Evaluating AI Explanations with Quantus

Fig. 1: The Quantus Logo. Quantus is an Open Source XAI evaluation library for Python hosted on https://github.com/understandable-machine-intelligence-lab/Quantus

The transparency of Artificial Intelligence (AI) models is an essential criterion for the deployment of AI in high-risk settings, such as medical applications. Consequently, numerous approaches for explaining AI systems have been proposed over the years (Samek et al., 2021). However, with a multitude of eXplainable AI (XAI) approaches at one's disposal, finding an answer to the question which method is most suitable for the application at hand is difficult to answer. This depends on a variety of factors, for example, whether the XAI method is compatible with the model to be explained, and beyond that, whether the *aspect* of the model's reasoning explained by the explainer fulfils the stakeholder's requirement. Once those points are dealt with, there still remains the question which method truly is the "best" choice.

Unfortunately, it often seems to be a common practice to make this choice based on qualitative criteria imposed on the explaining attribution maps, for example whether the explanation "looks good" or whether it represents some form of human expectation or intuition on how the model *should* make use of the available input features. Generally, it cannot be expected that a machine learning model trained on a limited pool of example data will use the same reasoning as a human after a lifetime of generalized learning across multiple sensory input domains. Furthermore, the formulation of a behavioural ground truth for machine learning models, being a difficult task on its own, might not apply to models of different architectures optimized on the same training data. If quantitative evaluations are made, there are many possible metrics available (faithfulness, complexity, localization, etc.), the results of which can be contradictory (e.g., the most faithful explanation can also be the most complex one). Thus, papers presenting quantitative results are often difficult to compare. In summary, picking the optimal explainer for the model at hand is a highly non-trivial, potentially error-prone and time-consuming task.

In order to help answering the question which explainer is best suited on a case-by-case basis, with as much objective information affecting the outcome, the AI Department of Fraunhofer Heinrich-Hertz-Institute, together with friends from the Berlin Institute of Technology, brings you **Quantus**: An Explainable AI Toolkit for Responsible Evaluation of Neural Network Explanations (see Fig. 1).

Quantus is a comprehensive, open-source toolkit in Python that includes a growing, well-organized collection of evaluation metrics (currently there are 27 metrics provided, from different evaluation categories) and tutorials for evaluating explainable methods. Quantus aims at unifying and objectivizing XAI evaluation procedures, by increasing reproducibility of XAI evaluation by providing an easy-to-use interface for automating extensive evaluation pipelines with only a few lines of code.

As a result, Quantus is able to provide rich summaries and comparisons of different prediction explanations, over various metrics, highlighting the relative strengths and weaknesses of (maybe apparently similar) attribution techniques at a glance, as shown in Figure 2. Quantus helps to get the "metric zoo" under control.

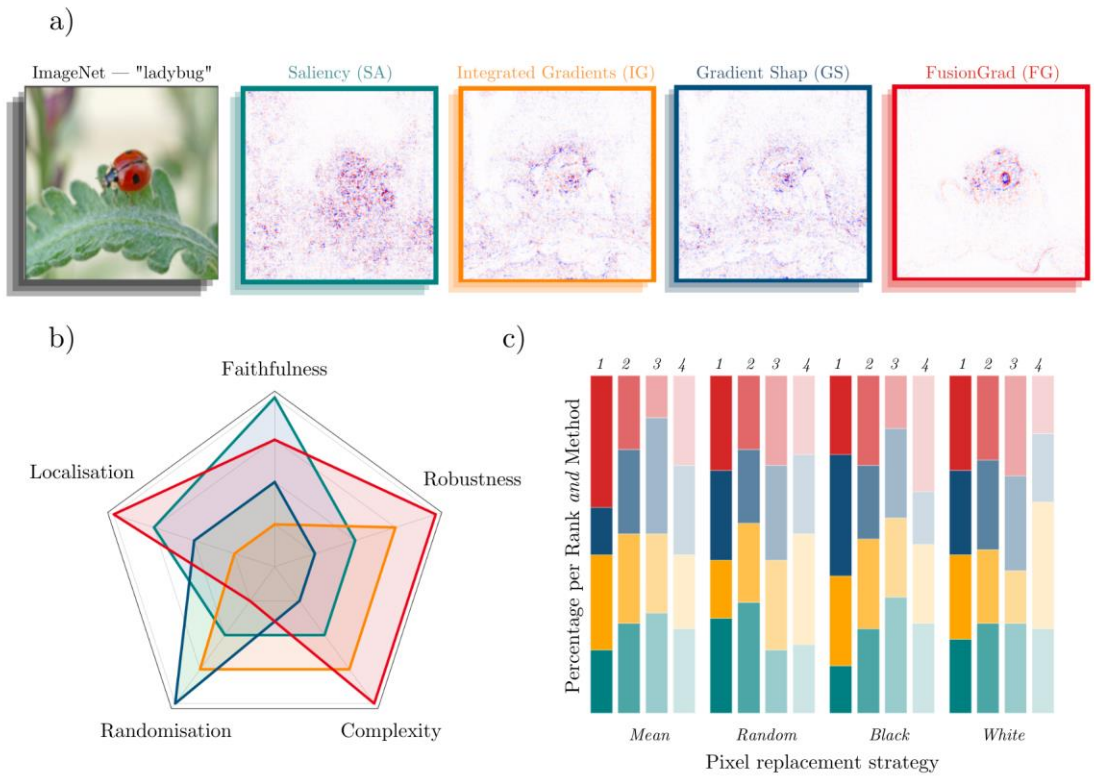

Fig. 2. With Quantus, we can obtain richer insights on how the methods compare a) for four gradient-based methods, e.g, b) by holistic quantification on several evaluation criteria and c) by providing sensitivity analysis of how a single parameter e.g., the pixel replacement strategy of a faithfulness test influences the ranking of the XAI methods.

While being agnostic to deep learning frameworks by design, Quantus is compatible with both tensorflow- and Pytorch models for the ad-hoc computation of explaining attribution maps. Quantus is available as an Open-Source implementation on https://github.com/understandable-machine-intelligence-lab/Quantus and comes with an introducing companion paper (Hedström et al. 2022).

Quantus allows its user to compile rich and information-dense summaries and comparisons of XAI methods, presenting the different strengths and weaknesses studied XAI methods in different explainable aspects in context of the relevant application.

Relevance to IToBoS

In iToBoS, many different AI systems will be trained for specific tasks, which in combination will culminate in an “AI Cognitive Assistant”. All those systems will need to be explained with suitable XAI approaches to elucidate all possible and required aspects of the systems’

decision making. With Quantus, we will be able to make well-informed choices regarding the application of XAI components throughout the project's machine learning core.

Authors

Sebastian Lapuschkin, Fraunhofer Heinrich-Hertz-Institute

Frederik Pahde, Fraunhofer Heinrich-Hertz-Institute

References

Samek W., Montavon G., Lapuschkin S., Anders C.J., Müller K.R.: „Explaining Deep Neural Networks and Beyond: A Review of Methods and Applications“. Proceedings of the IEEE 109 (3), pp 247-278

Hedström A., Weber L., Bareeva D., Motzkus F., Samek W., Lapuschkin S., Höhne M. M.-C.: „Quantus: An Explainable AI Toolkit for Responsible Evaluation of Neural Network Explanations “. arXiv preprint arxiv:2202.06861