Manipulating Model Explanations: Why you shouldn't trust me

Verena Heusser
verena.heusser@student.kit.edu
Karlsruhe Institute of Technology (KIT)
Karlsruhe, Germany

ABSTRACT

This paper reviews state-of-the-art approaches to model explanations with a focus on those techniques that try to fool these methods.

We don't only want the model to be good. We want it to be safe and interpretable. As machine learning models enter critical areas in human lives such as the criminal justice system, medicine or financial systems, the inability for humans to understand these models is dangerous and problematic. Advances in the rising research area of explainable AI seems to be a remedy. However, there is not yet a consensus about the validity and robustness of explanations methods themselves. The main suggestion of this paper is to be cautious about results of explanation methods. Explanations can be fooled just as the underlying machine learning models. So. in the end the question must be posed whether inexplainable mdoels should be used at all if we need other models to explain these models but are not valid themselves ..

KEYWORDS

Interpretability, Neural networks, adversarial training

ACM Reference Format:
Verena Heusser. 2020. Manipulating Model Explanations: Why you shouldn't trust me. In *Explainability '20: ACM Symposium on Neural Model Explanations, December 16–20, 2020, Karlsruhe, DE.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

In recent years deep learning models have demonstrated superior performance in a number of tasks. While the performance is still rising and more task domains are accomplished, these models still remain black boxes often uninterpretable even by experts. In many domains, neural networks currently are the state-of-the-art solution. However, their superior performance comes at the cost of complexity, as the models often employ millions to even billions of parameters in order to achieve universal function approximation. This complexity means a drawback in interpretability as the decision making process of such a network cannot be followed by humans without the help of further tools. For instance, withing object recognition one would like to assume that the presence (or absence) of an object in the image causes a model to decide for a specific object category, closely akin to how humans base their decision process.

At present, many concerns regard the coherence of automated decisions to ethical standards. Regarding the expanding number of tasks that automated computer algorithms are used for nowadays, this concern is becoming even more important, as machine learning models are moving out of the lab into the real world. The application of algorithms for prediction of recidivism rates are already applied at court [2], and the filtering of job applicants, piloting of self-driving cars, diagnoses of diseases or automated food recognition [7] are already in use.

Thus, automated interpretation methods are required to make sense of the reasoning process and stability of such deep learning based models and to ensure that a model makes decisions without unfair or hidden biases. The research field approaching the explanation or validation of machine learning models is called explainable artificial intelligence (XAI). Not knowing about the biases of a network the vastly advancing technology of machine learning to be used in high-stakes and safety critical applications and prevent real-life deployment of such systems. Furthermore, the rise in machine learning model deployments also caused the development of adversarial attacks. These attacks attempt to fool a machine learning model by providing deceptive input. Fooling refers to the resulting malfunction of the model.

Not knowing about attacks and data arranged to exploit specific vulnerabilities has contributed to a relatively new research field of XAI comprising topics of (1) *model explanations*, (2) *adversarial attacks*, or manipulation methods and (3) the field of *adversarial manipulations of model explanations.* All of this is also known by the name of robust machine learning or even explainable artificial intelligence, as all subfields have the common goal to make models more robust and safe for deployment. (1) refers to the development of techniques that can be used to understand and explain the decision making process of a machine learning model or even the development of models that are inherently interpretable. (2) is the field of detecting vulnerabilites in models that cause models to be deceived by altered input. (3) is the main topic of this paper, i.e. how to fool explanation models in order to detect vulnerabilities and malfunctions in explanation methods.

Detecting such vulnerabilities in models is most crucial

Fragility limits how much wecan trust and learn from the interpretations

Most research on explainability focuses on the application of computer vision tasks. Most works in the field of XAI focus on image classification task, mostly because visualizations of a neural networks prediction can be easily verified by a human. The general purpose of image classification is to detect what objects are in an image. If a model works can be checked rather easily (if an

Explainability '20, December 16–20, 2020, Karlsruhe, DE

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

image contains a cat, the prediction of a neural network should be cat and not some other object category). However, how it works (*interpretability*), i.e. based on which features in the image the decision is made or which parameters in the model influence the prediction most, is an entirely different matter (*explainability*).

More importantly, while a big motivation for the development of robust and explainable systems is to overcome biases in models, datasets with direct implication of biases are seldomly used and by far not treated as benchmarking scenarios for explainability analyses.

The overview presented in this article examines the existing literature and contributions in the field of XAI focusing on methods to manipulate explanation methods. The critical literature analysis might serve as a motivation and step towards the biggest problems in XAI: How to make sure that interpretations of models are truly valid. This paper is structured as follows...

2 INTERPRETATION METHODS

There exists a variety of definitions in the vastly expanding research field of XAI, and the concept of *interpretability* still has no formal commonly used technical meaning [6]. To build on the common ground of existing research, this paper follows the terminoloqy of [1]. The authors make a distinction between the related but different concepts of *interpretability* and *explainibility*: The former refers to a passive characteristic of an underlying model to be understandable by humans, which is also defined as transparency. Contrary, the latter refers to an active characteristic of a model, referring to any procedure with the aim of specifying or detailing the inner workings of a model.

Interpretability refers to the the extent to which cause and effect can be observed in a system. and thus relates to the level at which a human can understand a machine learning algorithm.

Explainability refers to the extent to which a machine learning model can be explained in human terms. So to say, explaining a model means to literally explain what is happening.

High interpretability is desired as it can help to uncover biases in the model. Suppose a machine learning model is to be deployed for the task of income prediction based on features such as age, race, gender, education and hours of work per week. The performance of the system would mainly be evaluated in terms oft the predictive accuracy and the fairness of the system. The former can be evaluated with metrics, such as accuracy on a held-out test set. For the latter, interpretability methods might be applied to observe which input features are used by the model to predict the income. If the model uses sensitive features, such as sex and race as important features, it is systematically biased thus unfair.

The notion of these two concepts can be nicely explained by the comparison between deep neural networks and (shallow) decision trees.

Within a decision tree, there exists a distinct set of rules. The date will be split at each node into subsets and the leaf node hold the predicted outcome. All edges are connected by a logical 'AND'. Thus, cause and effect in a shallow decision tree are easy to observe and visualize, and thus the model is interpretable. However, a deep neural network, holding millions to billions of parameters in several layers, there are many neurons impacting the prediction to in order to directly attribute the impact of individual input features.

Explanation methods aim at making complex and inherently uninterpretable black box models interpretable by creating human readable visualizations. A frequently used type of explanation methods are feature attributions mapping each input feature to a numeric score. This score should quantify the importance of the feature relative to the model output. The resulting attribution map is then visualized as a heatmap projected onto the input sample to interpret the input attributes regarding which ones are the most helpful for forming the final prediction.

Definition 1: Interpretation Method
We consider a neural network $N : \mathbb{R}^d \rightarrow \mathbb{R}^k$. For the task of image classification, N classifies an input image $\mathbf{x} \in \mathbb{R}$ in k categories where the prediction $f_N(\mathbf{x}) = y \in \{1, ..., K\}$ is given by $y = argmax_i f_N(\mathbf{x})_i$.

Given the neural network N, it's input vector $\mathbf{x} \in \mathbb{R}^d$ and the the neural networks prediction for input $\mathbf{x}$, $f_N(\mathbf{x}) = y$, an interpretation method $\mathcal{I}$ determines why label y has been chosen by N for input $\mathbf{x}$. The interpretation is given by an output vector $\mathbf{x} \in \mathbb{R}^d$ where each entry h_i is mostly a numeric value describing the relevance of an input dimension x_i of $\mathbf{x}$ for $f_N(\mathbf{x})$. As $\mathbf{x}$ has the same dimensions as the input $\mathbf{x}$ it can be mapped to the input, overlaying $\mathbf{x}$ as a heatmap, where the color value represents the importance of feature x_i towards the prediction $f_N(\mathbf{x})$.

An example is given in TODO. Higher values, implying a stronger relative importance for making the prediction $f_N(\mathbf{x})$, is depicted in TODD color.

While all explanation methods try to obtain importance measures for the network prediction, they differ with respect to how these measures are obtained. [9] propose two major categories for explanation strategies, namely *black-box* explanation methods and *white-box* explanation methods. While black-box interpretations assume no knowledge about the underlying model, white-box methods only work by using the model parameters.

This terminology of discriminating between bb and wb models may not be confused with the nature of the underlying models: Models still remain of black-box nature even though a white-box method may contribute to making the decision making process of such a model more insightful.

The following section details the two categories and will give examples of the state-of-the-art interpretation methods within each group.

2.1 Black-box Methods

Black-box interpretations assume no knowledge about the model thus treating it as a black-box. The underlying model is approximated by learning it's behavior with an interpretable model, e.g. a linear model. As the model itself does not need to be known for using such a model-agnostic approach, these approaches can be used in scenarios where the model itself is not directly accessible. A black-box interpretation offers the big advantage to be applicable to any model.

LIME LIME perturbs the input and observes how the predictions change. In image classification, LIME creates a set of perturbed instances by dividing the input image into interpretable components (contiguous superpixels), and runs each perturbed instance through the model26 to get a probability

SHapley Additive exPlanations). (SHAP) This method calculates an additive feature importance score for each prediction with a set of desirable properties, such as local accuracy or consistency that its antecendants lacked.

2.2 White-box Methods

On the other side are white-box interpretations, where the model is known with all its parameters. Thus, the explanations can be directly computed by using the model instead of relying on an approximation of f_N as within the black-box models. As within these white-box models, tha model parameters can be used for calculating the interpretations, these methods are also named gradient or saliency map based methods.

Notations A white-box interpretation method, in the following named interpreter $\mathcal{I}$, generates a heatmap

$$h_c^{\mathcal{I}}(\boldsymbol{\omega}) = \mathcal{I}(\mathbf{x}, c; \boldsymbol{\omega})$$

for a neural network with parameters $\boldsymbol{\omega}$ and class c. The heatmap is a vector $h_c^{\mathcal{I}}(\boldsymbol{\omega}) \in \mathbb{R}^d$, where the j-th value $h_{c,j}^{\mathcal{I}}(\boldsymbol{\omega})$ represents the importance score of the $j-th$ input feature x_j for the final score of class c.

Layer-wise Relevance Propagation (LRP) relies on a Taylor series close to the prediction point rather than partial derivatives atthe prediction point itself

Gradient-weighted Class Activation Mappin (Grad-CAM) [8] To further improve the quality of the visualization, attribution methods suchas heatmaps, saliency maps or class activation methods (GradCAM[292]) are used. Grad-CAM uses the gradients of any target concept, flowing into the final convolutional layer to produce acoarse localization map, highlighting the important regions in the image for predicting the concept

SimpleGrad (SimpleG)

3 MANIPULATION METHODS

The focus of most works is on adversarial attacks against the prediction of a machine learning model. Contrary, the focus of this paper is on the attacks on the interpretation without changing the prediction.

A manipulation method refers to a method influencing an explanation method $\mathcal{I}$ to yield a wrong explanation. This influence on the explanation method is also called *fooling* or an *attack*.

Definition 2: Interpretation Manipulation Method
A manipulation method $\mathcal{F}$ is defined as a method for altering the output of an explanation method while leaving the model performance of the neural network N roughly unchanged. This must be the case, as the attack is targeted to fool the explanation method, and is essentially not targeted to fool the model itself. Fooling tho model would only disclose the vulnerability of the model but would not allow to gain insight into the stability of the fooling method. Hence, the main criteria for an explanation manipulation method must be fulfilled:

1. The models prediction stays approximately the same, i.e. $f_N(x_{adv}) \approx f_N(x)$ or $f_{N_{adv}}(x) \approx f_N(x)$
2. The explanation map $h(x_{adv})$ is significantly different to the explanation map resulting from non adversarial models or inputs $h(x)$, i.e. $h(x_{adv})$
3. In case the attack is in the input domain of the model, the perturbation of input samples must be imperceptible by humans. According to [3], the norm of the added perturbation δ to an input sample x thus must be small, i.e. $\|\delta x\| = \|x_{adx} - x\| \ll 1$

Other criteria to a manipulation method are the following:

Effectiveness. The manipulation scheme is inexpensive to conduct. Input manipulations are by definition inexpensive, as the perturbation can be applied to single input samples. Model manipulations are more expensive as they require tho model parameters to be adapted. However, an adversarial model can be obtained by fine-tuning the model with an adapted objective function. This fine-tuning also has the advantage that the model is adapted to include a systematic bias and can thus be applied to fool explanation methods without further adapting the model or input samples.

Transferability. The manipulation does not only fool one type of explanation method, but it's effect transfers to other types.

Most of the explanation methods outlined in sec. section 2 have been shown to be vulnerable to adversarial perturbations. Manipulation methods often show that there exist small feature changes resulting in a change of the explanation methods output while the output of the model itself does not change. Most approaches aim at providing a relevance measure of the input features.

3.1 Manipulation Levels

Adversarial Input Manipulations. The general approach is to perturb input data while observing the effect of this perturbation. As found in TODO, visually-imperceptible perturbations of an input image can make explanations worse for the same model and interpreter.

Adversarial Model Manipulations. Contrary to input manipulations, model manipulations do not operate on the input space but rather on the model parameter space itself. As first introduced by Heo et al. [5] in 2017, this line of research is comparably new. Adversarial model manipulations are obtained by fine-tuning the model on the same data but with an adapted objective function. [5] propose the adapted loss function of

$$\mathcal{L} = \mathcal{L}_{CE}(\mathcal{D}; \omega) + \lambda \cdot \mathcal{L}(\mathcal{D}; \omega; \omega_0)$$

where $\mathcal{L}_{CE}$ would be the regular cross-entropy classification loss.

The authors find that perturbed model parameters can also make explanations worse for the same input images and interpreters.

3.1.1 *Transferability of Manipulations.*

[5] find that fooling one explanation method with a fooling scheme transfers to other methods.

3.2 Manipulation Targets

A further categorization ef explanation methods can be made based on the target of the explanation.

Untargeted Manipulations. The majority of manipulations is untargeted, meaning that the applied perturbations are mostly random and not designed to change the prediction for a specific portion of an input sample.

Targeted Manipulations. On the contrary, targeted manipulations are designed to specificly alter the explanation of a distinct portion. This specific portion might be an object in the input image in the context of image classification. Manipulations on the level of the model are mostly targeted, as the explanation methods are being fooled by adapting the model parameters

3.3 Evaluation Criteria

As outlined in section section 3, there exists a plethora of explanation methods differing in the assumption about the model character and also in style how explanations are obtained. Thus, reliable evaluation methods are required allowing for a choice of an appropriate and robust explanation method. Evaluations of the quality of an explanation method can be separated into qualitative and quantitative evaluations.

[5] propose to measure the quality of an explanation method by their stability with respect to adversarial model manipulations.

Qualitative Evaluations. As explanations are attributed to input features, the resulting explanation values l can be easily mapped to the input vector x. Looking at these evaluations for specific samples is informative albeit not usable to obtain a general statistic about the explanation methods quality.

Quantitative Evaluations. Heo et al. [5] introduce the concept of the Fooling Success Rate (FSR) with the aim to introduce a measure for systematically measuring the quality of a fooling method.

4 CHARACTERIZATION OF ROBUSTNESS

sec:robustness)

This section introduces common evaluation strategies designed to test the robustness of either model or interpreter towards an applied attack.

TODO: - auch auf entlarvungsmethoden eingehen?

5 TRANSFERABILITY OF PERTURBATIONS

sec:transferability)

input perturbations do not propagate to the whole validation set. On the contrary, model manipulations are non-local perturbations, meaning that they do not merely perturb an input sample but rather effect all samples in the way that the model itself is changed.

6 EXPLAINING MANIPULATIONS

There is an abundance of examples and scenarios where model explanations fail. However, there are few paper specifically targeting why these manipulations work in the first place.

7 EXPERIMENTS

In this section, several experiments are evaluated that were conducted to replicate findings of other studies. Furthermore these approaches are extended to other domains and datasets.

7.1 Explanation Methods

7.2 Manipulation Methods

7.3 Models

7.4 Datasets

ImageNet German Loan Dataset
Recidivism Dataset
Adult Income Dataset The adult income dataset [4] contains .. samples.

8 DISCUSSION

9 CONCLUSION

Finally, it must be noted that the suitability of a method depends on its application domain.

Much critique has been applied to methods aiming at interpreting complex and potentially non-interpretable models. Some researchers argue it is not worthwhile to study non interpretable systems while dismissing that using inherently interpretable models in the first place might be the better approach.

Adversarial attacks show that machine learning systems are still fundamentally fragile: They may be successful in a number of tasks, but fail to adapt to ood scenarios, i.e. when being applied to unfamiliar territory.

The findings about manipulating interpretations do not suggest that interpretations are completely meaningless, just as adversarial attacks on predictions models do not imply that machine learning models are useless. However, they suggest that there still are fundamental flaws in the way neural networks operate und that much caution and supervision sould be applied if they are to be deployed in the real world.

REFERENCES

[1] Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador García, Sergio Gil-López, Daniel Molina, Richard Benjamins, et al. 2020. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion* 58 (2020), 82–115.

[2] Alexandra Chouldechova. 2017. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data* 5, 2 (2017), 153–163.

[3] Ann-Kathrin Dombrowski, Maximillian Alber, Christopher Anders, Marcel Ackermann, Klaus-Robert Müller, and Pan Kessel. 2019. Explanations can be manipulated and geometry is to blame. In *Advances in Neural Information Processing Systems*. 13589–13600.

[4] Dheeru Dua and Casey Graff. 2017. UCI Machine Learning Repository. http://archive.ics.uci.edu/ml

[5] Juyeon Heo, Sunghwan Joo, and Taesup Moon. 2019. Fooling Neural Network Interpretations via Adversarial Model Manipulation. *CoRR* abs/1902.02041 (2019). arXiv:1902.02041 http://arxiv.org/abs/1902.02041

[6] Zachary C Lipton. 2018. The mythos of model interpretability. *Queue* 16, 3 (2018), 31–57.

[7] Robin Ruede, Verena Heusser, Lukas Frank, Alina Roitberg, Monica Haurilet, and Rainer Stiefelhagen. 2020. Multi-Task Learning for Calorie Prediction on a Novel Large-Scale Recipe Dataset Enriched with Nutritional Information. *arXiv preprint arXiv:2011.01082* (2020).

[8] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*. 618–626.

[9] Alexander Warnecke, Daniel Arp, Christian Wressnegger, and Konrad Rieck. 2020. Evaluating explanation methods for deep learning in security. In *2020 IEEE European Symposium on Security and Privacy (EuroS&P)*. IEEE, 158–174.