Investigation of the evaluation techniques and tools used for model-specific XAI models

Tanguy Marbot, 5614643 : T.P.Y.M.Marbot@student.tudelft.nl
Supervisor : Chhagan Lal. Responsible Professor: Mauro Conti
April - June 2022, Research Project | CSE30000

TUDelft

1. Background

- Within Artificial Intelligence (AI), the exponential rise of different models has opened many opportunities for assistance and automation, for example in self-automated driving, in the medical sector and media, with an urgent need to address issues such as transparency, trust and accountability in case of harmful impact.
- Additionally, they might need to make critical decisions in dangerous situations. In general, the black-box nature of AI models inhibits the access to crucial information.

2. Motivation and Research Question

- Explainable AI (XAI) techniques have been developed to tackle such issues by providing interpretability and safety. The recent research interests in this domain has led to a development of a taxonomy where two fundamental categories are distinguished: model-agnostic and model-specific techniques. The first type applies when the technique can be used generally. The second category is for techniques in focus of a AI technique in particular. It is thus worthwhile to focus research to one domain and evaluate the efficiency of different techniques in respect of intended purposes. This might help users to choose an appropriate technique depending on context. Our question is: **How are model-specific XAI techniques evaluated ?**

	Task	Subject	Cost	Evaluation metrics
Functionally-grounded	Proxy	Automated	Lower	Fidelity, Robustness, Correctness, Safety, Architectural complexity, Expressiveness
Human-grounded	Proxy	Humans	Higher	Simulatability, Trust, Preference, Comprehensibility, Time Efficiency, Amount of information, Debuggability, Model Validation, Time Efficiency
Application-grounded	Application Interactions	Humans	Highest	Performance, Satisfaction, Persuasiveness, Human Judgement, Novelty

Table 1: Taxonomy of the evaluation with three categories. Task refers to what is being directly assessed, Proxy meaning a mediate task to assess a specific property. The subject refers to the agent for the task.

Technique name and overview	Evaluation		
	Functionally-Grounded	Human-Grounded	Application-Grounded
TCAV Uses linear classification in any layer, and directional derivatives to achieve the quantification of the classification sensitivity of a concept given by a user through example samples.	**Fidelity**.	**Simulatability**.	No evaluation.
SIDU. Localizes entire object regions responsible for prediction. Applies to CNN models by using convolutional layer and mask generation.	**Fidelity** with causal tools of deletion and insertion that are measured using AUC. **Robustness**.	**Simulatability**, results are cross-examined with other XAI techniques using mathematical tools.	**Human Judgement** of medical experts.
ACE. An automated use of TCAV. Takes images of same class as input and a trained CNN to build image segments and cluster them together. TCAV will compute the importance score of the segments	**Importance** (can be understood as a form of **fidelity**) using Smallest sufficient concepts (SSC) or Smallest destroying concepts (SDC)	**Coherency** (is judged as a form of **comprehensibility**). **Meaningfulness** (is judged as a form of **comprehensibility**)	No evaluation
Net2Vec. Makes use of combination of filters which responses can construct vectorial embeddings to which semantic concepts can be mapped to.	**Fidelity** using IoU (Intersection over Union).	No evaluation	No evaluation
Concept Analysis with ILP. Derives symbolic knowledge in the inner layers of a DNN model and uses an ILP model to build explanation in the form of first-order rules	**Fidelity** of concepts importance using IoU metric and first-order rules explanations using accuracy and F1 metrics	No evaluation	No evaluation

Table 2: Evaluation of five model-specific techniques with their overview

3. Methodology

-We need to investigate the different metrics that can be assessed to ensure the expected behaviour of criteria in a literature review and compare their importance and trade-offs.
- Then, we can investigate the evaluation of five state-of-the art model-specific methods with regards of those metrics. Table 2 summarizes the techniques.
- Given the insights on the first part of the investigation, we can compare the five evaluation processes and make judgements.

4. Results

- Three types of metrics: functionally grounded, human-grounded and application grounded. Table 1 summarizes the differences and metrics for each category.
- We identified trade-offs between interpretability and fidelity and the importance of evaluating several criteria.
- Most evaluations are based on functionally grounded metrics as human and application grounded metrics are costly. Fidelity is the prioritized metric. Table 2 summarizes the evaluation of the five techniques.
- Tools for evaluation are diverse and can new ones can be defined for a specific technique.

5. Limitations, Future Work

-There are neglected tools of evaluation, notably concerning the robustness.
- There is a lack of a rigorous human or application grounded evaluation to assess the general subjective quality of explanation.
- Future work can focus on specific assessment of several techniques by researchers who have the material means (e. g. human subjects) and on standardization for the process.

References

[1] B. Kim, M. Wattenberg, J. Gilmer, C. Cai, J. Wexler, F. Viegas, and R. Sayres. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). volume 80, pages 2668,2677, July 2018.
[3] F. Chen, A. H. Gandomi, J. Zhou, and A. Holzinger. Evaluating the quality of machine learning explanations: A survey on methods and metrics. Electronics, 2021. doi:doi.org/10.3390/electronics10050593.
[4] S. Das, N. Agarwal, D. Venugopal, F. T, Sheldon, and Sajjan Shiva. Taxonomy and survey of interpretable machine learning method. 2021. doi:10.1109/SSCI47803.2020.9308404.
[5] S. Muddamsetty, M. Jahromi, A. Ciontos, L. Fenoy, and T. Moeslund. Introducing and assessing the explainable ai (xai) method: Sidu. January 2021. URL: arxiv.org/abs/2101.10710, doi:arXiv:2101.10710.
[6] Ghorbani A., Wexler J., Zou J.Y., and Kim B. Towards automatic concept-based explanations. volume 32, pages 9273,9282, 2019. URL: papers.nips.cc/paper/9126-towards-automatic-concept-based-explanations.
[7] Finale Doshi-Velez and Been Kim. Towards a rigorous science of interpretable machine learning. 2017. doi: arXiv:1702.08608v2.
[8] Mengnan D., Ninghao L., and Xia H.. Techniques for interpretable machine learning. In Communications of the ACM, volume 63, page 68–77, 2020.
[9] Y. Hailemariam, A. Yazdinejad, R. M. Parizi, G. Sri- vastava, and A. Dehghantanha. An empirical evaluation of ai deep explainable tools. IEEE, 2020. doi: 10.1109/GCWkshps50303.2020.9367541.
[10] Ruth Fong and Andrea Vedaldi. Net2vec: Quantifying and explaining how concepts are encoded by filters in deep neural networks. pages 8730,8738, 2018. URL: doi.org/10.1109/CVPR.2018.00910.
[11] Rabold J., Schwalbe G., and Schmid U. Expressive explanations of dnns by combining concept analysis with ilp. pages 148–162, 2020.
[12] X. Li, Y. Shi, H. Li, W. Bai, C. Chen Cao, and L. Chen. Quantitative evaluations on saliency methods: An experimental study. 2020. URL: doi.org/10.1145/3447548.3467148
[13] Y. Lin, W. Lee, and Z. Berkay Celik. Evaluation of explainable artificial intelligence (xai) interpretability through neural backdoors. 2020. doi: arXiv:2009.10639v1.