et al., 2019), oftentimes employing different parameterisations and various kinds of preprocessing and normalisations, each leading to different or even contrasting results, making evaluation outcomes difficult to interpret and compare. Critically, we note that it is common for XAI papers to base their conclusions on one-sided, sometimes methodologically questionable evaluation procedures, which we fear may hinder access to the current State-of-the-art (SOTA) in XAI and potentially hurt the perceived credibility of the field over time.

For these reasons, researchers often rely on a qualitative evaluation of explanation methods (e.g., Zeiler and Fergus (2014); Ribeiro et al. (2016); Shrikumar et al. (2017)). Although qualitative evaluation of XAI methods is an important and complementary type of evaluation analysis (Hoffman et al., 2018), the assumption that humans are able to recognise a correct explanation comes with a series of pitfalls: not only does the notion of an "accurate" explanation often depend on the specifics of the task at hand, humans are also questionable judges of quality (Wang et al., 2019; Rosenfeld, 2021). In addition, recent studies suggest that even quantitative evaluation of explainable methods is far from fault-proof (Bansal et al., 2020; Budding et al., 2021; Yona and Greenfeld, 2021; Hase and Bansal, 2020). In response to these issues, we developed `Quantus`, to provide the community with a versatile and comprehensive toolkit that collects, organises, and explains a wide range of evaluation metrics proposed for explanation methods. The library is designed to help automate the process of *XAI quantification*—by delivering speedy, easily digestible, and at the same time holistic summaries of the quality of the given explanations. As we see it, `Quantus` concludes an important, still missing contribution in today's XAI research by filling the gap between what the community produces and what it currently needs: a more quantitative, systematic and standardised evaluation of explanation methods.

2. Toolkit Overview

`Quantus` provides its intended users—practitioners and researchers interested in the domains of ML and XAI—with a steadily expanding list of 30+ reference metrics to evaluate explanations of ML predictions. Moreover, it offers comprehensive guidance on how to use these metrics, including information about potential pitfalls in their application.

Table 1: Comparison of four XAI libraries—(`AIX360` (Arya et al., 2019), `captum` (Kokhlikyan et al., 2020), `TorchRay` (Fong et al., 2019) and `Quantus`) in terms of the number of XAI evaluation methods for six different evaluation categories, as implemented in each library.

Library	Faithfulness	Robustness	Localisation	Complexity	Axiomatic	Randomisation
`Captum` (2)	1	1	0	0	0	0
`AIX360` (2)	2	0	0	0	0	0
`TorchRay` (1)	0	0	1	0	0	0
`Quantus` (27)	**9**	**4**	**6**	**3**	**3**	**2**

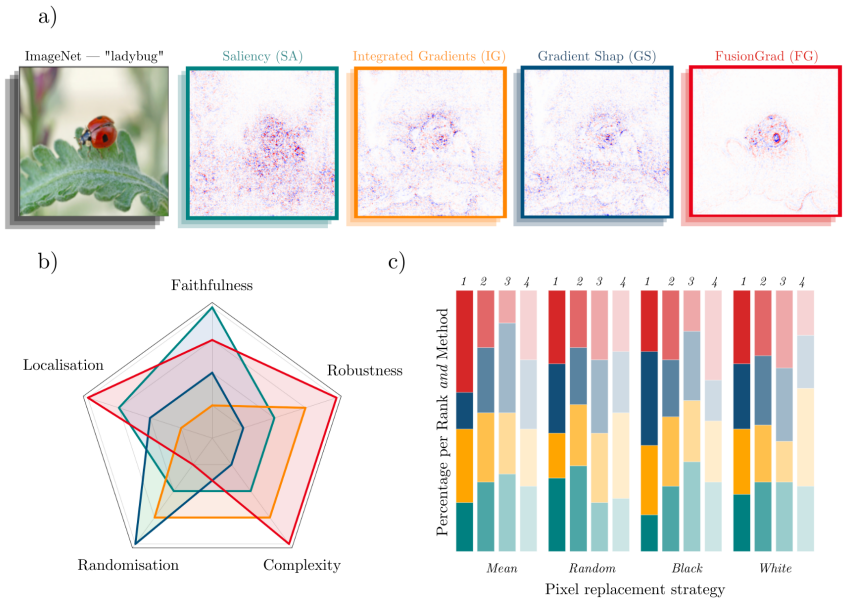

Figure 1: *a)* Simple *qualitative* comparison of XAI methods is often not sufficient to distinguish which gradient-based method—Saliency (Mørch et al., 1995; Baehrens et al., 2010), Integrated Gradients (Sundararajan et al., 2017), GradientShap (Lundberg and Lee, 2017) or FusionGrad (Bykov et al., 2021a) is preferred. With `Quantus`, we can obtain richer insights on how the methods compare *b)* by holistic quantification on several evaluation criteria and *c)* by providing sensitivity analysis of how a single parameter, e.g., pixel replacement strategy of a faithfulness test influences the ranking of explanation methods.

The library is thoroughly documented and includes tutorials covering multiple use-cases, data domains and tasks—from comparative analysis of XAI methods and attributions, to quantifying the extent evaluation outcomes are dependent on metrics' parameterisations. In Figure 1, we demonstrate some example analysis using ImageNet dataset (Russakovsky et al., 2015) that can be produced with `Quantus`[1]. The library provides an abstract layer between APIs of deep learning frameworks, e.g., `PyTorch` (Paszke et al., 2019) and `tensorflow` (Abadi et al., 2016) and can be employed iteratively both during and after model training. Code quality is ensured by thorough testing, using `pytest` and continuous integration (CI), where every new contribution is automatically checked for sufficient test coverage. We employ syntax formatting with `flake8`, `mypy` and `black` under various Python versions.

Unlike other XAI-related libraries[2], `Quantus` has its primary focus on evaluation and as such, supports a breadth of metrics, spanning various evaluation categories (see Table 1). A detailed description of the different evaluation categories can be found in the Appendix. The first iterations of the library mainly focus on attribution-based explanation techniques[3] for

1. The full experiment can be reproduced (and obtained) at the repository, under the `\tutorials` folder.
2. Related libraries were selected with respect to the XAI evaluation capabilities. Packages including no metrics for evaluation of explanation methods, e.g., `Alibi` (Klaise et al., 2021), `iNNvestigate` (Alber et al., 2019), `dalex` (Baniecki et al., 2021) and `zennit` (Anders et al., 2021) were excluded.
3. This category of explainable methods aims to assign an importance value to the model features and arguably, is the most studied group of explanations.

(but not limited to) image classification. In planned future releases, we are working towards extending the applicability of the library further, e.g., by developing additional metrics and functionality that will enable users to perform checks, verifications and sensitivity analyses on top of the metrics.

3. Library Design

The user-facing API of `Quantus` is designed with the aim of replacing an oftentimes lengthy and open-ended evaluation procedure with structure and speed—with a single line of code, the user can gain quantitative insights of how their explanations are behaving under various criteria. In the following code snippet, we demonstrate one way for how `Quantus` can be used to evaluate pre-computed explanations via a `PixelFlipping` experiment (Bach et al., 2015). In this example, we assume to have a pre-trained model (`model`), a batch of input and output pairs (`x_batch`, `y_batch`) and a set of attributions (`a_batch`).

```
import quantus
pixelflipping = quantus.PixelFlipping(perturb_baseline="black", abs=True)
scores = pixelflipping(model, x_batch, y_batch, a_batch, **params)
pixelflipping.plot(y_batch=y_batch, scores=scores)
```

Needless to say, XAI evaluation is intrinsically difficult and there is no one-size-fits-all metric for all tasks. Evaluation of explanations must, therefore, be understood and calibrated from its context: the application, data, model, and intended stakeholders (Chander and Srinivasan, 2018; Arras et al., 2022). To this end, we designed `Quantus` to be highly customisable and easily extendable—API documentation and examples on how to create new metrics as well as how to customise existing ones are included. Thanks to the API, any supporting functions of the evaluation procedure, e.g., `perturb_baseline` that determines the value that the input features should be iteratively masked with, can flexibly be replaced by a user-specified function to ensure that the evaluation procedure is appropriately contextualised.

It is practically well-known but not yet publicly recognised that evaluation outcomes of explanations can be highly sensitive to the parameterisation of metrics (Bansal et al., 2020; Agarwal and Nguyen, 2020) and other confounding factors introduced in the evaluation procedure (Hase et al., 2021; Yona and Greenfeld, 2021). Therefore, to encourage a thoughtful and responsible selection and parameterisation of metrics, we added mechanisms such as warnings, checks and user guidelines, cautioning users to reflect upon their choices.

4. Broader Impact

We built `Quantus` to raise the bar of *XAI quantification*—to substitute an ad-hoc and sometimes ineffective evaluation procedure with reproducibility, simplicity and transparency. From our perspective, `Quantus` contributes to the XAI development by helping researchers to speed up the development and application of explanation methods, dissolve existing ambiguities and enable more comparability. As we see it, steering efforts towards increasing objectiveness of evaluations and reproducibility in the field will prove rewarding for the community as a whole. We are convinced that a holistic, multidimensional take on XAI quantification will be imperative to the general success of (X)AI over time.

Acknowledgments

This work was partly funded by the German Ministry for Education and Research through project Explaining 4.0 (ref. 01IS20055) and BIFOLD (ref. 01IS18025A and ref. 01IS18037A), the Investitionsbank Berlin through BerDiBA (grant no. 10174498), as well as the European Union's Horizon 2020 programme through iToBoS (grant no. 965221).

Appendix

In most explainability contexts, ground-truth explanations are not available (Samek et al., 2017; Adebayo et al., 2020; Holzinger et al., 2020; Yona and Greenfeld, 2021; Arras et al., 2022), which makes the task of evaluating explanations non-trivial. Efforts on evaluating explanations have therefore been invested diversely. For better organisation, in the source code of `Quantus`, we therefore grouped the metrics into six categories based on their logical similarity—(a) faithfulness, (b) robustness, (c) localisation, (d) complexity, (e) randomisation and (f) axiomatic metrics.

In the following, we describe each of the categories briefly. A more in-depth description of each category, including an account of the underlying metrics, is documented in the repository. The direction of the arrow indicates whether higher or lower values are considered better (exceptions within each category exist, so please carefully read the docstrings of each individual metric prior to usage and/or interpretation).

(a) *Faithfulness* (↑) quantifies to what extent explanations follow the predictive behaviour of the model, asserting that more important features affect model decisions more strongly (Bhatt et al., 2020; Alvarez-Melis and Jaakkola, 2018; Arya et al., 2019; Nguyen and Martínez, 2020; Bach et al., 2015; Samek et al., 2017; Montavon et al., 2018; Ancona et al., 2018; Rieger and Hansen, 2020; Yeh et al., 2019; Rong et al., 2022; Dasgupta et al., 2022)

(b) *Robustness* (↓) measures to what extent explanations are stable when subject to slight perturbations in the input, assuming that the model output approximately stayed the same (Yeh et al., 2019; Montavon et al., 2018; Alvarez-Melis and Jaakkola, 2018; Dasgupta et al., 2022)

(c) *Localisation* (↑) tests if the explainable evidence is centred around a region of interest, which may be defined around an object by a bounding box, a segmentation mask or a cell within a grid (Zhang et al., 2018; Theiner et al., 2022; Kohlbrenner et al., 2020; Arras et al., 2022; Rong et al., 2022; Arias-Duart et al., 2021)

(d) *Complexity* (↓) captures to what extent explanations are concise, i.e., that few features are used to explain a model prediction (Chalasani et al., 2020; Bhatt et al., 2020; Nguyen and Martínez, 2020)

(e) *Randomisation* (↑) tests to what extent explanations deteriorate as the data labels or the model, e.g., its parameters are increasingly randomised (Adebayo et al., 2018; Sixt et al., 2020)

(f) *Axiomatic* (↑) measures if explanations fulfill certain axiomatic properties (Kindermans et al., 2019; Sundararajan et al., 2017; Nguyen and Martínez, 2020)

References

Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Gregory S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian J. Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Józefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dan Mané, Rajat Monga, Sherry Moore, Derek Gordon Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul A. Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda B. Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. Tensorflow: Large-scale machine learning on heterogeneous distributed systems, 2016.

Julius Adebayo, Justin Gilmer, Michael Muelly, Ian J. Goodfellow, Moritz Hardt, and Been Kim. Sanity checks for saliency maps. In Samy Bengio, Hanna M. Wallach, Hugo Larochelle, Kristen Grauman, Nicolò Cesa-Bianchi, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, pages 9525–9536, 2018.

Julius Adebayo, Michael Muelly, Ilaria Liccardi, and Been Kim. Debugging tests for model explanations. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin, editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020.

Chirag Agarwal and Anh Nguyen. Explaining image classifiers by removing input features using generative models. In Hiroshi Ishikawa, Cheng-Lin Liu, Tomás Pajdla, and Jianbo Shi, editors, *Computer Vision - ACCV 2020 - 15th Asian Conference on Computer Vision, Kyoto, Japan, November 30 - December 4, 2020, Revised Selected Papers, Part VI*, volume 12627 of *Lecture Notes in Computer Science*, pages 101–118. Springer, 2020.

Maximilian Alber, Sebastian Lapuschkin, Philipp Seegerer, Miriam Hägele, Kristof T. Schütt, Grégoire Montavon, Wojciech Samek, Klaus-Robert Müller, Sven Dähne, and Pieter-Jan Kindermans. innvestigate neural networks! *J. Mach. Learn. Res.*, 20:93:1–93:8, 2019.

David Alvarez-Melis and Tommi S. Jaakkola. Towards robust interpretability with self-explaining neural networks. In Samy Bengio, Hanna M. Wallach, Hugo Larochelle, Kristen Grauman, Nicolò Cesa-Bianchi, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, pages 7786–7795, 2018.

Marco Ancona, Enea Ceolini, Cengiz Öztireli, and Markus Gross. Towards better understanding of gradient-based attribution methods for deep neural networks. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada,*

April 30 - May 3, 2018, Conference Track Proceedings. OpenReview.net, 2018. URL https://openreview.net/forum?id=Sy21R9JAW.

Christopher J. Anders, David Neumann, Wojciech Samek, Klaus-Robert Müller, and Sebastian Lapuschkin. Software for dataset-wide xai: From local explanations to global insights with zennit, corelay, and virelay, 2021.

Anna Arias-Duart, Ferran Parés, Dario Garcia-Gasulla, and Victor Gimenez-Abalos. Focus! rating xai methods and finding biases. *CoRR*, abs/2203.02928, 2021. doi: 10.48550/arXiv.2109.15035.

Leila Arras, Ahmed Osman, and Wojciech Samek. Clevr-xai: A benchmark dataset for the ground truth evaluation of neural network explanations. *Information Fusion*, 81:14–40, 2022.

Vijay Arya, Rachel K. E. Bellamy, Pin-Yu Chen, Amit Dhurandhar, Michael Hind, Samuel C. Hoffman, Stephanie Houde, Q. Vera Liao, Ronny Luss, Aleksandra Mojsilovic, Sami Mourad, Pablo Pedemonte, Ramya Raghavendra, John T. Richards, Prasanna Sattigeri, Karthikeyan Shanmugam, Moninder Singh, Kush R. Varshney, Dennis Wei, and Yunfeng Zhang. One explanation does not fit all: A toolkit and taxonomy of AI explainability techniques, 2019.

Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller, and Wojciech Samek. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PloS one*, 10(7), 2015.

David Baehrens, Timon Schroeter, Stefan Harmeling, Motoaki Kawanabe, Katja Hansen, and Klaus-Robert Müller. How to explain individual classification decisions. *J. Mach. Learn. Res.*, 11:1803–1831, 2010.

Hubert Baniecki, Wojciech Kretowicz, Piotr Piatyszek, Jakub Wisniewski, and Przemyslaw Biecek. dalex: Responsible machine learning with interactive explainability and fairness in python. *J. Mach. Learn. Res.*, 22:214:1–214:7, 2021.

Naman Bansal, Chirag Agarwal, and Anh Nguyen. SAM: the sensitivity of attribution methods to hyperparameters. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 8670–8680. Computer Vision Foundation / IEEE, 2020.

Umang Bhatt, Adrian Weller, and José M. F. Moura. Evaluating and aggregating feature-based model explanations. In Christian Bessiere, editor, *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI 2020*, pages 3016–3022. ijcai.org, 2020.

Céline Budding, Fabian Eitel, Kerstin Ritter, and Stefan Haufe. Evaluating saliency methods on artificial data with different background types. *CoRR*, abs/2112.04882, 2021.

Kirill Bykov, Anna Hedström, Shinichi Nakajima, and Marina M.-C. Höhne. Noisegrad: enhancing explanations by introducing stochasticity to model weights. *CoRR*, abs/2106.10185, 2021a.

Kirill Bykov, Marina M.-C. Höhne, Adelaida Creosteanu, Klaus-Robert Müller, Frederick Klauschen, Shinichi Nakajima, and Marius Kloft. Explaining bayesian neural networks. *CoRR*, abs/2108.10346, 2021b.

Prasad Chalasani, Jiefeng Chen, Amrita Roy Chowdhury, Xi Wu, and Somesh Jha. Concise explanations of neural networks using adversarial training. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pages 1383–1391. PMLR, 2020.

Ajay Chander and Ramya Srinivasan. Evaluating explanations by cognitive value. In Andreas Holzinger, Peter Kieseberg, A Min Tjoa, and Edgar R. Weippl, editors, *Machine Learning and Knowledge Extraction - Second IFIP TC 5, TC 8/WG 8.4, 8.9, TC 12/WG 12.9 International Cross-Domain Conference, CD-MAKE 2018, Hamburg, Germany, August 27-30, 2018, Proceedings*, volume 11015 of *Lecture Notes in Computer Science*, pages 314–328. Springer, 2018.

Sanjoy Dasgupta, Nave Frost, and Michal Moshkovitz. Framework for evaluating faithfulness of local explanations. *CoRR*, abs/2202.00734, 2022. URL `https://arxiv.org/abs/2202.00734`.

Ruth Fong, Mandela Patrick, and Andrea Vedaldi. Understanding deep networks via extremal perturbations and smooth masks, 2019.

Peter Hase and Mohit Bansal. Evaluating explainable AI: which algorithmic explanations help users predict model behavior? In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel R. Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 5540–5552. Association for Computational Linguistics, 2020.

Peter Hase, Harry Xie, and Mohit Bansal. The out-of-distribution problem in explainability and search methods for feature importance explanations. *Advances in Neural Information Processing Systems*, 34, 2021.

Robert R. Hoffman, Shane T. Mueller, Gary Klein, and Jordan Litman. Metrics for explainable AI: challenges and prospects. *CoRR*, abs/1812.04608, 2018.

Andreas Holzinger, André M. Carrington, and Heimo Müller. Measuring the quality of explanations: The system causability scale (SCS). *Künstliche Intell.*, 34(2):193–198, 2020.

Pieter-Jan Kindermans, Sara Hooker, Julius Adebayo, Maximilian Alber, Kristof T. Schütt, Sven Dähne, Dumitru Erhan, and Been Kim. The (un)reliability of saliency methods. In Wojciech Samek, Grégoire Montavon, Andrea Vedaldi, Lars Kai Hansen, and Klaus-Robert Müller, editors, *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*, volume 11700 of *Lecture Notes in Computer Science*, pages 267–280. Springer, 2019.

Janis Klaise, Arnaud Van Looveren, Giovanni Vacanti, and Alexandru Coca. Alibi explain: Algorithms for explaining machine learning models. *J. Mach. Learn. Res.*, 22:181:1–181:7, 2021.

Maximilian Kohlbrenner, Alexander Bauer, Shinichi Nakajima, Alexander Binder, Wojciech Samek, and Sebastian Lapuschkin. Towards best practice in explaining neural network decisions with LRP. In *2020 International Joint Conference on Neural Networks, IJCNN 2020, Glasgow, United Kingdom, July 19-24, 2020*, pages 1–7. IEEE, 2020.

Narine Kokhlikyan, Vivek Miglani, Miguel Martin, Edward Wang, Bilal Alsallakh, Jonathan Reynolds, Alexander Melnikov, Natalia Kliushkina, Carlos Araya, Siqi Yan, and Orion Reblitz-Richardson. Captum: A unified and generic model interpretability library for pytorch, 2020.

Sebastian Lapuschkin, Stephan Wäldchen, Alexander Binder, Grégoire Montavon, Wojciech Samek, and Klaus-Robert Müller. Unmasking clever hans predictors and assessing what machines really learn. *CoRR*, abs/1902.10178, 2019.

Scott M. Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 4765–4774, 2017.

Grégoire Montavon, Wojciech Samek, and Klaus-Robert Müller. Methods for interpreting and understanding deep neural networks. *Digit. Signal Process.*, 73:1–15, 2018.

Niels J. S. Mørch, Ulrik Kjems, Lars Kai Hansen, Claus Svarer, Ian Law, Benny Lautrup, Stephen C. Strother, and Kelly Rehm. Visualization of neural networks using saliency maps. In *Proceedings of International Conference on Neural Networks (ICNN'95), Perth, WA, Australia, November 27 - December 1, 1995*, pages 2085–2090. IEEE, 1995.

An-phi Nguyen and María Rodríguez Martínez. On quantitative aspects of model interpretability. *CoRR*, abs/2007.07584, 2020. URL `https://arxiv.org/abs/2007.07584`.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Z. Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 8024–8035. NeurIPS, 2019.

Marco Túlio Ribeiro, Sameer Singh, and Carlos Guestrin. "why should I trust you?": Explaining the predictions of any classifier. In Balaji Krishnapuram, Mohak Shah, Alexander J. Smola, Charu C. Aggarwal, Dou Shen, and Rajeev Rastogi, editors, *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13-17, 2016*, pages 1135–1144. ACM, 2016.

Laura Rieger and Lars Kai Hansen. IROF: a low resource evaluation metric for explanation methods. *CoRR*, abs/2003.08747, 2020. URL `https://arxiv.org/abs/2003.08747`.

Yao Rong, Tobias Leemann, Vadim Borisov, Gjergji Kasneci, and Enkelejda Kasneci. Evaluating feature attribution: An information-theoretic perspective. *CoRR*, abs/2202.00449, 2022.

Avi Rosenfeld. Better metrics for evaluating explainable artificial intelligence. In Frank Dignum, Alessio Lomuscio, Ulle Endriss, and Ann Nowé, editors, *AAMAS '21: 20th International Conference on Autonomous Agents and Multiagent Systems, Virtual Event, United Kingdom, May 3-7, 2021*, pages 45–50. ACM, 2021.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael S. Bernstein, Alexander C. Berg, and Li Fei-Fei. Imagenet large scale visual recognition challenge. *Int. J. Comput. Vis.*, 115(3): 211–252, 2015.

Wojciech Samek, Alexander Binder, Grégoire Montavon, Sebastian Lapuschkin, and Klaus-Robert Müller. Evaluating the visualization of what a deep neural network has learned. *IEEE Trans. Neural Networks Learn. Syst.*, 28(11):2660–2673, 2017.

Wojciech Samek, Grégoire Montavon, Sebastian Lapuschkin, Christopher J. Anders, and Klaus-Robert Müller. Explaining deep neural networks and beyond: A review of methods and applications. *Proc. IEEE*, 109(3):247–278, 2021.

Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. Learning important features through propagating activation differences. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, volume 70 of *Proceedings of Machine Learning Research*, pages 3145–3153. PMLR, 2017.

Leon Sixt, Maximilian Granz, and Tim Landgraf. When explanations lie: Why many modified BP attributions fail. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pages 9046–9057. PMLR, 2020.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, volume 70 of *Proceedings of Machine Learning Research*, pages 3319–3328. PMLR, 2017.

Jonas Theiner, Eric Müller-Budack, and Ralph Ewerth. Interpretable semantic photo geolocation. In *IEEE/CVF Winter Conference on Applications of Computer Vision, WACV 2022, Waikoloa, HI, USA, January 3-8, 2022*, pages 1474–1484. IEEE, 2022.

Danding Wang, Qian Yang, Ashraf M. Abdul, and Brian Y. Lim. Designing theory-driven user-centric explainable AI. In Stephen A. Brewster, Geraldine Fitzpatrick, Anna L. Cox, and Vassilis Kostakos, editors, *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, CHI 2019, Glasgow, Scotland, UK, May 04-09, 2019*, page 601. ACM, 2019.

Mengjiao Yang and Been Kim. Benchmarking Attribution Methods with Relative Feature Importance. *CoRR*, abs/1907.09701, 2019.

Chih-Kuan Yeh, Cheng-Yu Hsieh, Arun Sai Suggala, David I. Inouye, and Pradeep Ravikumar. On the (in)fidelity and sensitivity of explanations. In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 10965–10976, 2019.

Gal Yona and Daniel Greenfeld. Revisiting sanity checks for saliency maps. *CoRR*, abs/2110.14297, 2021.

Matthew D. Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In David J. Fleet, Tomás Pajdla, Bernt Schiele, and Tinne Tuytelaars, editors, *Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part I*, volume 8689 of *Lecture Notes in Computer Science*, pages 818–833. Springer, 2014.

Jianming Zhang, Sarah Adel Bargal, Zhe Lin, Jonathan Brandt, Xiaohui Shen, and Stan Sclaroff. Top-down neural attention by excitation backprop. *Int. J. Comput. Vis.*, 126(10): 1084–1102, 2018.