Abstract

The field explainable artificial intelligence (XAI) has brought about an arsenal of methods to render Machine Learning (ML) predictions more interpretable. But how useful explanations provided by transparent ML methods are for humans remains difficult to assess. Here we investigate the quality of interpretable computer vision algorithms using techniques from psychophysics. In crowdsourced annotation tasks we study the impact of different interpretability approaches on annotation accuracy and task time. We compare these quality metrics with classical XAI, automated quality metrics. Our results demonstrate that psychophysical experiments allow for robust quality assessment of transparency in machine learning. Interestingly the quality metrics computed without humans in the loop did not provide a consistent ranking of interpretability methods nor were they representative for how useful an explanation was for humans. These findings highlight the potential of methods from classical psychophysics for modern machine learning applications. We hope that our results provide convincing arguments for evaluating interpretability in its natural habitat, human-ML interaction, if the goal is to obtain an authentic assessment of interpretability.

1. Introduction

In recent years complex machine learning (ML) models, many based on deep learning, have achieved surprising results in computer vision, natural language processing and many other domains. These models are difficult to interpret, which inspired many researchers to investigate ways to render ML models interpretable (Kim, 2015; Lipton, 2016; Doshi-Velez & Kim, 2017; Herman, 2017). There are many motivations for interpretable ML methods. Domain experts, data scientists or data engineers that control proper functioning of an ML pipeline need to be able to access the rules learned by a ML system in an intuitive manner in order to quickly spot the root causes of errors. More generally the main motivation for research on transparent ML is that intuitive human understanding of ML predictions can be a prerequisite for a healthy trust relationship between humans and assistive ML systems. In particular transparency is argued to prevent algorithm aversion as well as algorithmic bias. Algorithm aversion refers to cases when humans do not trust ML systems, even when they know that the model predictions are more accurate than those of a human (Dietvorst et al., 2015), algorithmic bias are cases of ethical or gender biases in ML predictions (Hajian et al., 2016). In the following we will also use the term algorithmic bias to refer to cases of too much trust into an ML prediction, for instance when a human interacting with assistive ML technology blindly follows its predictions. The usual narrative is that explanations of ML decisions can increase human trust in them (Sinha & Swearingen, 2002; Ribeiro et al., 2016).

A central problem with interpretability methods is that they are difficult to compare and evaluate. Most of the research compares methods using either proxy measures, that do not directly relate to interpretability by humans, as e.g. (Samek et al., 2017), or qualitative measures that render comparisons of results across studies difficult (Strumbelj & Kononenko, 2010). In this work we propose to use psychophysical methods to quantify and compare the quality of interpretability methods. We follow the ideas of (Schmidt & Biessmann, 2019) and base our approach on the assumption that the definition of interpretability is inherently tied to a human observer. Good interpretability methods should allow human observers to intuitively understand a ML prediction. Intuitive understanding of the rules learned by a ML system is reflected in how accurately and how fast humans make decisions when assisted with a transparent ML prediction. These two variables can be easily measured in psychophysical experiments that study the interaction between humans and ML systems.
The motivation for this work is twofold: For one this work aims at complementing previous work on measuring the quality of interpretability methods by establishing a quantitative measure of interpretability in the domain of computer vision that captures aspects of human cognition. Ultimately this will help practitioners to choose the right interpretability method for a given use case and researchers to devise novel objectives for better interpretability methods. Secondly the goal of this study is to validate to what extent existing approaches for measuring interpretability without humans in the loop reflect the interpretability metrics we measure in psychophysical experiments.

In the following we shortly highlight some of the related work and then describe an image annotation task, emotion recognition, as well as the ML model, the transparency approaches used and the experimental design for quantitatively evaluating interpretability with humans in the loop (HIL) and with no humans in the loop (NHIL). We compare the different interpretability approaches with respect to the HIL and NHIL metrics and analyze their relationship, in particular whether cheaper and more scalable machine based NHIL transparency metrics reflect the most relevant but more expensive HIL transparency metrics. We conclude with highlighting the implications of our results for practitioners that build systems with human-ML interaction or transparent ML.

2. Experiments

Annotation Task The annotation task was emotional expression classification on images. We used the extended Cohn-Kanade image data set (Lucey et al., 2010) which contains images for the classes, anger, contempt, disgust, fear, happiness, sadness, surprise. We reduced the data to a binary classification task in which annotators had to classify emotional expressions of anger and happiness. The annotators had the option of not providing an annotation in case they did not recognize the emotional expression.

Machine Learning Model We used a computer vision model from an open source python toolkit that achieves state of the art performance on emotional expression prediction from images (EmoPy).

Interpretability Methods We compared three different interpretability methods for the EmoPy computer vision model, a) the gradient of output w.r.t. the input image, b) Layerwise relevance propagation (lrp) (Lapuschkin et al., 2017) and c) Guided backpropagation (Springenberg et al., 2014). For all methods we used the implementation in the iNNvestigate! package (Alber et al., 2018). All methods were used with their default hyperparameters. For the LRP approach we used the sequential_preset_a variant provided in the package. The list of methods is not meant to be exhaustive, but we chose these methods based on recommendations of the iNNvestigate! package.

Quality of Explanations We compute two kinds of quality metric for each interpretability method, one with no humans in the loop (NHIL) and one approach based on psychophysical experiments with humans in the loop (HIL). In both settings we use all three interpretability methods to compute scores for each pixel in the image. We rank the pixels according to their score and mask a certain percentage of pixels. The percentages of shown pixels were ten logarithmically spaced values between 0 and 100 (Fechner, 1860). The masks showed 5, 6, 8, 11, 15, 19, 26, 34, 45, 60 percent of pixels of the image. Some example images for the emotional expression happiness are shown in Figure 1. These thresholds were based on initial experiments with different thresholds in which we determined the minimum number of pixels needed to detect the emotion and the number of pixels needed to enable most subjects to correctly classify the image.

Psychophysical human in the loop (HIL) metrics For each percentage of pixels shown we measured the annotation accuracy as well as reaction times in the emotional expression classification task.

No humans in the loop (NHIL) metrics Following standard perturbation approaches (Samek et al., 2017) we slightly modify the perturbations to match the conditions used in the psychophysics experiments. In particular we mask a certain percentage of pixels and feed the masked image to the convolutional neural network to obtain a prediction. To evaluate the interpretability quality we evaluate...
the predictive performance of the EmoPy model on masked images.

**User Interface and Experimental Design** We built the user interface using the open source library jsPsych (de Leeuw, 2015). In each trial of an experiment we show the same image with increasing percentages of pixels shown and ask the subjects to report whether the expression was happy, angry or not certain. As we used ten different mask sizes from 5% to 60% of all pixels in the image, subjects saw a series of ten images, an example is shown in Figure 1. At the last image, when 60% of the image was shown, all subjects correctly identified the emotional expression, see also Figure 3. The entire experiment was designed to be completed in about 10 minutes, based on a pilot experiment. For each label five images were shown, which resulted in \(5 \times 2 \times 3 \times 10 = 300\) images in total that were annotated by each subject. For each subject the order of the trials was randomized, so each subject has seen each interpretability method and source image in a random order, but the order of unmasking the image was always the same. In total 62 subjects participated in the experiment. The experiments were conducted on the crowdsourcing platform Amazon Mechanical Turk. We payed all subjects the minimum wage in the country of the research institution of the authors, 11 US$ per hour.

### 3. Results

In the following we compare the results of the psychophysical experiments with the results from the experiments without humans in the loop.

**Annotators’ uncertainty and interpretability** We investigated the impact of each interpretability approach on the uncertainty of annotators by counting how often they did not provide an annotation but just the *I don’t know* label. Our results show a clear effect of the interpretability method as shown in ???. The simplest gradient approach leads to the highest annotator uncertainty and least number of annotations up to mask sizes of 45% of all pixels. Guided BackProp (Springenberg et al., 2014) in contrast leads consistently to the lowest annotator uncertainty and the highest number of annotations. This finding highlights the importance of quantitatively comparing transparency approaches. The extent to which human users of ML can profit from transparency strongly depends on the quality of the explanation provided.

**Annotation accuracy distinguishes XAI methods** In Figure 3 we show the annotation accuracy, averaged across subjects, for increasing mask sizes and all three different transparency approaches. Explanations using the plain gradient approach consistently led to the lowest annotation accuracy. The layerwise relevance propagation approach (LRP) (Lapuschkin et al., 2017) yielded slightly better an-
Figure 5. XAI method quality ranked by annotation accuracy (ranker=human) or cross-entropy loss per image (ranker=AI). Left panel: automated XAI rankings are instable while human rankings show robust pattern. Guided BackProp explanations are consistently best. Middle and right panel: Ranks for each interpretability method computed on AI predictions and human annotators for each mask size. For most mask sizes human annotators’ accuracy was significantly higher for the Guided BackProp approach, there is no clear winner for the AI interpretability quality metric.

4. Conclusion

Reliable and quantitative measures for evaluating interpretability are however a fundamental prerequisite for designing and improving transparent ML systems (Adebayo et al., 2020). Often interpretability evaluations without humans in the loop (Samek et al., 2017; Adebayo et al., 2018) cannot be directly compared with humans in the loop metrics (Hase & Bansal, 2020). In this study we used psychophysical experiments to directly compare XAI quality metrics with and without humans in the loop.

Our findings demonstrate that not only do machine based interpretability metrics not allow to distinguish the three methods in terms of their interpretability. This is also reflected in the two right panels, which show the same data as in Figure 5(left), but split into all mask size conditions. The average ranks of the psychophysical experiments show the same clear pattern as the aggregate metrics, Guided BackProp is better than LRP which is in turn better than the plain Gradient explanation. In contrast it is difficult to single out the best interpretable explanation based on the machine based NHIL metrics; there is no significant difference between the methods for most thresholds.
Human Perception of Computer Vision Interpretability

References

Adebayo, J., Gilmer, J., Muelly, M., Goodfellow, I., Hardt, M., and Kim, B. Sanity checks for saliency maps. In Adv. Neural Inf. Process. Syst., volume 2018-Decem, pp. 9505–9515, 2018. URL https://goo.gl/hBmhDt.

Adebayo, J., Muelly, M., Liccardi, I., and Kim, B. Debugging tests for model explanations. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H. (eds.), Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/075b051ec3d22dac7b33f788da631fd4-Abstract.html.

Alber, M., Lapuschkin, S., Seegerer, P., Hägele, M., Schütz, K. T., Montavon, G., Samek, W., Müller, K.-R., Dähne, S., and Kindermans, P.-J. iNNvestigate neural networks! aug 2018. URL http://arxiv.org/abs/1808.04260.

de Leeuw, J. R. jsPsych: A JavaScript library for creating behavioral experiments in a Web browser. Behavior Research Methods, 47(1):1–12, mar 2015. ISSN 1554-3528. doi: 10.3758/s13428-014-0458-y. URL http://www.ncbi.nlm.nih.gov/pubmed/24683129http://link.springer.com/10.3758/s13428-014-0458-y.

Dietvorst, B. J., Simmons, J. P., and Massey, C. Algorithmic aversion: People erroneously avoid algorithms after seeing them err. Journal of Experimental Psychology: General, 144(1):114–126, feb 2015. ISSN 1939-2222. doi: 10.1037/xge0000033. URL http://www.ncbi.nlm.nih.gov/pubmed/25401381http://link.springer.com/10.1037/xge0000033.

Doshi-Velez, F. and Kim, B. Towards a rigorous science of interpretable machine learning. arXiv preprint arXiv:1702.08608, 2017.

EmoPy. https://github.com/thoughtworksarts/EmoPy.

Fechner, G. T. Elemente der Psychophysics. Breitkopf und Hartel, 1860.

Hajian, S., Bonchi, F., and Castillo, C. Algorithmic Bias. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '16, pp. 2125–2126, New York, New York, USA, 2016. ACM Press. ISBN 9781450342322. doi: 10.1145/2939672.2945386. URL http://dl.acm.org/citation.cfm?doid=2939672.2945386.

Hase, P. and Bansal, M. Evaluating explainable AI: which algorithmic explanations help users predict model behavior? In Jurafsky, D., Chai, J., Schluter, N., and Tetreault, J. R. (eds.), Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020, pp. 5540–5552. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.acl-main.491. URL https://doi.org/10.18653/v1/2020.acl-main.491.

Herman, B. The promise and peril of human evaluation for model interpretability. CoRR, abs/1711.07414, 2017.

Kim, B. Interactive and interpretable machine learning models for human machine collaboration. PhD thesis, Massachusetts Institute of Technology, 2015.

Lapuschkin, S., Binder, A., Müller, K.-R., and Samek, W. Understanding and Comparing Deep Neural Networks for Age and Gender Classification. aug 2017. URL http://arxiv.org/abs/1708.07689.

Lipton, Z. C. The mythos of model interpretability. arXiv preprint arXiv:1606.03490, 2016.

Lucey, P., Cohn, J. F., Kanade, T., Saragih, J., Ambadar, Z., and Matthews, I. The Extended Cohn-Kanade Dataset (CK+): A complete dataset for action unit and emotion-specific expression. In 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops, pp. 94–101. IEEE, jun 2010. ISBN 978-1-4244-7029-7. doi: 10.1109/CVPRW.2010.5543262. URL http://ieeexplore.ieee.org/document/5543262/.

Ribeiro, M. T., Singh, S., and Guestrin, C. ”why should I trust you?”: Explaining the predictions of any classifier. In SIGKDD, pp. 1135–1144, 2016.

Samek, W., Binder, A., Montavon, G., Lapuschkin, S., and Müller, K. Evaluating the visualization of what a deep neural network has learned. IEEE Trans. Neural Netw. Learning Syst., 28(11):2660–2673, 2017. doi: 10.1109/TNNLS.2016.2599820. URL https://doi.org/10.1109/TNNLS.2016.2599820.

Schmidt, P. and Biessmann, F. Quantifying interpretability and trust in machine learning systems. arXiv preprint arXiv:1901.08558, 2019.

Sinha, R. and Swearingen, K. The role of transparency in recommender systems. In CHI ’02 extended abstracts on Human factors in computing systems - CHI ’02, pp. 830, New York, New York, USA, 2002. ACM Press. ISBN 1581134541. doi: 10.1145/506443.506619. URL http://portal.acm.org/citation.cfm?doid=506443.506619.
Springenberg, J. T., Dosovitskiy, A., Brox, T., and Riedmiller, M. Striving for Simplicity: The All Convolutional Net. dec 2014. URL http://arxiv.org/abs/1412.6806.

Strumbelj, E. and Kononenko, I. An efficient explanation of individual classifications using game theory. Journal of Machine Learning Research, 11:1–18, 2010. doi: 10.1145/1756006.1756007. URL http://doi.acm.org/10.1145/1756006.1756007.