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.

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, 3 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 being 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.

Thus, automated interpretation methods are required to make sense of the reasoning process 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 decision processes 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 triggered a whole new research field consisting of the topics of (1) *model explanations*, (2) *adversarial attacks*, or manipulation methods and (3) the field of *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.

While most of the approaches to explainability focus on the application to computer vision tasks, other domains are seldomly chosen. 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.

This article reviews the current state of the art research in the field of model explanations and model manipulations. This paper is structured as follows...

2 EXPLANATION METHODS

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: Explanation Method
We consider a neural network $N : \mathbb{R}^d \rightarrow \mathbb{R}^k$. In case the task is image classification N classifies an input image $x \in \mathbb{R}$ in k categories where the prediction $f_N(x) = y$ is given by $k = argmax_i f_N(x)_i$.

Given the neural network N, it's input vector $x = (x_1, ..., x_d)$ and the the neural networks prediction to x $f_N(x) = y$, an explanation method $\mathcal{I}$ determines why label y has been chosen by N for input x. The explanation is given by an output vector $l = (l_1, ..., l_d)$ where each entry l_i is most often a numeric value describing the relevance of an input dimension x_i of x for $f_N(x)$. As l has the same

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

dimensions as the input x it can be mapped to the input, overlaying x as a heatmap where the color value represents the importance of feature x_i towards the prediction $f_N(x)$.

An example is given in TODO. Higher values, implying a stronger relative importance for making the prediction $f_N(x)$, 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. [3] propose two major categories for explanation strategies.

Black-box Explanations. Black-box explanations assume no knowledge about the neural network model thus treating it as a black-box. The model is approximated which makes these explanation methods applicable in scenarios where the model parameters are not directly accessible.

White-box Explanations. On the other side are white-box explanations, 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.

In this paper, I will focus exclusively on white-box methods.

Gradient Based Explanation Methods

3 MANIPULATION METHODS

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: Explanation 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 output of the network N stays approximately the same, i.e. $f_N(x_{adv}) \approx f_N(x)$
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 [1], 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. [2] 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. [2] 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. [2] 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.

[2] 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. [2] 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

7.4.1 *ImageNet.*

7.4.2 *Recidivism Dataset.*

7.4.3 *German dataset of..*

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.

REFERENCES

[1] 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.

[2] 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

[3] 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.