Why you shouldn't trust me: A survey on Adversarial Model Interpretation Manipulations.

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

Abstract

This paper provides a review on the subfield of manipulating interpretable machine learning, with the aim to provide insight into concepts, existing research and future directions. This paper reviews state-of-the-art approaches to model explanations with a focus on those techniques that try to fool these methods. Based on the findings in the reviewed studies, we conclude that knowledge of the defectiveness of neural networks and their highly sought explanations should be taken with caution. Therefore, the main purpose of this review is to raise awareness that explanations, while helpful, can can be fundamentally flawed.

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, Adversarial Machine Learning, Adversarial Model Interpretation Manipulations

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

1 Introduction

In recent years, deep learning models have demonstrated superior performance on a variety of tasks [? ? ?]. While performance is still increasing and more tasks are being handled, their performance comes at the cost of complexity: models often use millions to billions of parameters to achieve universal function approximation. This complexity means that they remain black boxes that cannot be interpreted even by experts.

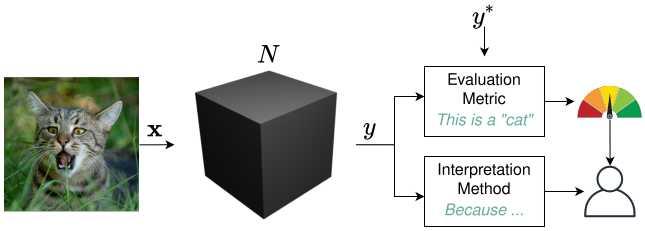


Figure 1: Prediction pipeline using a machine learning model, depicted as blach box. Typically, evaluation metrics require the prediction y and the ground truth label y^* allowing for the assessmant of the model's accuracy. Additionally answering the question *why*, i.e. making the model interpretable for a human, requires additional methods.

Such a black box is able to predict well for unseen yet similar data, answering the question of *what* is the most likely label for an input sample. However, most models will provide no answer to *why* or *how* the model chose this label for the instance and which features of the instance were crucial for making this prediction.

For example, if a machine learning model is tasked with classifying images of cats (as in **??**), one would like to assume that the presence (or absence) of a cat in an image is indicative (contraindicative) of the classification of the image into the "cat" ("not cat") category.

Suppose a deep neural network predicts the risk for cancer from a mammogram, which is an image of breast tissue. A doctor would only use the algorithm if there is a way to validate that (1) tha algorithm is accurate (which can be measured in terms of the predictive accuracy), and (2) if the model is also using the correct indications in the data for predicting the risk of cancer. (1) is the standard approach for validating the performance of machine learning models, but in this example, one can clearly see why predictive accuracy might not be enough in many areas. For approaching (2), i.e. uncovering *why* a model predicts a low / high risk of cancer, the research field of explainable artificial intelligence (XAI) offers a growing number of methods. Some research even suggests to allow 'peeking inside the black box' of deep learning models [?].

Automated algorithms are already in use in critical areas, such as medicine, chemistry, the criminal justice system, the financial sector or the piloting of self-driving cars [? ? ?]. Thus, as machine learning models are moving out of the lab into the real-world, the inability

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

of humans to understand these (black box) models seems even more problematic. Not knowing how a model makes predictions, and not being able to detect systematic biases in the model, prevents the vastly advancing technology of machine learning from being used in highly sensitive and safety-critical applications.

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.

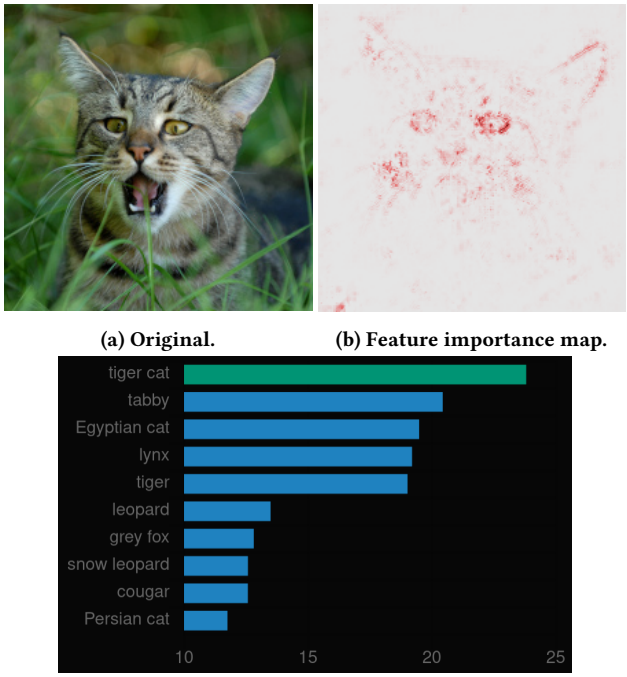
Not knowing about attacks and data arranged to exploit specific vulnerabilities has contributed to the field of XAI comprising topics of (1) *model interpretations*, (2) *adversarial attacks*, or manipulation methods and (3) the field of *adversarial manipulations of model interpretations*. All these 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 vulnerabilities in models that cause models to be deceived by altered input. (3) is the main topic of this paper, i.e. how to fool interpretation methods in order to detect vulnerabilities and malfunctions in interpretation methods. Ideally, an interpretation, or explanation method should indicate which pixels in the original image contribute to the prediction and also to what extent each pixel contributes. The extent to which each pixel contributes to a prediction is often called the *importance* score of a feature. **??** shows the importance map built by an explanation method for the deep neural network.

?? shows such a feature importance map produced by the interpretation method LRP [**?**] applied to the neural network model Inception-v3 [**?**]. The input image is from the ImageNet dataset [**?**]. As can be seen, the output of the interpretation method is projected onto the original image for better readability by humans. This importance map suggests that specific portions of the original cat image are important for the neural network make the high-confidence prediction (see..) of the category 'tiger cat'.

While interpretation methods are already used for analysis of computer vision systems [**? ? ?**], text and sequence analysis [**? ?**], and deep learning in security [**?**], there is still a lack in the the understanding of model interpretation methods.

In particular, it is unclear how the variety of proposed model interpretation methods are related and what common concepts can be used to evaluate and compare them. Many works are dedicated to establish a formal definition of what it means to be interpretable, and how to select, evaluate and compare methods for generating explanations of machine learning models. [**? ?**].

Most works in the field of XAI focus on image classification tasks, 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 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*).



(a) Original. **(b) Feature importance map.**

(c) Predictive accuracy.

Figure 2: Visualization of the feature importance map produced py the LRP interpreter applied to the image of a cat and an image classification model.

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 examines a topic at the intersection of explainable and adversarial machine learning research. 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.

?? introduces common interpretation methods for machine learning models, and offers a taxonomy of the variety of techniques and a brief outline of popular interpretation methods. In **??**, the main topic of this paper, namely manipulation methods for deceiving interpretive techniques, is outlined. A taxonomy of methods is proposed and possible evaluation criteria are listed.

?? provides a detailed review of important studies in the field of manipulation methods. The implacations of the current state-of-the-art in axplainable AI are discussed in section **??**.

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 [**?**]. To build on the common

ground of existing research, this work follows the terminology of Lipton et al. [**?**] and Arrieta et al. [**?**].

Broadly, interpretabiity focuses on *how* and *why* a machine learning model makes predictions. Simply put, interpretability is focused on getting some notion of an explanation for the decisions made by a machine learning model.

2.1 Why is Interpretability Important?

First, interpretability is helpful as it can help to build trust in a machine learning application.

The advantage of post-hoc interpretations is that they do not interfere with the training process of the model, and thus do not change the model. As the name says, post-hoc means that the techniques can be easily applied to an already trained model without much further computational overhead.

High interpretability is desired as it can help to 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 and thus unfair.

2.2 Terminology

The authors of [**?**] make a distinction between the related but different concepts of *interpretability* and *explainibility*. Lipton [**?**] further breaks down interpretability into *transparency* and *post-hoc* interpretability. The notion of *explainibility* from [**?**] can be related to Liptons's *transparency*, while Lipton's *post-hoc* interpretability is essentially *interpretability* as defined by Arrieta et al. [**?**].

Post-hoc Interpretability refers to the the extent to which cause and effect can be observed in a model, which can be translated to uncovering *why* a model made prediction y to an input $\mathbf{x}$, or how input and output ralate. Consider the example of image classification from **??**. Here, interpretability would mean that if a cat is present in the image (the cause), the model classifies it to the category 'cat' (the effect). Now imagine we find that the model takes the green meadow in the image as evidence to predict 'cat', and not the cat itself. This would imply a lack of interpretability, as the model learns to assign features to the concept 'cat' other than those related to 'cat' in the correct sense. This toy example emphasizes a common problem in image classification: [**?**] observe the over-reliance of models on image background, rather than on objects in the foreground.

Explainibility or Transparency on the other hand spans methods to uncover *how* a model makes predictions, meaning to observe the inner workings of a model and thus to literally explain what is happening in terms of understanding of the mechanisms by which a model works. Thus, tansparency refers to the model's inherent properties that can be known before the training process and that are helpful to understand the model.

While both concepts seem to be important for the general objective of explainable artificial intelligence, this paper focuses on post-hoc interpretability. There are essentially two ways to achieve interpretability: (1) to use inherently interpretable models or (2) to post-process a model in a fashion that allows to yield insights. The former is known as the development of *surrogate* models and more generally described as *model-agnostic* methods. Option (2) is known as *model-transparent* methods.

Local and Global Methods. A further categorization can be made based on the scope of interpretations: *Local* methods aim at providing interpretations that are true for a single data points and its neighbors. *Global* methods aim at gaining interpretations that are valid for most data points in a class [**? ? ?**]. The interpretation methods discussed within this paper mostly fall into the class of local explanation methods [**? ? ?**].

Feature Attribution Methods and Sample Attribution Methods. Interpretation 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. This allows humans to interpret which input attributes are the most helpful for making tha final prediction. Sample attribution methods on the other hand interpret the model performance in terms of the importance of training examples from the dataset.

We propose the following formal definition for interpretation method:

Definition 1: Interpretation Method.
We consider a neural network $N : \mathbb{R}^d \rightarrow \mathbb{R}^k$. For the an arbitrary classification task, N classifies an input sample $\mathbf{x} \in \mathbb{R}$ 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 $\mathbf{h_k} \in \mathbb{R}^d$ for a class k where each entry h_i is mostly a numeric value describing the relevance for the i-th input feature x_i of $\mathbf{x}$ for the final score $f_N(\mathbf{x})$.

As $\mathbf{h}$ 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 **??**. Higher values, implying a stronger relative importance for making the prediction $f_N(\mathbf{x})$, are depicted in dark red.

While all explanation methods try to obtain importance measures for the network prediction, they differ with respect to how these measures are obtained. [**?**] propose two major categories for interpretation strategies, namely *black-box*, in the following named *model-agnostic* methods and *white-box*, or interpretation 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 black-box and white-box methods 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.

2.3 Model-agnostic methods.

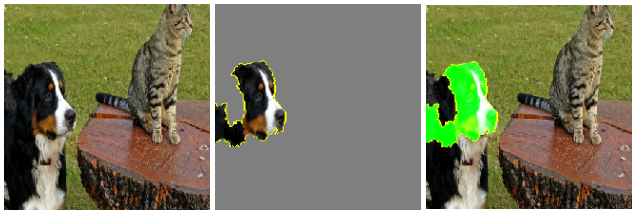
Model-agnostic interpretations assume no knowledge about the model thus treating it as a black box. The underlying model is approximated by learning it's behavior with an interpretable model, e.g. a linear model. The interpretable model is also dubbed the 'surrogate' model. The common approach for learning the surrogate is to approximate the relationship between the input samples and the corresponding prediction by tho model. As the model itself does not need to be known, these approaches can be used in scenarios where the model itself is not directly accessible. Model-agnostic interpretations are fairly popular and are used in a wide range of applications, ranging from finance and law to medicine and chemistry [**?** **?**].

A black-box interpretation offers the great advantage of being applicable to any model and offers simplicity because the interpretation is embedded in the model. However, this option of gaining interpretability might be costly for users that already have a high performing model. For this reason, growing need for methods exists that can be applied without retraining or modifying the underlying model. We will briefly describe two common model-agnostic approaches. Refer to the original papers for details.

Local Interpretable Model-agnostic Explanations (LIME). LIME [**?**] perturbs the input and observes how the predictions of a black box model change. For the task of image classification, LIME creates a set of perturbed instances by dividing the input image into interpretable components, which are technically contiguous superpixels, and runs each perturbed instance through the model to get the probability for how much the change in each superpixel influences the whole model prediction.

?? shows an example of LIME applied to the neural network model Inception-v3 [**?**]. Input image from the ImageNet dataset [**?**]. The superpixel in dog instances place is highlighted in green, which correctly indicates that this superpixel has a high influence on the prediction of the images correct class ('bernese mountain dog'). LIME also correctly indicates that the superpixel in the cat's place does not indicate the correct class label.

SHapley Additive exPlanations (SHAP). SHAP [**?**] builds on Shapley analysis, which is essentially about judging the importence of attibutes. The model is trained on a number of subsets of all available features, and the feature importance scores are calculated by evaluating the effects that the omissions of the specific features have on the model prediction. An example is shown in **??**. Image regions highlighted in green are found to be important for predicting the correct label.



(a) Original **(b) Mask.** **(c) Saliency Map.**

Figure 3: Visualization of the output of the LIME interpreter applied to an image and image classification model.

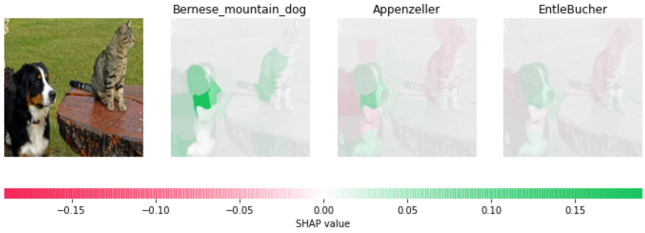


Figure 4: Visualization of the output of the SHAP interpreter applied to applied to an image and image classification model.

2.4 Model-transparent methods.

The other group of interpretations are model-transparent, or white-box methods, where the underlying model is known with all its parameters. Thus, the interpretation can be directly computed by using the model instead of relying on an approximation of f_N as within the black-box methods. These methods typically rely on the relationship between an input sample, the underlying model's prediction and the associated activations of the models hidden layers. Methods within this group are for example propagation-based and gradient-based approaches. The former propagate back the model's prediction back through the model. The latter make use of the information provided by the gradients of the loss function, which contain sensitive information about the prediction and tha features. A few example methods within this group are listed below.

Layer-wise Relevance Propagation (LRP). While many approaches in the group of model-transparent interpretations are designed only for image classification, or convolutional neural networks, this method [**?**] is an exception. This method [**?**] propagates relevance values backwards through the network. It relies on a Taylor series close to the prediction point rather than partial derivatives at the prediction point itself. An example of LRP applied to the cat image example is in fig:lrp$_c at_l rp$.

Gradient-weighted Class Activation Mapping (Grad-CAM) [**?**] 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)

Smooth Grad

Please see the original publications for more detailed information.

3 Manipulation Methods

As outlined in **??**, there are a variety of explanation methods readily available as frameworks and open source implementations. However, there is still little analysis on the robustness and reliability of such methods. While it is already common practice to test machine learning models against adversarial attacks in a number of domains [**? ?**], the same is not yet the standard for interpretation methods. We argue that interpretation techniques should not be used in critical applications without basic testing of interpretation techniques against adversarial settings.

3.1 Adversarial Setting

Adversarial Attacks on Models. Adversarial examples, as first introduced by [**?**], are clever manipulations of an input by an adversary which aims at causing misclassification and fooling of applications. They are mostly used to fool or attack machine learning models. We formally define adversarial attacks by the following properties:

Definition 2: Model Manipulation Method

1. *Impercpetibility of Perturbation*: The adversal example is similar to original data, i.e. the norm of the added perturbation δ to an input sample $\mathbf{x}$ thus must be small, i.e.
$$\|\mathbf{x}+\delta\| = \|(\mathbf{x}+\delta)-\mathbf{x}\|_{\text{inf}} \leq \epsilon$$
2. *Prediction dissimilarity*: The prediction of the model is significantly different to the prediction on the non-adversarial example:
$$f_N(\mathbf{x}+\delta) \approx f_N(\mathbf{x})$$

Note, that within adversarial foolings of models, the perturbation is mostly applied to the input data, and not to the model itself.

Evicendce from many studios shows that deep learning models can be easily tricked by adversarial exapmles. Albeit there are not yet as many studies, there also exists evidence that many interpratation methods are aslo fragile with respect to small changes to input data [**? ? ?**]. This fooling of interpretation methods is outlined below.

Adversarial Attacks on Model Interpreters. Contrary to adversarial attacks on machine learning models, the focus of this paper is on the attacks on interpretation techniques without changing the prediction of the model. An adversarial attack on a model interpretation is in the following also called a *manipulation* method. The goal is to apply perturbations to either an input sample or the model to change the output of an interpretation technique while leaving the model prediction unchanged. The last condition is important because adversarial interpreter manipulations aim to fool the interpretation method and essentially not the model itself. Fooling the model would only disclose the vulnerability of the model but would not allow to gain insight into the stability of the interpretation method. Again, the problem can be formally defined as:

Definition 3: Interpretation Manipulation Method.

A manipulation method $\mathcal{F}$ is defined as a method for altering the output of an explanation method $\mathcal{I}$ while leaving the model performance of the neural network N roughly unchanged. As manipulations can be applied on the input or the model level (see **??**), $\mathbf{x}+\delta$ denotes a perturbed input sample regarding the input level manipulation, while $N+\delta$ denotes a model with altered parameters, referring to the model level manipulation.

A manipulation method is successful in fooling an interpreter, if the following properties hold:

1. *Prediction similarity*: The model prediction stays approximately the same, i.e.
$$f_N(\mathbf{x}+\delta) \approx f_N(\mathbf{x}), \text{ or } f_{N+\delta}(\mathbf{x}) \approx f_N(\mathbf{x})$$
2. *Interpretation dissimilarity*: The explanation map $h(\mathbf{x}+\delta)$ is significantly different to the explanation map resulting from non adversarial models or inputs $h(\mathbf{x})$, i.e. $h(\mathbf{x}+\delta)$ or
$$\arg\max_{\delta} = \mathcal{D}(\mathcal{I}(x_i, \boldsymbol{\omega}), \mathcal{I}(x_i+\delta, \boldsymbol{\omega})$$
3. *Impercpetibility of Perturbation*: In case the attack is in the input domain of the model, the perturbation of input samples must be imperceptible by humans. According to [**?**], the norm of the added perturbation δ to an input sample $\mathbf{x}$ thus must be small, i.e.
$$\|\mathbf{x}+\delta\| = \|(\mathbf{x}+\delta)-\mathbf{x}\|_{\text{inf}} \leq \epsilon$$

These measures are to be seen as comparison between a baseline model N and a model that is applied an the adversarial setting (i.e. either N is not changed but applied to adversarially altered data $x+\delta$, or N is adversarially trained thus becoming $N+\delta$).

[**?**] extend these properties to include the so called *model similarity*. This measure extends the *prediction accuracy* to span the accuracy difference in between the baseline model and the new model, and also the mismatch of data points where the predictions of both models differ.

3.2 Taxonomy of Manipulation Methods

3.2.1 *Manipulation Levels*

TODO

Adversarial Input Manipulation. The general approach is to perturb or alter input data while observing the effect of this perturbation on the model prediction. As found in TODO, visually-imperceptible perturbations of an input image can make explanations worse for the same model and interpreter. This concept is visualized in **??**.

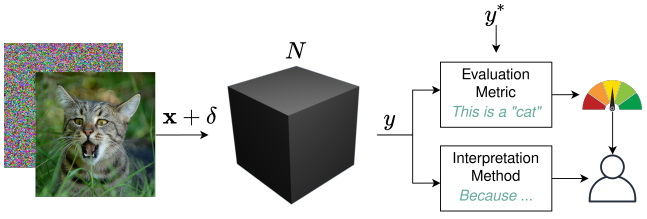


Figure 5: Depiction of an adversarial input manipulation. The model is fine-tuned with altered input samples, which are indicated by $\mathbf{x}+\delta$.

Adversarial Model Manipulation. 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. [**?**] in 2019, 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. [**?**] propose the adapted loss function for the task of image classification 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 standard cross-entropy classification loss. Adversarial model manipulation is depicted in **??** where the altered model is indicated by $N + \delta$.

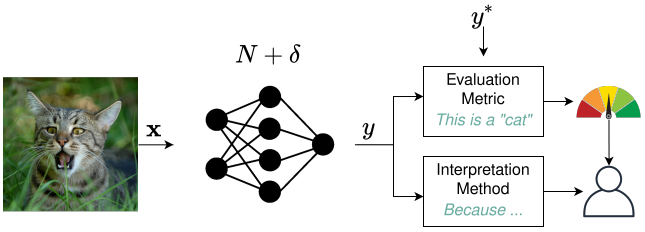


Figure 6: Depiction of an adversarial model manipulation. The model is fine-tuned with the same distribution of input data and a fooling loss, thus yielding the biased model $N+\delta$.

3.2.2 Manipulation Targets

In addition to the categorization of manipulation methods based on the manipulation level, the methods can further be categorized based on the target of their perturbation. The first possibility is *untargeted* perturbation, the second is *targeted* perturbation. Both these styles can be applied on either model and input level.

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 distinct features of an input instance. Such a specific feature might be an object in the input image in the context of image classification. For instance [**?**] introduce a fooling scheme in which the interpretations of the classes elephant and school bus are swapped. 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

Basides the necessary properties of a successful interpretation manipulation method, other evaluation criteria are importent to access the successfulness of a fooling method and to enable the comparison between different fooling methods. These criteria are infermally defined in 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 after the fine-tuning step. Furthermore, this systematic bias is hidden in the model, and is hard to uncover. Input manipulations can only fool the model when the inputs are always manipulated.

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

Generalisation. Generalization of an attack refers to the transfer of fooling to other test samples. This is noteworthy since a manipulation method might only perturb the decision boundary locally around the training points, i.e. only influencing trainang instances and their neighbors. However, it is desired that the explanations of unseen samples are affected as well. Furthermore, not only unseen samples intarpretations, but also samples that are far away in the feature space, should be affected,

TODO

Most of the explanation methods outlined in sec. **??** 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.

Creation of a fooled model by fine-tuning the model with a fooling loss.

As outlined in section **??**, there exists a plethora of interpretations methods differing in the assumption about the model character and also with respect to how interpretations are obtained. Thus, reliable evaluation methods are required allowing for a choice of an appropriate and robust interpretation method. Ultimately, the accordance to these evaluations should naturally allow for chosing an appropriate and robust intarpretation method. Evaluations of the quality of an explanation method are separated into qualitative and quantitative evaluations.

Qualitative Evaluation. Inspection and ramdom sampling are commonly used techniques to obtain an intuition about the effect of manipulations. As interpretations are attributed to input features, the resulting relevance values l can be easily mapped to the input vector $\mathbf{x}$. Visual inspection of these evaluations for specific samples is informative, but does not allow for general statistics and validation of manipulation effects. Thus, quantitative evaluations are required.

Quantitative Evaluation. As the goal of interpreter manipulations is to fool an interpreter, thus altering the output of in interpreter, it is straightforward to compare interpretations and data samples before and after perturbation [**?**]. Common metrics for quantitatively measuring similarities are the following:

- **Spearman's rank order correlation.** As interpretation methods rank the features based on their importance, the rank correlation [?] is a natural meusure for comparing interpretations.
- **Intersection of the top-k features.** For some tasks, only the top-k features are relevant, such that a comparison between these top-k features is insightful.
- **Fooling Success Rate (FSR).** [?] introduce the concept of the Fooling Success Rate (FSR) with the aim to create a measure for systematically measuring the robustness of an interpretation method to adversarial model manipulations. The FSR measures on how many samples from a dataset an interpretation method $\mathcal{I}$ is fooled by a model manipulation. The higher the FSR, the more often an inetpretation method is fooled.
- **Structural Similarity Index (SSIM).** SSIM values are relative similarity measure in the range $[0, 1]$, where larger values indicate higher similarity.
- **Pearson Correlation Coefficien (PCC).** PCC is also a relative similarity measure returning values in the range $[0, 1]$. Larger values indicate higher similarity.
- **Mean Squared Error (MSE).** As an absolute error measure, values close to zero indicate high similarity.
- **AOPC** TODO

Note that these are only examples without demand for completeness. For further information see [?]. Normalizing these measures to yield values in $[0, 1]$ with a sum of one is good practice.

4 Interpreter Manipulation Method Examples

After introducing the terminology of model interpreters and manipulation methods in the previous chapters, this chapter gives detailed information about recent manipulation methods. This section also provides insight into major findings in the field of manipulating model interpretations. First, input level manipulations are discussed, followed by model level manipulations.

4.1 Input Level Manipulations

[?] propose a framework for fooling the model-agnostic interpreters LIME and SHAP. Their method successfully hides the biases of models trained on adversarial examples by... The authors take a statistical approach model-agnostic interpretation methods. They examined the data produced by LIME and SHAP and showed that the perturbed samples are out-of-distrubution samples compared to the distribution of the regular trainging data.

?? shows that the data samples generated by LIME are distributed differently than the data original, non-perturbed data samples.

4.2 Model Level Manipulations

[?] were the first to introduce adversarial model manipulations for fooling interpretation methods. Their research quesiton is if state-of-the-art interpretation methods be fooled with adversarial model manipulations. For this, adapted the fine-tuning of image-classification models by using an altered loss function. After fine-tuning the model with the regular classification loss combined with an adversarial loss, they investigated if the interpreter results

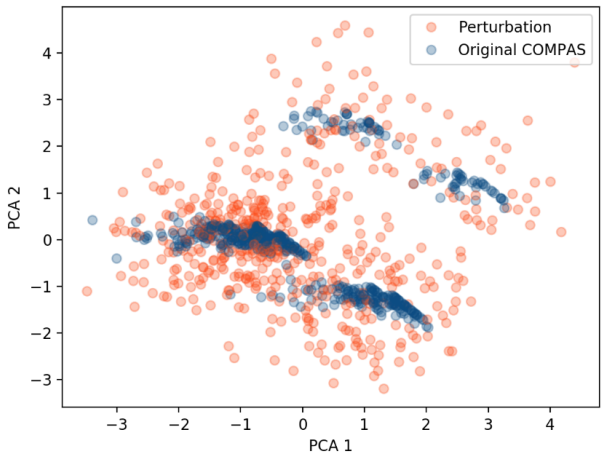


Figure 7: Principal Component Analysis (PCA) on the original data (orange) of the COMPAS dataset and the samples perturbed by LIME (blue). Note that the differences are obvious even in the space reduced to two dimensions.

change as a function of model parameters. Two types of fooling methods were introduced, namely *active* and *passile* fooling.

[?] examine the relation of interpretation methods and the concept of fairness. They propose to learn a modified model with concealed unfairness. This is done by fine-tuning a classification model with a loss function extended by an explanation loss.

Their approach differs methodologically to [?] as follows: [?] adapt the standard cross entropy loss function by taking the gradient of the correct label ellement from the logits layer, while [?] use the gradient of the cross-entopy loss. Taking the gradient of the cross-entropy loss conveys more information about other classes, which may contribute to an improved generalization across different interpretation methods and first of all across different test samples.

While their approach takes the gradient of the onecorrect label element from the logits layer just before thesoftmax output, we take the gradient of the cross-entropyloss.

They define adversarial models that focus only on sensitive features which are not informative for the ground truth decision.

5 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.

6 Uncovering Manipulations

7 Sanity Checks for Interpreters

Checking the robustness, scope and hence the quality of model interpreters has become an indispensable step in explainable machine learning.

[?] propose a number of randomization tests for saliency-map based interpreters. The authors find that most methods fail their tests, and warn of the danger of visual assessment.

Additionally, in order to account for adversarial model manipulations, Heo et al. [**?**] propose to expand the criteria for checking the robustness of interpreters further.

8 Benchmarking Interpretations

Evaluating explanation and interpretation methods is difficult as ground truth is mostly lacking. In most applications, it is not known which input features are most important.

9 Discussion

This paper summarizes the current approaches to manipulating model interpretation methods. On the one hand, the findings suggest that our models aro not fully aligned with how human information processing works. If machine learning models would decide by the criteria we humans employ for tasks such as image classification, there would be no fooling of interpretation models by input or model manipulations. On the other hand, it was shown that advances in machine learning models has led to models that rely too much on the data they are trained on, thus showing a high susceptibility to ood properties or properties that are highly correlated with labels in the dataset but are not distinctive in the real world (such as image backgrounds). Models and interpreters can still be misled in a large and systematic manner.

However, we hope that the vastly expanding and progressing field of XAI will help to move towards more robust, reliable ond human-aligned machine learning models.

While there exists a number of review papers on XAI and it's various subfields, this report is to the best of our knowledge the first one to comprehensively review manipulation methods for interpreters.

A growing number of studies gives evidence for model interpretation methods. Among these are the studies outlined in **??**. Other studies also raise concerns about if standard deep learning practices are valid, such as the work on fooling the attention mechanism.

We believe that identifying risks and adversaries helps to open up research on more robust interpretation methods.

Interpretation methods can be categorized based on if they maintain local consistency among explanations (i.e. finding an explanation that is true for single data samples and their neighbors) or based on if they try to find global explanations, being true for all samples of a class. As there exist model manipulations methods, that structurally alter the models by adapting tre loss function, this line of global model fooling though being approached is still in its infancy.

9.1 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 wolrd .

do notsuggest that interpretations are meaningless, just as adversarial attacks on predictions do not imply that neural net-works are useless.

This paper follows the footseps of [**?**], trying to caution against blindly putting faith into post-hoc explanation methods. Moreover, we propose that checking the robustness of interpretation methods not only with respect to adversarial input manipulations but also with respect to adversarial model manipulation should be an proof of concept.

9.2 Future Work

We see several possible future directions of future work. Firstly, for approaching the discrepency of in-lab and real-life applications of, more focus might be laid in the development better performance metrics for both measuring the performance of machine learning models as well as their interpreters. More specifically, it might be fruitful to further investigate the correlation between ood samples and the performance of an interpretation method. So far, most of these findings are limited to specific experimental settings.

There is also no work proposing a benchmarking for ...

What is currently sparse is the comparison between different interpretation techniques and the relation of interpretation fragility to the model class, interpretation method and the task type and dataset structure.