Explainability for neural networks

Rieger, Laura

Publication date:
2020

Document Version
Publisher's PDF, also known as Version of record

Ph.D. Thesis
Doctor of Philosophy

DTU Compute
Department of Applied Mathematics and Computer Science

Explainability for neural networks

Laura Rieger

Kongens Lyngby 2020

DTU Compute
Department of Applied Mathematics and Computer Science
Technical University of Denmark

Richard Petersen Plads
Building 324
2800 Kongens Lyngby, Denmark
compute@compute.dtu.dk
www.compute.dtu.dk

English Summary

Neural networks are nowadays used for more and more applications, often without the end-user even being aware of it. However, because of the number of parameters and amount of training data, they effectively become a black box. We have no way of knowing for what reasons a decision was output. When we use neural networks to make high-impact decisions, this becomes dangerous.

In this thesis, we address this need for explainability with a focus on application. In particular, we note the gap between the amount of research into interpretability and its application in practice and contribute to close this gap. To this end, we contribute a chapter with a historic perspective on prior approaches towards explainability.

Furthermore, we introduce a novel approach to evaluate explanation methods for a given use case, enabling practitioners to identify the best method for their task. Using this approach, we confirm that there are inherent differences in the interpretability of several popular network architectures. This insight implies that it may be favorable to use a less accurate architecture that is more interpretable for a task where the neural network classification is secondary.

We show that an ensemble of explanation methods not only better explains a neural network but is also much more robust to adversarial manipulation. Especially when explanations are a factor in the final decision or used to determine whether a given decision was aligned with ethical requirements, making them robust to malicious manipulation is vital.

Lastly, we contribute a method to infuse prior knowledge into a neural network via explanations. The end-to-end nature of neural networks prevents the use of domain knowledge in the algorithm and requires unbiased, independent, and identically distributed data for training. Our method gives practitioners the ability to restrict learning of unintentionally informative features that 'give away' the label, enabling the use of neural networks in the more realistic scenario with qualitatively worse data.

Danish Summary

Neurale netværk bruges i dag til flere og flere applikationer, ofte uden at slutbrugeren overhovedet er klar over det. Men på grund af antallet af parametre og mængden af træningsdata bliver de effektivt en sort boks. Vi har ingen måde at vide, af hvilke grunde en beslutning blev truffet. Når vi bruger neurale netværk til at træffe vigtige beslutninger, bliver dette farligt. I denne afhandling behandler vi dette behov for forklarbarhed med fokus på anvendelse. Vi bemærker især forskellen mellem mængden af forskning i fortolkningsevne og dens anvendelse i praksis og bidrager til at lukke dette hul. Til dette formål bidrager vi med et kapitel med et historisk perspektiv på tidligere tilgange til forklarbarhed. Desuden introducerer vi en ny tilgang til evaluering af forklaringsmetoder for en given anvendelse, så de praktiserende kan identificere den bedste metode til deres opgave. Ved hjælp af denne tilgang bekræfter vi, at der er naturlige forskelle i adskillige populære netværksarkitekturers fortolkning. Denne indsigt indebærer, at det kan være gunstigt at bruge en mindre nøjagtig arkitektur, der er mere fortolkelig til en opgave, hvor den neurale netværksklassifikation er sekundær. Vi viser, at et ensemble af forklaringsmetoder ikke kun forklarer et neuralt netværk bedre, men også er meget mere robust over for kontradiktorisk manipulation. Især når forklaringer indgår i den endelige beslutning eller bruges til at afgøre, om en given beslutning var i overensstemmelse med etiske krav, er det afgørende at gøre dem robuste til ondsindet manipulation. Afslutningsvis bidrager vi med en metode til at tilføre forudgående viden til et neuralt netværk via forklaringer. End-to-end-karakteren af neurale netværk forhindrer brugen af domænekendskab i algoritmen og kræver upartiske, uafhængige og identisk fordelt data til træning. Vores metode giver de praktiserende mulighed for at begrænse indlæring af utilsigtet informative attributter, der 'afslører' klassen, hvilket tillader brug af neurale netværk i det mere realistiske scenarie med kvalitativt dårligere data.

German Summary

Neuronale Netze werden heutzutage für immer mehr Anwendungen eingesetzt, oft ohne dass der Endbenutzer sich dessen überhaupt bewusst ist. Aufgrund der Anzahl der Parameter und der Menge der Trainingsdaten werden sie jedoch effektiv zu einer Black Box. Wir haben keine Möglichkeit zu erfahren, aus welchen Gründen eine bestimmte Entscheidung getroffen wurde. Wenn wir uns für Entscheidungen mit großer Tragweite auf neuronale Netze verlassen, ist dies gefährlich.

In dieser Arbeit gehen wir auf das Bedürfnis nach Interpretierbarkeit ein und konzentrieren uns dabei auf die Anwendung. Insbesondere stellen wir die Lücke zwischen dem Umfang der Forschung zur Interpretierbarkeit und ihrer Anwendung in der Praxis fest und tragen dazu bei, diese Lücke zu schließen. Zu diesem Zweck steuern wir ein Kapitel mit einer historischen Perspektive auf frühere Ansätze zur Erklärbarkeit bei.

Darüber hinaus stellen wir einen innovativen Ansatz zur Bewertung von Erklärungsmethoden für einen bestimmten Anwendungsfall vor, der es Praktikern ermöglicht, die beste Methode für ihre zu identifizieren. Mit diesem Ansatz bestätigen wir, dass es inhärente Unterschiede in der Interpretierbarkeit verschiedener populärer Netzwerkarchitekturen gibt. Diese Erkenntnis impliziert, dass es vorteilhaft sein kann, eine weniger genaue aber besser interpretierbare Architektur zu verwenden.

Wir zeigen, dass ein Ensemble von Erklärungsmethoden ein neuronales Netz nicht nur besser erklärt, sondern auch robuster gegenüber böswilliger Manipulation ist. Insbesondere wenn Erklärungen ein Faktor bei der endgültigen Entscheidung sind oder dazu dienen, festzustellen, ob eine bestimmte Entscheidung mit ethischen Anforderungen in Einklang ist, ist es von entscheidender Bedeutung, sie gegenüber böswilliger Manipulation robust zu machen.

Schlussendlich steuern wir eine Methode bei, mit der Domänenwissen mithilfe von Erklärungen in ein neuronales Netz eingebracht werden kann. Normalerweise verhindert die End-to-End-Natur neuronaler Netze die Verwendung von Domänenwissen im Algorithmus und erfordert unabhängige und identisch verteilte Daten für das Training. Unsere Methode gibt Praktikern die Möglichkeit, das Erlernen unbeabsichtigter informativer Merkmale, die die Klasse “verraten”, einzuschränken, wodurch die Verwendung neuronaler Netze in dem realistischeren Szenario mit qualitativ schlechteren Daten ermöglicht wird.

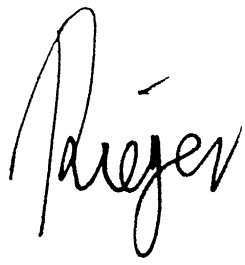
Preface

This Ph.D. thesis was prepared at the Cognitive Systems section at the Technical University of Denmark in partial fulfillment of the requirements for acquiring a Ph.D. degree in computer science.

This thesis consists of a summary report and a collection of four papers: one book chapter, one conference paper, and two workshop papers. The Ph.D. project was carried out at DTU during the period of September 2017 - October 2020 except for two one-month leaves of absence and a five-month research stay at UC Berkeley under Bin Yu.

The work presented in this thesis was funded by DTU Compute. It was supervised by Professor Lars Kai Hansen and co-supervised by Associate Professor Finn Årup Nielsen.

Kongens Lyngby, October 31, 2020



Laura Rieger

Acknowledgments

First, I would like to thank my supervisor Lars Kai Hansen for the extraordinary guidance and encouragement throughout this project. I am especially thankful for you taking on the project together with me in the first place.

Additionally, I would also like to thank my co-supervisor Finn Årup Nielsen, in particular for enlightening discussions on knowledge bases.

Thank you to my colleagues in the Cognitive Systems group at DTU for a great research environment and interesting discussions.

I would like to thank Bin Yu, Chandan Singh, Jamie Murdoch, and the rest of the Yu Group for their warm welcome and an educational stay at UC Berkeley.

I am grateful to Otto Mønsted Fond and Augustinusfonden for their financial support towards my stay abroad.

Thank you to my family and friends for their ongoing love and support throughout my life. I would especially like to thank my brother Phillip for proof-reading this thesis and look forward to returning the favor.

I would like to thank Morten for his unwavering support and love during my Ph.D. studies. In particular, his patient correction of Danish pronunciation, manuscripts, and of course this thesis was invaluable.

List of Publications

All publications are peer-reviewed.

Included in thesis

A Lars Kai Hansen and Laura Rieger. Interpretability in Intelligent Systems - A New Concept? In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 11700 LNCS, pages 41–49. Springer, Cham, 2019

B Laura Rieger and Lars Kai Hansen. IROF: a low resource evaluation metric for explanation methods. In *Workshop AI for Affordable Healthcare at ICLR 2020, Addis Ababa, Ethiopia*, 2020

C Laura Rieger and Lars Kai Hansen. A simple defense against adversarial attacks on heatmap explanations. In *Workshop on Human Interpretability in Machine Learning (WHI) at ICML*, 2020

D Laura Rieger, Chandan Singh, W James Murdoch, and Bin Yu. Interpretations are Useful: Penalizing Explanations to Align Neural Networks with Prior Knowledge. In *Proceedings of the International Conference on Machine Learning*, 2020

Not included in thesis

E Laura Rieger. Separable explanations of neural network decisions. In *Proceedings Workshop on Interpreting, Explaining and Visualizing Deep Learning (at NIPS)*. Proceedings Workshop on Interpreting, Explaining and Visualizing Deep Learning (at NIPS), 2017

F Laura Rieger, Pattarawat Chormai, Grégoire Montavon, Lars Kai Hansen, and Klaus-Robert Müller. Structuring Neural Networks for More Explainable Predictions. In *Explainable and Interpretable Models in Computer Vision and Machine Learning*, pages 115–131. Springer, Cham, 2018

G Laura Rieger, Rasmus M. Th. Høegh, and Lars K. Hansen. Client Adaptation improves Federated Learning with Simulated Non-IID Clients. *International Workshop on Federated Learning for User Privacy and Data Confidentiality in Conjunction with ICML 2020 (FL-ICML'20)*, 2020

Contents

List of Abbreviations

CD Contextual Decomposition

CDEP Contextual Decomposition Explanation Penalization

DARPA U.S. Defense Advanced Research Projects Agency

GAN Generative Adversarial Network

GDPR General Data Protection Regulation

GPU Graphics Processing Unit

ICML International Conference on Machine Learning

IROF Iterative Removal Of Features

LIME Local Interpretable Model-agnostic Explanations

LRP Layerwise Relevance Propagation

LSTM Long Short Term Memory

ReLU Rectified Linear Unit

ROAR RemOve And Retrain

SENN Self Explaining Neural Network

SHAP Shapley Additive Explanations

TCAV Testing with Concept Activation Vectors

VAE Variational Autoencoder

CHAPTER 1

Introduction

In this chapter we outline and motivate the work presented in this thesis.

1.1 Scope of this thesis

The topic of this thesis is the deployment of interpretation methods for neural networks in practical use cases. Our goal is to make explainability for neural networks more approachable and useful for practitioners.

Explainability has enjoyed a lot of attention in recent years. In a 2017 roll-out, the U.S. Defense Advanced Research Projects Agency (DARPA) declared explainability as one of the key factors of third-wave AI, pledging two billion USD into its development. In 2018, the General Data Protection Regulation (GDPR) came into effect, including a 'right to explanation' and fostering a lot of debate about the extent, usefulness, and actual consequences of this right [45, 88, 144]. At the recent International Conference on Machine Learning (ICML) 2020, no fewer than three workshops were dedicated to explainability in AI. As a result, there is no shortage of approaches, definitions, and methods for explainability.

However, in real-life uses of AI, explainability is still rarely considered or deployed, pointing us to a gap between the development of explainability methods in theory and their adoption in practice [16]. This thesis addresses multiple aspects of explainability in practice.

- In Appendix A we connect the field of explainability to its history, giving a perspective on how the view on explainability has changed in the current time.
- Appendix B introduces a low-resource method to evaluate explanation methods for a specific use case.
- Appendix C shows that aggregations of explanation methods have multiple advantages, circumventing the need to choose a specific method and making explanations more robust against malicious attacks.
- Appendix D introduces a method to enforce prior knowledge in a neural network via explanations.

Three other contributions are included in the appendix but are not a part of this thesis.

1.2 Motivation

1.2.1 Neural networks are black boxes

Although neural networks have been invented a long time ago, increasing their depth to the extent necessary to model complex functions was infeasible until recently due to computational restrictions. With the recognition that Graphics Processing Units (GPUs) can be utilized to parallelize many small computations, deep neural networks were suddenly possible to train.

This break-through enabled neural networks to become the state of the art in many different tasks such as image recognition or natural language processing. As an example, in the ImageNet 2012 challenge, a convolutional neural network achieved a top-5 error of only 15.3%, more than 10% lower than the next-best method. This was widely considered as the start of the deep learning renaissance. The ImageNet challenge was a yearly contest in image classification. The goal is to classify images from 1000 categories as correctly as possible. The training data contains more than 14 million annotated images. By now, neural networks achieve superhuman performance in this task [53, 100].

Neural networks have also been shown to achieve good and sometimes even superhuman performance in diverse tasks such as playing games, recognizing skin cancer, transcribing speech, or driving cars [33, 84, 89, 123]. However, to train neural networks on these tasks, a vast amount of training data is necessary. The amount of training data needed means that it is no longer possible to manually inspect each training point. For example, it was only very recently discovered that the aforementioned and extremely popular ImageNet database included racist categories and tags [129].

Since current neural network architectures contain many millions of parameters, we can no longer control or even check what the neural network has learned and what input features are important for classification. Combined with the fact that we cannot inspect the training data either due to the sheer mass, we can check whether the output is correct but we have no way of checking if it was made for the reasons we expect. Neural networks effectively become black boxes.

1.2.2 Black boxes are dangerous

> "Ginny!" said Mr. Weasley, flabbergasted. "Haven't I taught you anything? What have I always told you? Never trust anything that can think for itself if you can't see where it keeps its brain."

While Mr. Weasley in 'Harry Potter and the Chamber of Secrets' by Rowling [111] was referring to a part of an evil wizard's soul, the statement is just as valid for modern machine learning. Particularly for decisions with high-impact, it is dangerous to assume that high classification accuracy implies that the neural network or any other machine learning algorithm is using the same reasoning as we humans do. The difference between the two might be due to unknown irregularities in the data, previous discrimination that is reflected in the data but that we no longer wish to carry forward or intentional manipulation of the input. A skilled human can pick up on those discrepancies between what the algorithm should base the decision on and what it does base it on, given background knowledge and an explanation.

It should be noted that an algorithm can be biased even when the attribute in question is not fed into the algorithm, simply due to dependencies in the data distribution that the algorithm can learn. In 2017, Amazon discontinued an AI-based hiring tool after the model continued discriminating against women after multiple attempts to fix the bias by masking out input were unsuccessful [69]. If terms such as 'women's chess club' were removed, the model picked up on male-favored descriptions such as 'executed' instead of focusing on the actual skills listed. In another well-known example, ProPublica showed that COMPAS, a tool for aiding judges with determining bail, was biased against people of color [74]. On average, people of color were most likely to be mistakenly predicted to re-offend from all ethnicities, twice as likely as white people. Conversely, white people were most likely to be wrongly predicted to not re-offend, twice as likely as people of color. It was also recently found that the algorithm judging credit card limits for Apple Card seems to be severely biased against women, giving them a much lower credit card limit when all other factors are equal [98].

In all these cases, deploying the model without extensive inspection simply due to its high accuracy would be and has been harmful. Moreover, all these cases also have in common that the people whose lives are influenced by these decisions have no way of challenging or even knowing about the biased algorithm they have been subjected to. Without internal inspection or people knowing a comparable person with only the relevant attribute changed, i.e. an unintentional counterfactual explanation, the harmful bias in the model would not be noticed. Having an explanation along with the output decisions would not only enable people to challenge the decisions they are subject to, but it would also enable much quicker debugging of the machine learning model in regards to what was learned from the data. It has also been shown that explanations are useful for gaining new scientific insights in physics, chemistry, or biology [109].

In summary, having explainability along with high classification accuracy is desirable and in some cases even obligatory because it enables us to check that the neural network is working as intended, examine whether the data contained unfair biases, and learn from the correlations that the neural network picked up on. Common reasons that motivate explainability are therefore ensuring fairness, debugging models, and

the discovery of novel physical effects.

1.2.3 Looking into the black box

Given the laid out reasons, it should be no surprise that explainability has gained a lot of academic attention in recent years. However, despite the clear need for explainability as well as the number of approaches designed to solve it, the application of explainability in practice is surprisingly sparse. Querying thirty companies revealed that not a single one is currently showing explanations of decisions to their users [18]. In a more encouraging result, Bhatt et al. [18] however also found that explainability methods are increasingly used internally to debug or sanity-check models during development.

While the focus on explainability in recent years has translated into a lot of research, most papers introduce completely new approaches instead of building on prior work. Moreover, even comparisons to different previous approaches remain the exception, not the norm. While this leads to a wide variety of approaches, it also makes it hard to determine the current state of progress. It also makes it harder to choose a suitable explanation method for a task as a practitioner or simply a machine learning researcher not familiar with this particular sub-field. In fact, at the time of writing this thesis, no widely accepted benchmark tasks or datasets exist at all.

A reason for this could be the wide range of use cases for explanations. An engineer debugging a model will look for a different explanation than a user trying to raise their credit score or a doctor using a neural network for diagnosis. To be useful, explanations need to not only be aligned to the model they explain but also take the receiver of the explanation into account.

Particularly for visual explanations for convolutional neural networks, the use case for the resulting explanations is often not explicitly discussed when introducing a new method or approach, leaving the reader with no clear objective against which to evaluate the approach. As a result, clear explanations that resemble the original input, are often favored even though they are not necessarily better [1].

In summary, while there has been a great amount of interest and work in explainability, many open questions remain. The lack of a united view and goal may contribute to the currently sparse adoption of explainability in practice. This is despite the large potential of explanations and the risks of deploying black box models without explanations. The accidental uncoverings of biased algorithms that already have very real consequences further highlight the need for systematic inspection of the reasoning happening within machine learning algorithms, including neural networks, in deployment.

1.3 Thesis structure

As the papers comprising this thesis look at different aspects of explainability, Chapter 1 will be followed by basic background knowledge in Chapter 2 and a comprehensive overview of explainability in Chapter 3.

To unify the vocabulary, Section 3.1 in Chapter 3 will give an overview of current position papers, surveys, and definitions of explainability. Since fairness is frequently named as a motivation for explainability, Section 3.2 sketches out current fairness measures. Sections 3.3 to 3.5 contain current explainability methods, ascending from single classification explanations to model and dataset explanations. Finally, Section 3.6 contains ways to process, attack or evaluate explanations.

Having set a structure of the current field, Chapter 4 will embed our contributions to the field in this structure. For each paper, we will motivate the work contained in the paper, summarize the contributions of the paper, look at the subsequent progress and sum up the wider implications of the work, including usability and potential dangers.

Chapter 5 concludes this thesis. We summarize and evaluate the past three years of research, and give an outlook into the future of explainability.

CHAPTER 2

Background

This chapter will lay out the background knowledge for the rest of the thesis, covering common neural network architectures and ways of training.

2.1 Supervised learning

Fundamentally, we can distinguish between *supervised learning* and *unsupervised learning* in machine learning. Whereas the goal of unsupervised learning is to discover structures in the data without labels, the goal of supervised learning is to learn a function that maps a given input $\boldsymbol{X} \in \mathbb{R}^{n \times m}$ to an output $\boldsymbol{y} \in \mathbb{R}^n$. In general, we talk about a classification problem when the outputs are discrete $\boldsymbol{y} \in \mathbb{Z}^n$ such as in digit classification. When the outputs are real-valued $\boldsymbol{y} \in \mathbb{R}^n$, we talk about a regression problem. An example would be to predict the market value of a house based on its features.

Given a particular algorithm f_θ and the data, the goal is to learn the best parameters θ for the algorithm. What is best is often measured with the *cost function* or *loss function* $L(f_\theta(\boldsymbol{X}), \boldsymbol{y})$. We want to find the set of parameters θ for which

$$\theta = \text{argmin}_\theta L(f_\theta(\boldsymbol{X}), \boldsymbol{y})$$

holds. The choice of the loss function is dependent on the task, dataset, and algorithm at hand.

For some algorithms such as linear or logistic regression, the optimization can be solved in closed form. For others such as neural networks this is intractable and the parameters need to be iteratively updated.

2.2 Neural networks

Feed-forward neural networks or multi-layer perceptrons are general function approximators that can approximate any function, given enough training data. Generally, they are comprised of an input layer, any number of hidden layers and an output layer, each representing a simpler transformation. They are very loosely inspired or

are often compared to the way brains work. As with brains, neural networks contain a number of neurons that are connected to each other.

2.2.1 Common layers

We describe the most common layer types for neural network architectures.

Dense layers or fully connected layers have each input neuron connected to each output neuron.

$\boldsymbol{A} = \boldsymbol{W} \times \boldsymbol{x} + \boldsymbol{b}$

with $\boldsymbol{A} \in \mathbb{R}^a$, $\boldsymbol{W} \in \mathbb{R}^{a \times b}$, $\boldsymbol{x} \in \mathbb{R}^b$ and the bias $\boldsymbol{b} \in \mathbb{R}^a$.

Convolutional layers are a type of layer that drastically reduces the number of parameters by exploiting a known grid-like structure in the input such as in images. Each channel has a smaller receptive field than the input that is sliding over the entire input. For a two-dimensional input, each output is computed by

$$A_{i,j} = \sum_m \sum_n X_{m,n} * W_{i-m,j-n}$$

Commonly, each convolutional layer has multiple channels. The step size, number of channels and size of the receptive field are hyperparameters.

Non-linearities To enable the neural network to represent complex functions, linear transformation layers are followed by a non-linear function, enabling the neural network to model any given function. The most common ones are the Rectified Linear Unit (ReLU) and softmax.

$\mathrm{ReLU}(\boldsymbol{x}) = \max(\boldsymbol{0}, \boldsymbol{x})$

$\mathrm{Softmax}(\boldsymbol{x}) = \frac{e^{\boldsymbol{x}}}{\sum_j e^{x_j}}$

The softmax is most often used in the last layer to compress the output to be between 0 and 1 and all outputs to sum up to one.

Pooling layers reduce the dimension of the internal representation. The values of the given receptive field (for 2D input, this is commonly 2×2) are either averaged or more commonly the maximum value of the receptive field is taken.

Recurrent layers are used to process sequences of an undetermined length. Along with the output, each neuron also retains an inner state that influences subsequent computations. They are commonly used for language processing. Arguably the most common type is the Long Short Term Memory (LSTM) unit [60]. An LSTM unit has

learnable gates that control how the input is incorporated into the internal state as well as how much the input and internal state control the output.

Residual layers add skip connections by adding the input identity to the output of the layer. This circumvents the problem of vanishing gradients and enables training of much deeper architectures than previously possible [54]. In the original paper by He et al. [54], skip connections were added in with bottleneck modules. In each residual building block comprised of three layers, the input and output size of the building block are equivalent but the representation size within the building block is first increased to permit more complexity in the representation and then decreased with the subsequent layers.

2.2.2 Training

Neural networks are trained by iteratively updating weights. Since each layer transformation is differentiable, the entire neural network function can be automatically differentiated in regard to the loss function. Most commonly, the loss function is the categorical cross-entropy between the output of the neural network and the true label for each sample.

As the entire neural network is fully automatically differentiable, the gradient of the loss function in regards to each weight parameter can be calculated for each sample in the dataset. During training, each weight is updated according to their gradients in the direction that minimizes the loss function. While the direction of the update is determined by the gradient, the scale of the update is determined by an *optimizer*. For each of the following optimizers, $\theta_{n,t}$ denotes a weight parameter with index n at time step t.

Stochastic Gradient Descent is the simplest way to update the weights. In each iteration, the gradient updates from a number of samples from the training dataset are summed up and the weights are updated accordingly. Each parameter θ_n of the neural network gets updated according to

$$\Delta\theta_{n,t} = -\eta \frac{\partial L}{\partial \theta_{n,t}}$$

with η being the learning rate.

In a modification taking **momentum** into account, each weight parameter has a running average that saves previous gradients. In each update, the momentum for each weight gets updated with the current gradient and the weights are then updated according to the momentum. This adds another hyperparameter α, controlling how much the momentum gets updated each iteration.

$$v_{n,t} = \alpha v_{n,t-1} + (1-\alpha)\frac{\partial L}{\partial \theta_{n,t}}$$
$$\Delta\theta_{n,t} = -\eta v_{n,t}$$

Both of these approaches have the disadvantage that the update rate for each weight is the same. Ideally, the learning rate for weights which are rarely updated should be higher than for weights which are often updated.

Adagrad addresses this issue [31, 112]. For each parameter, the universal learning rate is divided by the square root of the square of all past updates to this weight parameter plus a smoothing constant. In this way, learning rates for parameters which are frequently updated decay much faster than learning rates for parameters which are rarely updated. A disadvantage is that the learning rate for each parameter decreases monotonically.

$$g_{n,t} = g_{n,t-1} + \left(\frac{\partial L}{\partial \theta_{n,t}}\right)^2$$
$$\Delta\theta_{n,t} = -\eta \frac{\partial L}{\partial \theta_{n,t}} \frac{1}{\sqrt{g_{n,t}} + \epsilon}$$

RMSProp attempts to resolve this by using the moving average rather than the sum of the squares of all past gradients [59, 112].

$$g_{n,t} = 0.9 g_{n,t-1} + 0.1 \left(\frac{\partial L}{\partial \theta_{n,t}}\right)^2$$
$$\Delta\theta_{n,t} = -\eta \frac{\partial L}{\partial \theta_{n,t}} \frac{1}{\sqrt{g_{n,t} + \epsilon}}$$

Adam extends on this by also keeping a running average of the gradient update itself along with the learning rate [73, 112]. To our knowledge, Adam is currently the most commonly used optimizer. β_1 and β_2 are hyperparameters, controlling how fast the moving averages get updated.

$$m_{n,t} = \beta_1 m_{n,t-1} + (1-\beta_1)\frac{\partial L}{\partial \theta_{n,t}}$$
$$g_{n,t} = \beta_2 g_{n,t-1} + (1-\beta_2)\left(\frac{\partial L}{\partial \theta_{n,t}}\right)^2$$

To account for the parameters' bias towards zero in the initial period, the update is done with bias-corrected parameters:

$$\hat{m}_{n,t} = \frac{m_{n,t}}{1-\beta_1^t}$$

$$\hat{g}_{n,t} = \frac{g_{n,t}}{1 - \beta_2^t}$$

$$\Delta\theta_{n,t} = -\eta \frac{\hat{m}_{n,t}}{\sqrt{\hat{g}_{n,t}} + \epsilon}$$

All methods covered only take the first-order differential into account. There are a few optimization schemes that also use the second-order differential, the Hessian, for the weight updates.

Before training it is not clear how many epochs, i.e. how many complete passes through the training data set, are needed for the loss to converge. Therefore, it is common to do *early-stopping* where the training is stopped once the validation loss is no longer improving.

All methods have hyperparameters controlling the training in addition to the architecture choice. As validating on hyperparameters and architectures is costly especially with large neural networks, optimal hyperparameter choice is an active field of research.

2.2.3 Common regularizations and modifications

There are many approaches to regularize neural networks during training. It is possible to regularize the weights with **L1 (Lasso)** or **L2 (Ridge)** regularization where the absolute or squared values of all weights are added to the loss function to pull the weights towards zero.

Another possible way to regularize neural networks is via **dropout** [131]. In a dropout layer, each input neuron is set to 0 with probability p, commonly e.g. with $p = 0.5$ and the remaining neurons getting their signal multiplied with $\frac{1}{p}$. This avoids co-adaptation of neurons, translates to effectively sampling from a large number of neural networks and is effective at preventing overfitting during training. Gal and Ghahramani [39] have suggested that dropout could also be used at test time to approximate a Bayesian neural network. By running an input through the neural network with dropout, we can effectively sample from a distribution of weights.

Batch normalization has been suggested to speed up the training of neural networks [63]. Batch normalization addresses the problem of the input distribution of each layer shifting during training. Ideally, the input of each input neuron has an expectation of zero and a standard deviation of one. Batch normalization addresses this by normalizing for each input variable the input over the mini-batch. This significantly speeds up training as layers do not have to adapt to the changing training distribution.

CHAPTER 3

Related work

Explainability touches on many different areas in machine learning and is itself not easily defined. In Section 3.1 we give an overview of papers that motivate or argue against explainability as a necessity and attempt to define it.

A frequent motivation for having explainability for your machine learning algorithms is to ensure that the algorithm is 'fair'. An example of this would be the need to prove that minority groups are not discriminated against. By having an explanation either for a single example, a group with a specific feature or the entire model, it is possible to check whether the algorithm is using discriminative attributes that we do not wish to use in the decision, for example because of fairness or because they are an artifact of the dataset. In Section 3.2 we review the most common fairness measures.

Our work is mainly concerned with post-hoc explainability, i.e. algorithms that create an explanation based on an already existing neural network with a given architecture as opposed to designing the neural network with inherent explainability. In Sections 3.3 to 3.5 we will cover works dealing with post-hoc explainability. Finally, we introduce applications of explainability in Section 3.6.

3.1 Definitions and position papers

In contrast to accuracy or other traditional performance measures, explainability is a much more vague goal with no immediate definition. As a result, there are a number of papers giving different definitions of explainability. We give a short overview of recent influential papers in the field[1]. This serves also to establish the vocabulary for the later parts of the thesis.

The words *interpretability* and *explainability* are often used interchangeably. Unless otherwise noted, we follow the convention of Gilpin et al. [43] and see explainability as a sub-field of interpretability: Interpretability refers to the general 'understandability' of a model. As an example, we often assume that the weights of a linear classifier are often intuitively understandable for a human though this may not necessarily be true [52]. An explanation and accordingly *explainability* refers to a summary of important factors for a specific objective.

[1] Older papers will be covered in Section 4.1

In 2018 the GDPR came into effect in the European Union (EU) [144]. Among other protections, it also included the 'right to explanation', stating according to Wachter et al. [144] that if a person is subjected to an algorithmic decision, they have the right to an explanation of this decision. This has been the subject of much debate as the formulation in the GDPR was relatively vague [32, 45, 88]. Mittelstadt et al. [88] pointed out that there are several caveats to this right. As we cover in this section, the definition of an explanation is not clear in itself. In the GDPR it could either refer to an explanation of how the algorithm works or an explanation of the decision itself. Along with other caveats, this severely weakens the GDPR as motivation for work on explainability.

Lipton [83] recognizes five comprehensive desiderata that are frequently named as the motivation of interpretability research:
Interpretability research should foster *trust* in the model, by increasing our understanding of what the model considers relevant. Second, it should enable us to make hypotheses about the *causality* in the real world that the machine learning algorithm is modeling. Similarly, we can use interpretability to ensure that the learned model has *transferability*, i.e. that the model is not exploiting spurious correlations specific to the training set as demonstrated in [62]. Interpretability should enable us to gain more information about the model of the world that the algorithm has learned and potentially about the world itself - it should be *informative* for humans. Lastly, interpretability can enable us to ensure that the decisions made by the algorithm are *fair*.

Gilpin et al. [43] states that an explanation can be evaluated in two, sometimes at-odds, ways: An explanation should be *interpretable*. This means that it should "describe the internals of a system in a way that is understandable to humans." [43] Second, the explanation should be *complete* and correct. It is easy to imagine explanations that fulfill only one of the two desiderata perfectly. The weights of an neural network are a fully complete explanation of classification but a human will not understand this explanation. Indicating the input dimension with the maximum relevance to the decision as an explanation will be understandable but not complete. Gilpin et al. [43] states that explanation methods should be classified and evaluated according to those two desiderata for an explanation.

Doshi-Velez and Kim [28] note that all current interpretability research relies to some extent on "you'll know it when you see it." Stating that an explanation has to be understandable just leads us to the question of "what does it mean to be understandable?" They classify evaluation approaches for interpretability into three categories [28]. The first is functionally grounded evaluation which does not rely on humans but only on the fulfillment of quantitative metrics of a definition of interpretability. The second category is human-grounded evaluation which evaluates the quality of the interpretation algorithm on simple tasks. This includes qualitative evaluation. The most costly but also most useful option is to evaluate how well the interpretability method supports humans in actual tasks. An example would be to

measure how fast a doctor can diagnose a disease when given an explanation of a machine learning algorithm diagnosis for the same disease.

Taking a different route, Rudin [113] asserts that using models that are not interpretable by design such as neural networks always carries an inherent risk and should not be used for high-stakes decisions. She argues that an explanation for a model too complex for humans to understand can by definition not be faithful to the original decision. Instead, black-box models should only be used as a performance baseline for a subsequent inherently explainable model.

In recent work, Murdoch et al. [92] defined machine learning interpretability as "the use of machine-learning models for the extraction of relevant knowledge about domain relationships contained in the data". In that sense, interpretability becomes a natural part in the data science life cycle after a model has been trained. Now the model can be used and examined via interpretability methods to determine whether the given goal has been reached.

Summing up the current work on defining explainability, there have been comparatively few papers on what the goal of explainability research is compared to papers on how we reach this goal. Many definitions of explainability seem to be circular, requiring understandability which is hard to define in itself [28]. At the same time, humans have an intuitive understanding of what an explanation is. Similarly, there are many potential benefits of explainability with one of the most frequently named motivations being the need to ensure fairness [28, 29, 43, 83]

3.2 Fairness

Fairness and explainability are closely linked as research fields, especially since fairness and the examination thereof is often given as a motivation for explainability. Given that fairness in itself is not easily defined, explainable algorithms can serve as tools to gain more insight into the statistical relations between input and output and decide on what measure of fairness is most suitable to the current application.

As such, we give a short overview of common measures of fairness without claim to completeness. In general, we assume that there is a binary outcome classifier and that we want to test whether a particular group or attribute is potentially being discriminated against. We also assume that there is a ground truth (that we may not know about) for the outcome.

An easy criterion to fulfil is "fairness through unawareness". The attribute in question is not included in the input features for the algorithm.

Equal parity indicates that each group is equally represented in the positive outcome group, regardless of their percentage in the true positive group or the general population [95]. For each subgroup, the total sum of false and true positives should be close to each other. *Proportional parity* indicates that the proportion of positive

outcome in each subgroup should be equal or similar to each other. Neither of these two take the true positive rates in each subgroup into account [95]. As a result, they are relatively easy to implement and check. Assuming that the rates of true positive and true negatives in each sub-population is different, we may want to ensure that the number of falsely classified samples from each subgroup is similar. This is referred to as *equalized odds* [50]. Depending on whether the consequence of the classified outcome is positive or negative for the subjects, we can focus on either *false-negative rate parity*, i.e. equal sensitivity, or *false-positive rate parity*, i.e. equal specificity, for all subgroups.

All fairness measures can only be fulfilled at the same time if the true incidence rates are equal for the groups to be compared. As a result, if fairness is a concern, a choice has to be made on the deciding fairness measure. This choice on which fairness measure to pursue has to be made individually for each situation, i.e. there is no universally appropriate fairness measure. Factors to take into account are the desired ideal outcome as well as the consequences for individuals if they are falsely classified.

3.3 Single example explainability

In recent years a lot of approaches for single example explainability have been proposed. The term here refers to approaches that explain a single classification by a neural network. Some methods are architecture-agnostic, others only work for a specific architecture or a specific input type such as text or images. While there are a lot of different methods, they can be categorized according to different criteria.

In Table 3.1 we give an overview of the usual advantages and disadvantages with the different approaches.

We will go on to describe representative methods from each category.

Table 3.1: Different approaches for single example explainability have different advantages. The right approach depends on the needs for the specific application.

	understandable	sufficient	low overhead	efficient
Backpropagation	-	o	+	+
Local explanations	+	o	+	-
High-Level	+	-	-	o

Backpropagation based methods are explainability methods that utilize the weights of the neural network to trace the output decision back to the input of the neural network. As they backpropagate relevance back onto the input features, they have the disadvantage of only representing relevance in linear combinations of the input. If, as in most cases, the classification task is dependent on non-linear combinations

of the input they require the practitioner to be able to partition and make out these combinations of input features. As such they are naturally less understandable and not sufficient on their own without prior knowledge. As backpropagation is highly optimized in most modern frameworks, they are generally computationally efficient.

A basic way to do that is to take the gradients of the output in the input dimensions [124]. Intuitively, if an input dimension, for example a pixel in an image, has a high gradient, i.e. if a small change in this pixel would result in a relatively big change in the network output, this pixel is important for the classification. Taking the gradient of the output over the input results in an explanation with the same dimensions as the input. As shown in Simonyan et al. [124], this results in a heatmap of relevances that does to some degree align with what humans would consider important in an image. However, there will be a lot of noise in the explanation.

Based on this idea, a number of variations have been proposed: Selvaraju et al. [117] introduced Guided Backpropagation, where a ReLU operation is applied after each layer. This reduces the noise in the resulting explanation considerably. Grad-CAM, which can be applied to convolutional neural networks only applies the backpropagation through the dense layers of the convolutional neural network [117]. Through the convolutional layers, the relevances are up-sampled to match the resolution of the input, resulting in a more coarse explanation that highlights large important areas in the image.

Integrated Gradients, as proposed by Sundararajan et al. [133], integrates sampled gradients over the path from a chosen baseline to the input to be interpreted. With e.g. a completely black picture as a baseline, we take the average of the gradients of a number of images on the linear interpolation between the black image and the original image. For neural networks trained on images, this leads to considerably less noisy explanations that highlight the object to be classified.

Layerwise Relevance Propagation (LRP) is another backpropagated approach where the relevance is in each layer split up onto the neurons of the previous layer according to how much they contributed [10]. They propose multiple alternative approaches as to how relevance is distributed in each layer. In a follow-up Arras et al. [9] extended the method to also work on LSTM network structures.

The aim of DeepLIFT aim is to explain the difference of the output from a reference value in terms of the difference of the input from a reference value [122]. The choice of the input neutral value is dependent on the problem at hand. DeepLIFT then recursively calculates the intermediate layers contribution scores with the constraint that the contributions summed up in each layer have to be equal to the output difference.

Smilkov et al. [130] proposed SmoothGrad as an augmentation to existing explanation techniques to decrease relevance on irrelevant parts of the input by sampling many inputs with Gaussian noise added. For each noisy sample an explanation with one of the previously introduced methods is extracted. By averaging out local variations on

the gradient, the final explanation more meaningfully highlights important parts of the input.

In summary, most backpropagation-based methods have in common that they have a relatively low computational cost as they only rely on backward-passes through the neural network. Modern frameworks are designed to minimize the cost of backward-passes, as they are essential for training. SmoothGrad and Integrated Gradients require multiple backward passes and therefore take considerably longer than other approaches.

A disadvantage is the need to know beforehand what should be important to compare with the explanation. This is especially pronounced for images where each pixel has its own importance. In contrast, humans will not assign importance to each pixel but rather to features in the image. This makes it hard for humans to evaluate if a pixel-wise explanation is aligned with our understanding. Indeed, Adebayo et al. [1] showed that many popular explanation methods are hard to distinguish from simple edge detectors.

Local approximation methods are explanation methods that approximate the neural network with a less complex local model around the current point to be explained.

One of the most well-known methods is Local Interpretable Model-agnostic Explanations (LIME) [101]. It works by approximating the neural network (or another machine learning algorithm) with a linear model around the point to be explained. For high-dimensional inputs such as images, they first divide the input into a small number of semantically meaningful parts. They then learn a linear classifier that approximates the neural network output while removing parts of the input. Since there are 2^n possible combinations with n being the number of parts in the input, the necessary sampling leads to a relatively high variance in the explanation [152].

Shapley Additive Explanations (SHAP) approximates the contribution of each input feature in a game theory approach by representing the output classification as a game and measuring the marginal contribution of each feature [86]. For neural networks they also propose DeepSHAP, a combination of DeepLIFT and SHAP.

High-level explanation methods are methods that utilize the neural network architecture in a different way to explain decisions.

Alvarez-Melis and Jaakkola [7] proposed explicitly enforcing explainability during training with Self Explaining Neural Networks (SENNs). Starting from the idea that linear classifiers are always understandable, their goal is to make them gradually more complex while retaining the interpretability. This is done by applying the linear classifier not on the input directly but on learned relevances of a (small) number of learned features. By adding a reconstruction loss on an autoencoder, they enforce the learned features to be meaningful.

Contextual Decomposition (CD) as introduced by Murdoch et al. [91], Singh et al. [125] follows a different approach. Contextual Decomposition requires that the practitioner first divides the input to be explained into two parts. The classification is then explained as the relevance of those two parts compared to each other. This is done by essentially passing both parts separately through the neural network, linearizing intermediate layers as necessary. While the original method only returned an explanation in terms of two parts, the method was later extended into hierarchical explanations, specifically for text [125].

Koh and Liang [75] uses influence functions to determine how single training samples can harmfully change the prediction of multiple or single test images. Other works also deal with single example explanations [2, 23, 47, 154].

In summary, there has been a lot of interest and work with various approaches into single-example explainability. As of current, there is no definite state of the art. In practice, local approximation methods and LIME in particular seem to be preferred, perhaps due to the relative simplicity of the approach.

3.4 Feature explainability

Another intriguing sub-field of explainability with several aspects is feature explainability.

For generative models such as Generative Adversarial Networks (GANs)[2] we are often interested in how a particular feature in the image is influenced by the latent space [44, 64, 85, 119, 143]. The challenge here lies in finding meaningful directions in the latent space that correlate to human-recognizable features in the output image. Current work often relies on finely-grained knowledge about potential directions to discover them. Autonomous discovery of interpretable direction, while potentially not possible, remains an open field with many exciting research questions [85].

Counterfactual explanations are another possible approach. In contrast to the meaning causal inference, we here refer to a minimally changed example with a different classification (without necessarily building a causal graph). In general, it is also a requirement that the changed example stays on the data manifold, i.e. is realistic. s Counterfactual explanations answer the question “What would have to change for the output classification to change?” or “What is the closest meaningful sample with a different classification? They have the advantage of being intuitive to humans [90]. Additionally, counterfactual explanations are well-suited to applications in fairness as they enable us to inspect whether discriminative attributes were a contributing factor in a decision.

[2]Since GANs are often used and well-known for image generation, we will also focus on this use case in the subsequent text.

Fong and Vedaldi [38] proposes to learn minimal masks that cover informative content in the image. They propose alternative ways of infilling the informative region by blurring, noise or a constant value.

Chang et al. [22] proposed creating counterfactual image explanations by in-filling parts of the input image that would most change the output classification. Compared to heuristics such as blurring or using one reference value this leads to more natural counterfactual explanations that are close to the data manifold. Their method, FIDO-CA, uses a GAN to fill in informative regions conditioned on the remaining image.

Goyal et al. [46] also create counterfactual explanations. In contrast to Chang et al. [22], they use patches from the training set to "cover up" informative regions of the original input image. While this guarantees that the new regions of the image are in the data manifold, the method relies on the classes in the dataset not being too fundamentally different from each other so that composite images still look natural.

Antorán et al. [8] proposed CLUE (Counterfactual Latent Uncertainty Explanations), a method to create counterfactual explanations for the uncertainty of the output in Bayesian neural networks. They utilize the decoder of an auxiliary Variational Autoencoder (VAE) to enforce that the counterfactual explanation stays close to the data manifold the Bayesian neural network was trained on. They then optimize a loss function comprised of the uncertainty of the output and a difference measure to the original input to obtain counterfactual explanations of uncertainty. Explanation of uncertainty, while interesting, has otherwise been sparsely explored.

In recent work, Singla et al. [126] proposed an approach that, while not strictly counterfactual, follows a similar idea: By learning a walk that negates the originally predicted property, they produce a range of images that together are a counterfactual explanation. For the example of smiling, they generate images that interpolate between a smiling and a frowning face while requiring the images to still look plausible to an adversarial network.

Captioning models have been used for explanations in natural human language, sometimes in a multi-modal fashion [55, 56, 97]. In general they require a finely annotated dataset with a list of named features available for each image. This enables recognizing features common or unusual for each class. Speech generation is then used to create a natural sentence, e.g. "It is not a Scarlet Tanager because it does not have black wings." [56]. A very similar approach was recently proposed by Koh et al. [76]. Instead of training a neural network end-to-end, they train a feature classifier on the annotations and then add a linear classifier based on the feature vector. This enable easy counterfactual explanation generation but is severely limited by the reliance on knowledge of contributing features for each class and annotations of those for a large part of the dataset.

In summary, there has been research into feature explainability both for unsupervised as well as for supervised learning. For supervised learning a clear advantage is that

the end-user is not required to already know about and recognize meaningful features such as in Section 3.3.

3.5 Dataset and model explainability

Quite often we are interested in what the neural network has learned as a whole. Since training is done with data, this also extends into learning what correlations there are in the dataset. Following Gilpin et al. [43] as introduced in Section 3.1, explanations should be evaluated based on their interpretability and completeness. The latter requirement, completeness, is considerably harder to fulfil, as it is harder to explain an entire model rather than a single point. Conversely, model explainability is often of interest when checking a model for fairness and safety or for better understanding of architectures and training effects in general [83].

Bau et al. [14] is a large study of where in the neural network concepts such as color or structure are learned. They assembled a large dataset of images representing different categories (scenes, objects, parts, material, texture and color) and tracked where and how the activation of a neuron in the neural network correlated with the representation of those concepts. In line with our later work in Appendix B, they found that architecture and training modifications such as dropout do have an effect on the interpretability or disentangledness of concept representations in the neural network. An expected finding was that different concept categories emerge in different parts of the network with complex concepts emerging later than simple concepts such as color.

Geirhos et al. [40] found that modern neural networks highly prefer image texture compared to shape. This is at odds with how humans perceive an image. By synthesizing images with e.g. a cat *shape* but elephant skin *texture* they quantified which cue is more important, in effect 'explaining' what cues the neural network had learned. Training a neural network to prefer shape makes them more robust towards adversarial attacks. Conversely, making neural networks more robust also decreases the preference towards texture bias, indicating that there is a robust connection between vulnerability and preference for texture [3].

Kim et al. [71] proposed to quantify how much neural networks rely on particular concepts with Testing with Concept Activation Vectors (TCAV). This approach is similar to earlier work by Alain and Bengio [5], albeit with more complex neural networks. TCAV requires assembling a dataset of inputs[3] *with* and *without* the concept present. For each layer they learn a linear classifier that separates the activations when presented with images with and without the concept. This enables quantification of how important a concept is for classification and comparison of how similar two concepts are for the neural network. Particularly the two latter are intriguing from a fairness standpoint, as it permits questions such as "Is gender or attire more important

[3] As with the generation of data, images are the predominantly interesting domain

for classification of an image as 'doctor' or 'nurse'?". However, the assembly of datasets hinges on the practitioner being aware of potential bias and being able to gather enough data to train the linear classifier, making this a very expensive approach. The authors later expanded on this work with the automated discovery of concepts by segmenting images into super-pixels and then finding super-pixels that activate similar neuron configurations [42].

In other work, van Steenkiste et al. [140] found that disentangled representations, along with being conducive to human understanding, also appear to aid better performance of latter tasks. Besserve et al. [15] examined the inner representations of generative models with counterfactual manipulation of the latent representation. They found that disentangled representations to some extent do emerge in generative models during training, enabling the meaningful traversal of output features without further training. In a very comprehensive 'activation atlas', Carter et al. [21] visualized what concepts a neural network learn by feature inversion, i.e. investigating what inputs particularly activate sets of neurons.

Recently attention mechanisms, specifically transformers, have been used to great success in NLP and, as of this year, in Computer Vision as well [141]. Attention is based on the idea that specific parts of an input have differing importance and should be attended to accordingly. Broadly, for a given input, a vector with values is output that indicate how 'important' specific parts of the input are [4]. This mechanism lends itself to reading it as a natural explanation. If attention on a particular part of the input was high, this part of the input should be important for the output. There have been several papers interpreting attention as explanation [24, 148]. On the other hand, there have also been several papers showing that interpreting attention as an explanation can be misleading [65, 99, 118].

Bastani et al. [13] proposed model extraction to relearn a much simpler and interpretable model from a blackbox model. While the original model is not interpretable, its outputs can be used as targets to train a less complex and thus more interpretable model.

3.6 Processing explanations

Having covered the need for explainability (Section 3.1) as well as the many approaches with different goals (Sections 3.3 to 3.5), we will now look at the way explanations are processed. This covers the evaluation of different explanation approaches (including follow-up papers with weaknesses of particular methods), attacks on explanation methods, as well as their use in cooperative systems between human and machine.

[4] For a more specific and all-around much better explanation of attention in general and transformers in particular, we refer the reader to the tutorials by Jay Alammar [6, 66]

Evaluation of explanation methods So far we have covered a plethora of diverse approaches to explainability. A natural question to ask is which method is best for a particular case or in general. This question has been covered by a number of works [1, 17, 61, 72, 96, 114, 127, 138, 149]. Relatively early, Kindermans et al. [72] criticized that many explanation methods are not input invariant, i.e. that the absolute value of the mean influences what explanations look like despite the neural network being trained is not influenced by a constant bias on all inputs.

Adebayo et al. [1] proposed two sanity checks for single-example explainability methods, namely parameter and data randomization. If an explainability method actually picks up on meaningful information about the classification, there should be substantial difference between an explanation from a trained neural network and a neural network with the parameters shuffled, as there is no meaningful information in the latter. Equally, a neural network can be fitted to data with the labels shuffled, effectively memorizing all the labels. Since there is no information about the classes in the labels, we'd expect the neural network to not have learned anything meaningful and again for the explanations to look very different from explanations for a trained model. Surprisingly their results showed that a popular method, GuidedBackprop is very robust to both randomizations. This implies that GuidedBackprop is in fact not conveying meaningful information about the neural network classification.

In recent findings, Sixt et al. [127] confirmed those results. They showed that a number of modified backpropagation methods, including Guided Backpropagation, LRP and Pattern Attribution, in fact do not rely on the parameters of the neural network. This finding also highlighted the fact that evaluation of explanation methods has mainly relied on visual inspection of the resulting explanations. As humans are primed to expect an explanation that aligns to our own shape-based classification while traditional neural network mainly rely on texture [40], visual inspection is clearly not sufficient for evaluation.

Samek et al. [114] proposed a quantitative approach for comparing explanation methods. To evaluate an explanation method, they iteratively replace the most important inputs (according to the explanation method) with white noise and reclassify this manipulated input. If the explanation method correctly identifies important features, the output value will fall quickly. Done iteratively, this results in a curve that can be compared for different methods. In Section 4.2 we improve on this approach by utilizing image segmentation to remove image segments at a time and using a better replacement value to keep the inputs close to the training manifold.

Hooker et al. [61] pointed out that replacing parts of the input will result in input samples that are off the data manifold that the neural network has been trained on. To resolve this issue, they retrain the neural network with the entire training set where the most important n pixels[5] are replaced with a uniform pixel value in their approach RemOve And Retrain (ROAR). The accuracy drop of the retrained neural

[5]Hooker et al. [61] only evaluated their approach on images

network indicates how accurate the explanation method is. Notably, this approach does not take redundant features into account. Due to the retraining, ROAR will evaluate what features are important for all possible models weights with the current architectures, not the one model at hand.

Fairwashing If explanation methods are used to evaluate how fair a given neural network is, a malicious entity would naturally be interested in how to disguise unfairness [80]. The same goes for intentional manipulation of the input where the attacker might also be interested in disguising their attack from explanation methods. As a result, attacks and defenses of explainability have been a small but growing subfield in the last years [4, 26, 27, 41, 57, 94, 99, 108, 128, 142].

In general, all methods assume that the neural network weights are fully known. Some works have considered attacks where the attacker can manipulate the neural network weights as well [57, 128, 142]. Aïvodji et al. [4] under perhaps the most realistic assumptions introduces *LaundryML*, an algorithm that finds a fair-looking model while approximating a discriminative blackbox model but do not consider neural networks.

Ghorbani et al. [41] showed that given the neural network weights, an input can be iteratively manipulated such that the output does not change and the input image only imperceptibly changes while a given explanation is manipulated, for example to remove all relevance in a given part of the input. Similar to regular adversarial attacks, they simply define a loss function for the given manipulation while regularizing the output and the input to stay close to the original image.

Dombrowski et al. [27] showed that, given the same basic technique, an explanation can also be manipulated to recreate a target explanation. In the paper they argue that this vulnerability is due to the large curvature of the output around the data manifold, i.e. is related to the severe over-parametrization of neural networks. By changing the ReLU non-linearities to a smooth approximation, the vulnerability is severely reduced. They also show that for this smoothed version of the neural network, explanations are identical to the ones obtained with SmoothGrad for one-layer networks.

Multiple papers have also pointed out the connection between robustness and increased interpretability [94, 108, 139]. Ilyas et al. [62] also showed that adversarial examples themselves transfer remarkably well between architectures, implying that noisy explanations may be due to the actual classification happening on grounds of features that humans do not pick up on. This would indicate that robustness is more accurately the neural networks being forced to 'look' at the same features as humans do and the increased explainability merely an effect of this.

Cooperation with humans In most cases, explanations are meant to be shown to humans. An obvious question to ask is therefore whether humans benefit from seeing explanations, be it by increased trust, more well-calibrated uncertainty on the

decision of the AI, or better cooperation with the AI [11, 12, 18, 37, 51, 67, 70, 77, 79, 80, 93, 115, 137, 153].

An often asked question is whether having explanations will increase a user's trust in a model. Narayanan et al. [93] looked at how the complexity of the explanation influenced a person's ability to verify the consistency of an explanation, unsurprisingly finding that explanation complexity makes it harder for humans to verify that an explanation makes sense.

Lai and Tan [79] showed that providing explanations will greatly raise trust in the AI when a human is making an assisted decision. However, showing explanations alone (without the classification) only slightly raised performance, indicating that showing an explanation raises trust but not necessarily competence. Indeed, Lakkaraju and Bastani [80] showed that explanations will also raise trust in the model when the explanation is intentionally misleading, indicating that humans can not in fact recognize a 'bad' explanation but will blindly trust a model more when more information is provided, regardless of whether the information is correct.

In contrast, there is evidence that explanations can help humans decide whether a neural network learned correct features in general:
This year, Degrave et al. [25] showed that, while a convolutional neural network can get high accuracy and AUC for diagnosing COVID-19, explainability maps highlighted lateral markers in the x-ray image. A skilled (or in this case even unskilled) human can recognize that certain features as commonly highlighted by an explanation are certainly not causally related to the outcome class and that a neural network has therefore learned 'bad' features and should not be deployed

Bhatt et al. [18] found in a comprehensive study that explanations are indeed used in companies to debug machine learning models. However, no company reported showing explanations to end users or using them systematically to make deployed machine learning models more transparent.

Bansal et al. [12] suggest that increasing trust may indeed be the only benefit of explanations. If the human is less competent than the AI, i.e. ideally the human would always follow the recommendation of the AI, they found that providing explanations will increase human performance. However, the performance is still lower than the AI by itself. Creating explanations that increase the performance of human and AI above each alone seems to remain an open problem, raising several exciting open questions for the future design of explanation methods.

Tomsett et al. [137] looks at the different stakeholders, i.e. engineers, practitioners and end users, and their requirements for an explanation. They then consider the possibility that different user groups might have different requirements for an explanation to be sufficient. Asking the right questions is vital for getting the right answers, making this an important consideration in aiding cooperation between humans and AI.

CHAPTER 4
Contributions

In this chapter we describe the contributions that make up this thesis. For each paper we will motivate the research question underlying the manuscript, briefly outline the contributions made by the paper and talk about later work in the same field. We then connect the contributions in this paper to subsequent work and give a broader context into the current state, implications, and use cases for the contribution. The manuscripts themselves are found in the appendix.

4.1 Interpretability in Intelligent Systems - A New Concept?

As covered in Chapter 3, there has been a lot of recent developments in the field of explainability for neural networks along with the research into neural networks themselves. While the idea of neural networks is far from new, the interest in them has been greatly revived with the deep learning renaissance in the last few years. In the same way, the demand for explainability has been around for a long time and there is a rich literature of requirements for explainability to build on.

In this paper we take a look at the history of explainability, connect past approaches to current ideas and approaches and point out where looking backwards might be useful for moving forward. We focus on two areas, desiderata for explainability and the quantification of uncertainty within importance maps.

In Section 3.1 we covered papers concerned with definitions and categorizations of explainability in recent times and mainly focusing on neural networks. In fact, there has been ample prior work on requirements for artificial intelligence. In 1983 Swartout [134] already noted that "trust in a system is developed not only by the quality of its results, but also by clear description of how they were derived". Yet, current position papers on explainability for neural networks tend to be myopic, only looking at the most recent work and issues.

In this paper we cover five desiderata for explanations of AI as originally proposed by Swartout and Moore [135]. Swartout and Moore [135] state that explanations should be (i) true to the original system, (ii) understandable, (iii) sufficient to justify the original decision, (iv) have low construction overhead and (v) be efficiently obtainable.

Various of these desiderata have in concurrent work been redefined as essential, highlighting the usefulness of a look into the past [28, 43, 58, 83].

In the same way, Lacave and Diez [78] in adaptation from results of the 1988 AAAI Workshop on Explanation [147] already proposed to classify explanations according to (i) content, (ii) communication and (iii) adaptation to the user. Notably different to concurrent work, it is seen as an important factor that the explanation is understandable to the user and can adapt to the current user's knowledge of the system. A recent line of work focuses on usability and usefulness of explanations, essentially rediscovering those requirements [12, 18, 51].

Intriguingly, explainability was seen as essential for the adoption of AI in many use cases, resulting in the development of Explainable Expert Systems (EES). An example of this is MYCIN, a medical expert system to classify bacteria causing infections [120, 121]. In contrast to the statistical machine learning that we focus on, the system used hand-coded rules and was able to display an explanation for what factors were most responsible for a particular diagnosis.

The second focus of our paper is the inclusion of uncertainty in the context of explanations. For this, we distinguish between aleatoric and epistemic uncertainty, with epistemic uncertainty also being addressed in app:aggreg. In the context of neuroimaging, quantification of uncertainty for visualizations of brain activity has been of interest for a long time. Accordingly, there are prior approaches that address this. In particular, we focus on NPAIRS as a previous approach to provide uncertainty estimates of aleatoric uncertainty via resampling techniques [132]. For epistemic uncertainty, Lange et al. [81] and Hansen et al. [49] provide ways to combine estimates from multiple models to reduce and quantify uncertainty.

In this work we looked at previous approaches to explainability in the history of artificial intelligence. We covered shared and differing requirements on explainable AI in the past and present and considered the importance of handling uncertainty also for explanations. While the deep learning renaissance is recent, the need for explainability is not and many ideas have previously been covered. We hope that this paper presents an accessible though by no means complete overview of previous work, providing inspiration for future work.

4.2 IROF: a low resource evaluation metric for explanation methods

We present a novel way to evaluate explanation methods objectively. In recent years explainability has gained far more attention, resulting in a slew of explanation methods being proposed. The majority of them are only evaluated qualitatively. Recent literature has covered this issue [1, 127]. It points to a much deeper problem with a lot of explainability research, the lack of a clearly stated objective. The need

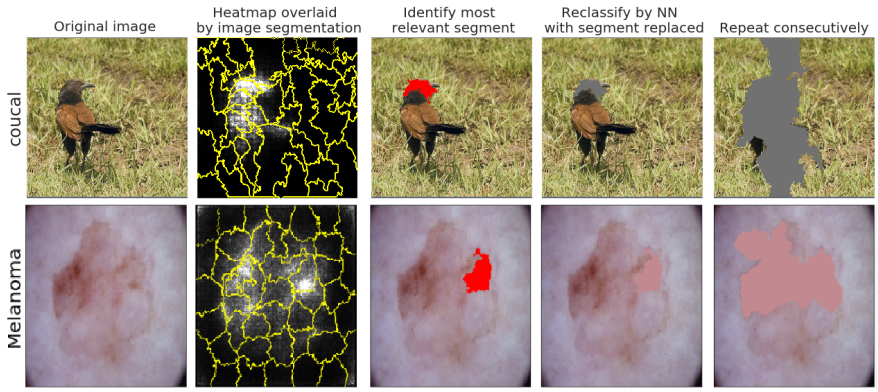


Figure 4.1: Figure from Appendix B. Visualization of IROF approach. Each image in the evaluation set is first divided into coherent segments. For each explanation method, segments judged as important get removed iteratively.

for explainability is intuitively understandable and many people would hesitate to deploy a black-box algorithm. Yet, the questions "What is explainability?" and "What constitutes a good explanation?" is difficult to answer. Exacerbating this problem, few of the current explainability methods build on or systematically compare to previous work. As there is no agreed upon standard on how to compare or evaluate explanation methods, this however is understandable.

Many papers show visualizations of the explanations they produce, particularly for convolutional neural networks, as evidence of the validity of their methods. As laid out in previous work in Chapter 3, qualitative evaluation by humans is not a good measure of accuracy, since humans are bad at distinguishing good from bad explanations. Following Gilpin et al. [43] we therefore want to measure quantitatively how complete the explanation is, i.e. how many important features are captured. The other dimension, understandability by humans, remains an open problem.

The research question behind this paper is "how can we compare two or more explanation methods without human bias?". Since the goal is to measure completeness, intuitively we want to quantify how well the explanation method identifies parts of the input that were important for classification. Many explanation methods for neural networks return an explanation with the same dimensions as the input to the neural network. Particularly for images, where there is high redundancy in information in neighboring pixels, this presents a challenge. In this paper we present two contributions. We propose a new way of quantifying explanation methods in terms of completeness that does not rely on human evaluation. Using this approach, Iterative Removal Of Features (IROF) we evaluate several common explanation methods and summarize the results.

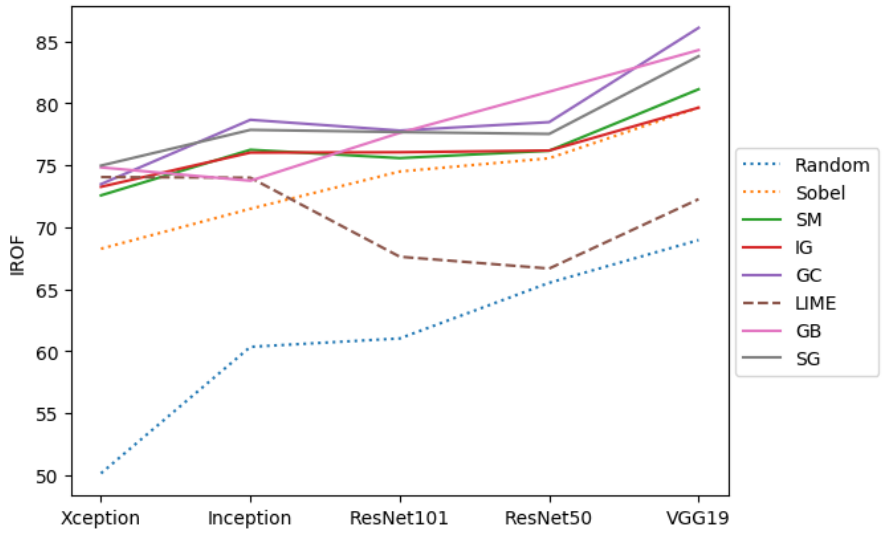


Figure 4.2: Figure from Appendix B. Performance of common explanation methods measured with IROF on common pretrained architectures. The ranking between architectures stays consistent. Architectures ordered after decreasing accuracy. Accuracy and explainability are inversely correlated.

In summary, IROF as shown in Figure 4.1 first segments each image in the test set into super-pixels. Comparing the importance according to an explanation method, important super-pixels iteratively get removed from the image. Measuring the area under the curve for the degrading output score gives one comparable measure for each explanation method. If an explanation method works well, the output score will decrease fast as information is removed, enabling us to compare two or more methods directly. For an exact description of the method we refer to the paper. Using IROF, we found evidence that explainability is dependent on the network architecture itself and may be inversely related to accuracy. This implies that for tasks where explainability is essential, choosing a less accurate but more explainable network architecture may be advisable. We also found that the ranking of explanation methods stayed remarkably constant between architectures trained for the same task but differed between tasks. This implies that there may not be one universally best method but that different explanations may be most suitable for different tasks.

Since this paper first came out, evaluation of explainability has gained more interest as evidence mounted that human evaluation is not enough [17, 87, 96, 127, 145]. In particular, Bhatt et al. [17] extends on previous desirables for explanations by also giving quantitative measures for low sensitivity and low complexity.

There has also been work that show that some backpropagation-based explanations

do not change substantially when randomizing weights of the neural network [127]. In combination with our work, this underlines the importance of dataset choice when evaluating explanation methods. It also emphasizes that we need different datasets for evaluating explanation methods in particular:
Imagenet, the most common benchmark for evaluating convolutional neural networks, consists of non-overlapping classes [68]. In most cases the object to be classified is the clear focus of the image. This results in a Sobel edge filter or a randomized convolutional neural network (which functions as an edge detector) being a very strong baseline with IROF (see Figure 4.2), as it also correctly identifies important parts of the image. Since an explanation does not need to 'decide' between several salient objects, the dataset makes explanations very easy and consequently distinguishing between two explanation methods harder.

To conclude, we presented a computationally cheap approach for evaluating explanation methods for a given specific task. New developments point to evidence that current datasets may not be sufficient for accurately evaluating explanations, as they are too easy to explain even without knowledge of the correct class, e.g. by an edge detector.

In order to accurately evaluate explanation methods as to whether they pick up semantic differences, we may need to assemble more challenging datasets that contain multiple and misleading objects. This is left to future research.

4.3 A simple defense against adversarial attacks on heatmap explanations

As covered in Chapter 3, there is a multitude of explanation methods already available [133]. Following different underlying philosophies, some methods treat the model as a black box and determine the most important parts of the input by sampling the same input with some input dimensions removed [101], backpropagate the output onto the input or use game theory to determine the attributions. Particularly when the explanations are intended for practitioners without knowledge of how machine learning algorithms work, fostering not only trust but also mistrust is important. Concretely, an explanation should enable a practitioner with domain knowledge to recognize a wrong decision.

Naturally, the question arises of which method to use for a specific problem. In Appendix B we introduced a way to measure the performance of explanation methods and determine which method is right for a specific task. However, a more elegant way to solve this issue is to circumvent this choice altogether. Combining multiple weak estimators to form one better estimator is a well-explored strategy in machine learning, motivating our use of it for explanations.

Additionally, attacking explanations, i.e. 'fairwashing' is becoming increasingly of interest as explainability is becoming more main-stream. Multiple works have found

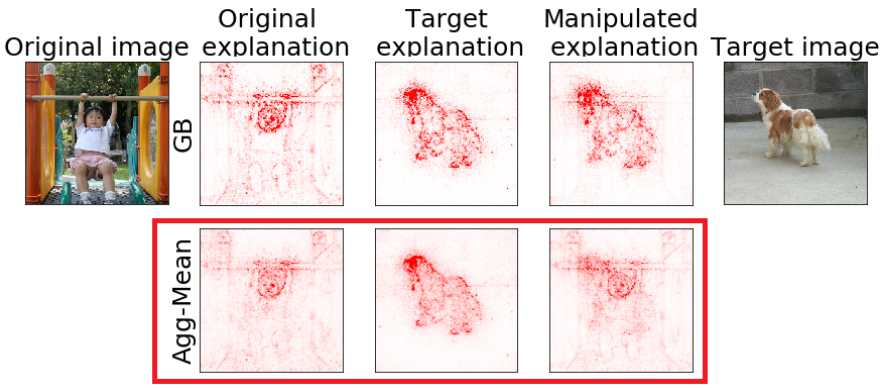


Figure 4.3: Figure from Appendix C. Common explanation techniques (Guided Backpropagation in the upper row) are vulnerable to adversarial attacks and can be manipulated to look like any given target explanation map without visible changes in the input or output. Aggregations (lower row) are much more robust. In the example, the original explanation is still clearly recognizable.

that explanations are as vulnerable to adversarial manipulation as classifications, making it easy to manipulate many well-known methods into showing arbitrarily changed explanations without perceptible change of the input or the output [41, 128]. A malicious attacker might have motivation to hide that an input was manipulated in an adversarial way for a different output. The deployer of an algorithm might have motivation to hide discriminative reasoning.

In this paper we explore and evaluate this approach of aggregating multiple methods for explainability for increased stability. We assume that there exists a 'ground-truth' explanation, i.e. an explanation that perfectly explains the neural network decision, and that different explanation methods are imperfect estimators of this ground-truth. By combining multiple imperfect methods the resulting aggregate comes closer to the 'perfect' explanation than either method on its own.

Attacks on explanations, similar to traditional attacks on the output of an neural network can be executed by iteratively updating the input with a loss defined to change the output maximally with minimal input change [136]. In our article, we show that using multiple explanation methods is an effective defense against adversarial attacks: While single methods can easily be manipulated to be virtually indistinguishable from the target explanation, an aggregated explanation stays robust. This holds, even when the attacker knows the exact methods that will be used by the defender and has complete knowledge of the neural network being used.

Our contribution in this paper is two-fold. We show theoretically that averaging over

multiple explanations will get an error lower than the average error of the contributing explanation methods. Empirically we also give evidence that an aggregate will also outperform the best single explanation method in most cases. Second, we show that an aggregate is remarkably more robust to adversarial manipulation as visible in Figure 4.3 and measured quantitatively via distance to the original explanation. In this paper we use our defense against the attack introduced in Dombrowski et al. [27]. In experiments we also see that the defense holds against a different attack as proposed by Ghorbani et al. [41]. In both cases, the attacker has full knowledge of the network weights and full control of the input (restricted by the input range of the neural network under the assumption that inputs outside of the data range would be checked for in a realistic use-case). In all cases, our approach of aggregating explanation methods into an ensemble far outperforms any single method in robustness.

Since our paper came out, Bhatt et al. [17] in an approach very similar to ours also proposed aggregating explanations for reduced sensitivity and reduced complexity. Given a criterion for the optimal explanation to optimize, they show that an optimal convex combination of explanation methods will be at most as sensitive as either of the explanation methods. Based on this, they propose an alternative to SHAP [86] that aggregates the SHAP explanations of the nearest neighbors in data space. Manupriya et al. [87] also explored non-linearly aggregating explanation methods with sub-modular functions.

Zhang et al. [151] has recently shown that output and explanation can also be manipulated simultaneously. This enables the attacker to change the output decision without it being detectable, making the problem of developing good defenses against adversarial attacks more urgent. As research has already shown that adversarial attacks or more generally malicious manipulation on the output of a neural network can be executed in real life, recognizing, defending against and conversely from the other side obscuring those attacks remains a very open research field [36]. There is ample motivation to obscure the 'real' reasons behind a neural networks decision, either as the deployer of the algorithm or as the entity supplying the input.

While adversarial attacks on the output of a neural network are a concern for fully autonomous agents such as in self-driving cars, adversarial attacks on explanations of neural networks are unsurprisingly more of a concern for high-impact decisions where the explanation is expected to be examined. This could be the case with applications where a human is using the explanation for further information, e.g. in medical care, or applications where a single decision or the neural network in general is inspected for discriminative bias.

Making neural networks robust to adversarial manipulation for either output or explanation is still an open research question [116]. In this work we show that a simple aggregation may be an effective way to ward against naive attacks aiming to manipulate what features are shown to be important for the decision.

4.4 Interpretations are useful: penalizing explanations to align neural networks with prior knowledge

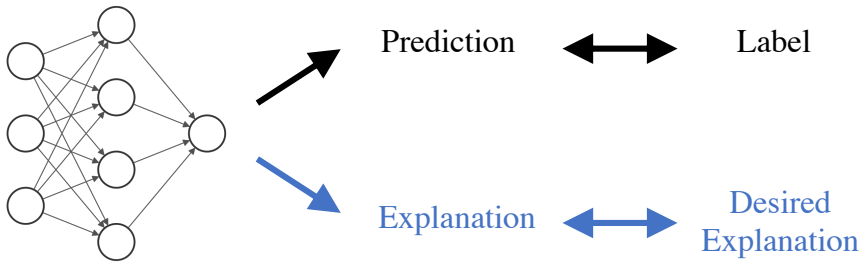


Figure 4.4: Figure from Appendix D. Traditionally, neural networks are trained with labeled data. With our method, it is also possible to train a neural network with explanations to infuse prior knowledge.

Being able to know *why* a particular classification was made is extremely useful for the reasons outlined before. In some cases, knowing the underlying reasons may then lead us to discover that the neural network has a systemic bias encoded. Particularly with large datasets as used for training neural network, it is not possible to analyze the entire dataset for the underlying patterns. Ideally, explanations can also help us to discover unwanted correlations [92]. With real life datasets that reflect the current reality, they will also reflect unwanted aspects. In some cases we may also only be able to collect a selected part of the data, i.e. the data is not independent and identically distributed.

A popular machine learning urban legend tells of the US Army training an neural network to classify their own versus enemy tanks [20]. While the neural network achieved almost perfect accuracy on the training set, it also made puzzling mistakes. Later it turned out that all pictures of enemy tanks had been taken during gloomy weather whereas all pictures of ally tanks had been taken during sunny weather. This resulted in the neural network focusing only on the weather conditions as opposed to details on the tank. While the story is likely embellished or not true, we know of many cases where a neural network picked up on such unwanted correlations:

In an innocuous example, the popular PASCAL VOC image dataset includes a category 'horse' for which a trained neural network can get high accuracy [35]. However, by inspecting explanations for images of the class we see that the neural network does not focus on the horse itself but rather on an area in the bottom of the image [82]. Many of the images for 'horse' were scraped from the internet and contain the source tag of the author in the image, a bias in the dataset that would have gone unnoticed without explainability. We can also find many such examples in the medical domain:

In diagnosing pneumonia, Zech et al. [150] found that a trained neural network can predict the hospital the x-ray was taken in with high accuracy due to local differences in how x-rays are taken. Thus they can learn to exploit site-individual incidence rates and will be biased against patients from a particular hospital. In another very concurrent example, Degrave et al. [25] found that while convolutional neural networks can reach high accuracy in the diagnosis of COVID-19 from chest x-rays, they also rely on confounders such as laterality markers and will drastically drop in accuracy otherwise. For skin cancer detection, a unique challenge for medical machine learning in that large-scale datasets are available, Bissoto et al. [19] identified seven potential confounders in a popular dataset, among them rulers left next only to malignant skin lesions (when the doctor already knew that a skin lesion was malignant, they would often use a ruler to have an indicator of size for subsequent treatment decisions). Neural networks trained on this dataset will learn to recognize and pick up on these confounders [19].

The harm these kind of exploits can cause if left unexamined and uncorrected is obvious. However, while explainability methods enable us to find out about those unwanted learned relationships, there was no remedy to the bias. The practitioner was left with the choice to accept this bias or to curate a new dataset, retrain the model and evaluate again [92]. Our work remedies this by enabling practitioners to specify prior knowledge that is not present in the data but should be learned regardless. Our main contribution is Contextual Decomposition Explanation Penalization (CDEP), a method to train neural networks to rely on the right reasons via explanations as visualized in Figure 4.4).

We assume that we have access to prior knowledge and wish to incorporate this prior knowledge into the neural network even though it is not expressed in the training data. As it is rare to know how important a particular feature should be exactly, we focus on cases with a known spurious bias, i.e. a correlation in the training data that should not be reflected in the trained neural networks. In this case we do know exactly how important the feature should be - not at all. This bias could exist because we were only able to collect training data in certain circumstances, was collected before it was known to be used for deep learning, is expensive to obtain or reflects a discriminative bias that we explicitly do not want to continue.

In the paper we demonstrate the effectiveness of CDEP in removing bias on toy datasets as well as on real-life datasets. In Figure 4.5 we show an example of a skin cancer dataset where many of the non-cancer samples had a bright patch in the image. While the vanilla neural network learned to rely on the patch, the neural network trained with CDEP learned to ignore it. During our experiments we also found that the neural network learned to rely on rulers, marking them as relevant in explanations. This further highlights the ability of neural networks to pick up on correlations that humans would not find in the dataset, underlining the need for approaches to address this.

An often asked question is whether CDEP will work when prior information is not

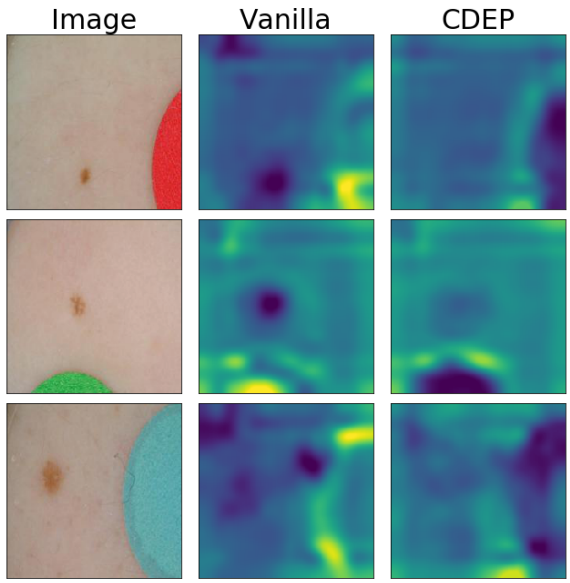


Figure 4.5: Figure from Appendix D. The 'normal' network learned to detect bright patches. This is an expected but undesirable effect given the training data. With CDEP regularizing this feature, the patches are less important and the network focuses on skin.

available. Since it works by incorporating prior knowledge into the neural network, having access to prior knowledge is absolutely essential. We show this with experiments on the COMPAS dataset: With the COMPAS dataset, the goal is to predict re-offenders based on a number of attributes like age, gender and race. Conventional training will result in a model that is biased against black people as measured by the wrongful conviction rate. We show that CDEP can make the model more fair with the right explanations but conversely also more unfair with different explanations. This highlights not only that it is vital to incorporate the right knowledge into the model but also the need to ask the right questions when using explanation methods to check your model for unintended bias.

Since our paper was published, Weinberger et al. [146] proposed a new approach to regularize a neural network when only uncertain prior information is known. By jointly learning the neural network and a model that emphasizes features from a set of potential features, they circumvent needing to know exactly what features should be emphasized. However, introducing a prior in that way may suppress features that are important for only a small proportion of the data distribution. In another similar approach, Etmann et al. [34] proposed regularizing the model to be sparse for more robust and interpretable tumor type classification.

The basic idea to use gradients to regularize a neural network is not new. In fact, to our knowledge, the first paper that proposes this idea is 'Improving Generalization Performance Using Double Backpropagation.' from 1992 [30]. However, the idea to explicitly enforce prior knowledge, not general regularization, in the model is relatively recent [110]. As deep learning is adopted for more specialized tasks with small and biased datasets and more focus is put on the potential harm of uncorrected bias within machine learning systems, we expect the idea to gain popularity.

Our paper proposes a way to use biased datasets without having this bias reflected in the trained model. We hope that this will ease the adoption of machine learning particularly in medical decision support systems where such bias is common [19, 25].

CHAPTER 5

Discussion and Conclusion

In this chapter we summarize the work presented, reflect on the Ph.D. studies as well as general progress in the topic and give an outlook for potential future work.

5.1 Summary

In this thesis we have introduced the reader to and tied in major parts of three years of research.

In Chapter 1 we laid out the urgent need for explainability in practice and contrasted it with the current state in deployment. After covering base concepts of machine learning and deep learning in particular in Chapter 2, we introduced concurrent relevant work in sub-fields of and adjacent to explainability in Chapter 3. In Chapter 4 we presented our own contributions and connect them to the thesis topic. In all the contributions in this thesis, the common theme is the focus on how explainability can be made easier to use or more useful for practitioners.

Appendix A took a look at previous research on explainability. In particular, we introduced earlier approaches to make intelligent systems explainable and noted the strong focus on interactive communication to the receiver of the explanation that was present in earlier work.

Appendix B introduced IROF, a method to comparatively evaluate two or more explanation methods based on how well they identify parts of the input important for the neural network decision. This enables a practitioner to identify the explanation method best suited for a particular task.

In Appendix C we showed that this choice can in fact be circumvented in some cases, as an aggregation of multiple explanation methods tends to be more accurate than any single explanation method. In addition, it is much more robust against adversarial attacks, which single methods are very susceptible to.

Finally, in Appendix D we showed that explanations can not only be used to gain new knowledge from a neural network but also to infuse prior knowledge into the

neural network. In addition to the original reasons for explainability that we covered in Chapter 1, this also constitutes a novel motivation for explainability.

5.2 Reflection and Outlook

Looking back on three years of development in the field of explainability, including our own work, it seems clear that explainability hinges at least as much on the recipient, i.e. the human than the neural network. In Section 3.1 we covered Gilpin et al. [43] with the view that explanations should be judged on two axes: their completeness and their comprehensibility for humans. The most complete explanation for a model is the model itself, however this explanation is not comprehensible. As we reduce the complexity of the explanation, the explanation should become more understandable. We think that classifying the goal of a paper on explainability according to this schema would be useful to better evaluate the current progress in explainability and the marginal benefit that a new approach will bring to the field.

One aspect of explainability that we only briefly touched on is the connection to causality when using explanations to discover causal effects. In general, neural networks as we cover them in this thesis do not attempt to build a causal model but build a statistical model of correlations. However, counterfactual explanations as covered in Section 3.4 still build or assume a causal model by answering "What would need to change in the input for the output to change?" Extracting or connecting the knowledge learnt by a neural network with a causal graph is an exciting future research direction for explainability but also for more grounded neural networks by themselves.

Originally, the topic of this thesis was explainability of uncertainty. The main motivation for this at the start of my Ph.D. was to help along the adoption of machine learning in critical applications. During the course of the Ph.D. studies, this was slowly put out of focus as it became clear that even explainability of point estimates was currently not used in applications. In Chapter 1 we outlined this gap between academic research and practice. Contributing to closing this gap is likely to have more of a positive impact in the future adoption of machine learning for critical applications than the original goal, leading to the refocus of these Ph.D. studies.

5.3 Conclusion

In these Ph.D. studies we have tackled explainability from different angles. With a refocus from the original research question, we have achieved the underlying goal of developing methods to make explainability more accessible and useful for practitioners.

While there has been a lot of progress in the field in the past three years, many open questions remain. In the coming years, we envision that the questions that we want to answer with explainability will be more targeted, with explanations becoming

more specific to specific stakeholders and usability being a strong focus of subsequent work. In the same line, we envision more interactive explanations with the option to adjust faithfulness and complexity according to requirements rather than explanation methods being fixed in those two dimensions.

Bibliography

[1] Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. Sanity checks for saliency maps. In *Advances in Neural Information Processing Systems*, volume 2018-Decem, pages 9505–9515, 2018.

[2] Tameem Adel, Zoubin Ghahramani, and Adrian Wcller. Discovering interpretable representations for both deep generative and discriminative models. In *35th International Conference on Machine Learning, ICML 2018*, volume 1, pages 86–101, 2018. ISBN 9781510867963.

[3] Chirag Agarwal, Peijie Chen, and Anh Nguyen. Intriguing generalization and simplicity of adversarially trained neural networks. Technical report, jun 2020.

[4] Ulrich Aïvodji, Hiromi Arai, Olivier Fortineau, Sébastien Gambs, Satoshi Hara, and Alain Tapp. Fairwashing: The risk of rationalization. In *36th International Conference on Machine Learning, ICML 2019*, volume 2019-June, pages 240–252, jan 2019. ISBN 9781510886988.

[5] Guillaume Alain and Yoshua Bengio. Understanding intermediate layers using linear classifier probes. Technical report, oct 2016.

[6] Jay Alammar. The Illustrated Transformer, 2019. URL http://jalammar.github.io/illustrated-transformer/.

[7] David Alvarez-Melis and Tommi S. Jaakkola. Towards robust interpretability with self-explaining neural networks. Technical report, 2018.

[8] Javier Antorán, Umang Bhatt, Tameem Adel, Adrian Weller, and José Miguel Hernández-Lobato. Getting a CLUE: A Method for Explaining Uncertainty Estimates. Technical report, 2020.

[9] Leila Arras, Grégoire Montavon, Klaus-Robert Müller, and Wojciech Samek. Explaining Recurrent Neural Network Predictions in Sentiment Analysis. In *Proceedings of the 8th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 159–168. Association for Computational Linguistics (ACL), jun 2017. doi: 10.18653/v1/w17-5221.

[10] Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus Robert Müller, and Wojciech Samek. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PLoS ONE*, 10(7), 2015. ISSN 19326203. doi: 10.1371/journal.pone.0130140.

[11] Gagan Bansal, Besmira Nushi, Ece Kamar, Daniel S. Weld, Walter S. Lasecki, and Eric Horvitz. Updates in human-ai teams: Understanding and addressing the performance/compatibility tradeoff. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):2429–2437, jul 2019. ISSN 2159-5399. doi: 10.1609/aaai.v33i01.33012429.

[12] Gagan Bansal, Tongshuang Wu, Joyce Zhou, Raymond Fok, Besmira Nushi, Ece Kamar, Marco Tulio Ribeiro, and Daniel S. Weld. Does the Whole Exceed its Parts? The Effect of AI Explanations on Complementary Team Performance. Technical report, jun 2020. URL http://arxiv.org/abs/2006.14779.

[13] Osbert Bastani, Carolyn Kim, and Hamsa Bastani. Interpretability via Model Extraction. Technical report, 2017.

[14] David Bau, Bolei Zhou, Aditya Khosla, Aude Oliva, and Antonio Torralba. Network dissection: Quantifying interpretability of deep visual representations. In *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, volume 2017-Janua, pages 3319–3327, apr 2017. ISBN 9781538604571. doi: 10.1109/CVPR.2017.354.

[15] Michel Besserve, Arash Mehrjou, Rémy Sun, and Bernhard Schölkopf. Counterfactuals uncover the modular structure of deep generative models. In *International Conference on Learning Representations*, dec 2019. URL http://arxiv.org/abs/1812.03253.

[16] Umang Bhatt, McKane Andrus, Adrian Weller, and Alice Xiang. Machine Learning Explainability for External Stakeholders. Technical report, jul 2020.

[17] Umang Bhatt, Adrian Weller, and José M. F. Moura. Evaluating and Aggregating Feature-based Model Explanations. Technical report, 2020.

[18] Umang Bhatt, Alice Xiang, Shubham Sharma, Adrian Weller, Ankur Taly, Yunhan Jia, Joydeep Ghosh, Ruchir Puri, José M.F. Moura, and Peter Eckersley. Explainable machine learning in deployment. In *FAT* 2020 - Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 648–657, sep 2020. ISBN 9781450369367. doi: 10.1145/3351095.3375624.

[19] Alceu Bissoto, Eduardo Valle, and Sandra Avila. Debiasing skin lesion datasets and models? not so fast. Technical report, 2020.

[20] Gwern Branwen. The Neural Net Tank Urban Legend, 2011.

[21] Shan Carter, Zan Armstrong, Ludwig Schubert, Ian Johnson, and Chris Olah. Activation Atlas. *Distill*, 4(3):e15, mar 2019. ISSN 2476-0757. doi: 10.23915/distill.00015.

[22] Chun Hao Chang, Elliot Creager, Anna Goldenberg, and David Duvenaud. Explaining image classifiers by counterfactual generation. In *7th International Conference on Learning Representations, ICLR 2019*, number 2017, pages 1–15, jul 2019.

[23] Aditya Chattopadhyay, Piyushi Manupriya, Anirban Sarkar, and Vineeth N. Balasubramanian. Neural network attributions: A causal perspective. In *36th International Conference on Machine Learning, ICML 2019*, volume 2019-June, pages 1660–1676, feb 2019. ISBN 9781510886988.

[24] Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D. Manning. What Does BERT Look At? An Analysis of BERT's Attention. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 276–286. Association for Computational Linguistics (ACL), jun 2019.

[25] Alex J Degrave, Joseph D Janizek, Su-In Lee, and Paul G Allen. AI for radiographic COVID-19 detection selects shortcuts over signal. *medRxiv*, page 2020.09.13.20193565, sep 2020. doi: 10.1101/2020.09.13.20193565.

[26] Botty Dimanov, Umang Bhatt, Mateja Jamnik, and Adrian Weller. You shouldn't trust me: Learning models which conceal unfairness from multiple explanation methods. Technical report, 2020.

[27] Ann Kathrin Dombrowski, Maximilian Alber, Christopher J. Anders, Marcel Ackermann, Klaus Robert Müller, and Pan Kessel. Explanations can be manipulated and geometry is to blame. In *Advances in Neural Information Processing Systems*, volume 32, pages 1–34, 2019.

[28] Finale Doshi-Velez and Been Kim. Towards A Rigorous Science of Interpretable Machine Learning. Technical report, feb 2017.

[29] Finale Doshi-Velez, Mason Kortz, Ryan Budish, Christopher Bavitz, Samuel J. Gershman, David O'Brien, Stuart Shieber, Jim Waldo, David Weinberger, and Alexandra Wood. Accountability of AI Under the Law: The Role of Explanation. Technical report, nov 2017.

[30] Harris Drucker and Yann Le Cun. Improving Generalization Performance Using Double Backpropagation. *IEEE Transactions on Neural Networks*, 3(6):991–997, 1992. ISSN 19410093. doi: 10.1109/72.165600.

[31] John Duchi JDUCHI and Yoram Singer. Adaptive Subgradient Methods for Online Learning and Stochastic Optimization. Technical report, 2011.

[32] Lilian Edwards and Michael Veale. *Slave to the Algorithm? Why a Right to Explanationn is Probably Not the Remedy You are Looking for*, volume 2017. 2017. ISBN 3540445668. doi: 10.2139/ssrn.2972855.

[33] Andre Esteva, Brett Kuprel, Roberto A. Novoa, Justin Ko, Susan M. Swetter, Helen M. Blau, and Sebastian Thrun. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, 542(7639):115–118, 2017. ISSN 14764687. doi: 10.1038/nature21056. URL https://licensing.eri.ed.ac.uk/i/.

[34] Christian Etmann, Maximilian Schmidt, Jens Behrmann, Tobias Boskamp, Lena Hauberg-Lotte, Annette Peter, Rita Casadonte, Jörg Kriegsmann, and Peter Maass. Deep Relevance Regularization: Interpretable and Robust Tumor Typing of Imaging Mass Spectrometry Data. Technical report, 2019.

[35] Mark Everingham, Luc Van Gool, Christopher K.I. Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (VOC) challenge. *International Journal of Computer Vision*, 88(2):303–338, jun 2010. ISSN 09205691. doi: 10.1007/s11263-009-0275-4.

[36] Kevin Eykholt, Ivan Evtimov, Earlence Fernandes, Bo Li, Amir Rahmati, Chaowei Xiao, Atul Prakash, Tadayoshi Kohno, and Dawn Song. Robust Physical-World Attacks on Deep Learning Models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1625–1634, 2018. URL http://arxiv.org/abs/1707.08945.

[37] Shi Feng and Jordan Boyd-Graber. What can Ai do for me? Evaluating machine learning interpretations in cooperative play. In *International Conference on Intelligent User Interfaces, Proceedings IUI*, volume Part F1476, pages 229–239. Association for Computing Machinery, 2019. ISBN 9781450362726. doi: 10.1145/3301275.3302265.

[38] Ruth C. Fong and Andrea Vedaldi. Interpretable Explanations of Black Boxes by Meaningful Perturbation. In *Proceedings of the IEEE International Conference on Computer Vision*, volume 2017-Octob, pages 3449–3457, apr 2017. ISBN 9781538610329. doi: 10.1109/ICCV.2017.371.

[39] Yarin Gal and Zoubin Ghahramani. Dropout as a Bayesian Approximation: Appendix. In *33rd International Conference on Machine Learning, ICML 2016*, volume 3, pages 1661–1680, 2016. ISBN 9781510829008.

[40] Robert Geirhos, Claudio Michaelis, Felix A. Wichmann, Patricia Rubisch, Matthias Bethge, and Wieland Brendel. Imagenet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. In *7th International Conference on Learning Representations, ICLR 2019*, 2019.

[41] Amirata Ghorbani, Abubakar Abid, and James Zou. Interpretation of neural networks is fragile. In *33rd AAAI Conference on Artificial Intelligence*, volume 33, pages 3681–3688. Association for the Advancement of Artificial Intelligence (AAAI), oct 2019. ISBN 9781577358091. doi: 10.1609/aaai.v33i01.33013681.

[42] Amirata Ghorbani, James Wexler, James Zou, and Been Kim. Towards automatic concept-based explanations. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

[43] Leilani H. Gilpin, David Bau, Ben Z. Yuan, Ayesha Bajwa, Michael Specter, and Lalana Kagal. Explaining explanations: An overview of interpretability of machine learning. In *Proceedings - 2018 IEEE 5th International Conference on Data Science and Advanced Analytics, DSAA 2018*, pages 80–89, 2019. ISBN 9781538650905. doi: 10.1109/DSAA.2018.00018.

[44] Lore Goetschalckx, Alex Andonian, Aude Oliva, and Phillip Isola. GANalyze: Toward visual definitions of cognitive image properties. In *Proceedings of the IEEE International Conference on Computer Vision*, volume 2019-Octob, pages 5743–5752, 2019. ISBN 9781728148038. doi: 10.1109/ICCV.2019.00584.

[45] Bryce Goodman and Seth Flaxman. European union regulations on algorithmic decision making and a "right to explanation". *AI Magazine*, 38(3):50–57, sep 2017. ISSN 07384602. doi: 10.1609/aimag.v38i3.2741.

[46] Yash Goyal, Ziyan Wu, Jan Ernst, Dhruv Batra, Devi Parikh, and Stefan Lee. Counterfactual visual explanations. In *36th International Conference on Machine Learning, ICML 2019*, volume 2019-June, pages 4254–4262, apr 2019. ISBN 9781510886988.

[47] Arushi Gupta and Sanjeev Arora. A Simple Saliency Method That Passes the Sanity Checks. Technical report, may 2019.

[48] Lars Kai Hansen and Laura Rieger. Interpretability in Intelligent Systems - A New Concept? In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 11700 LNCS, pages 41–49. Springer, Cham, 2019.

[49] Lars Kai Hansen, Finn Årup Nielsen, Stephen C. Strother, and Nicholas Lange. Consensus inference in neuroimaging. *NeuroImage*, 13(6):1212–1218, jun 2001. ISSN 10538119. doi: 10.1006/nimg.2000.0718.

[50] Moritz Hardt, Eric Price, and Nathan Srebro. Equality of opportunity in supervised learning. In *Advances in Neural Information Processing Systems*, pages 3323–3331, 2016.

[51] Peter Hase and Mohit Bansal. Evaluating Explainable AI: Which Algorithmic Explanations Help Users Predict Model Behavior? Technical report, may 2020.

[52] Stefan Haufe, Frank Meinecke, Kai Görgen, Sven Dähne, John Dylan Haynes, Benjamin Blankertz, and Felix Bießmann. On the interpretation of weight vectors of linear models in multivariate neuroimaging. *NeuroImage*, 87:96–110, feb 2014. ISSN 10959572. doi: 10.1016/j.neuroimage.2013.10.067.

[53] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification. In *Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV)*, pages 1026–1034, 2015.

[54] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 2016-Decem, pages 770–778, 2016. ISBN 9781467388504. doi: 10.1109/CVPR.2016.90.

[55] Lisa Anne Hendricks, Ronghang Hu, Trevor Darrell, and Zeynep Akata. Grounding Visual Explanations, 2018. ISSN 16113349.

[56] Lisa Anne Hendricks, Ronghang Hu, Trevor Darrell, and Zeynep Akata. Generating Counterfactual Explanations with Natural Language. In *ICML Workshop on Human Interpretability in Machine Learning*, pages 95—-98, 2018. ISBN 1806.09809v1.

[57] Juyeon Heo, Sunghwan Joo, and Taesup Moon. Fooling neural network interpretations via adversarial model manipulation. In *Advances in Neural Information Processing Systems*, volume 32, feb 2019.

[58] Bernease Herman. The Promise and Peril of Human Evaluation for Model Interpretability. Technical report, nov 2017.

[59] Geoffrey Hinton, Nitish Srivastava, and Kevin Swersky. Neural Networks for Machine Learning Lecture 6a Overview of miniÂŋ-batch gradient descent. Technical report.

[60] Sepp Hochreiter and Jürgen Schmidhuber. Long Short-Term Memory. *Neural Computation*, 9(8):1735–1780, nov 1997. ISSN 08997667. doi: 10.1162/neco.1997.9.8.1735.

[61] Sara Hooker, Dumitru Erhan, Pieter Jan Kindermans, and Been Kim. A benchmark for interpretability methods in deep neural networks. *Advances in Neural Information Processing Systems*, 32, jun 2019. ISSN 10495258.

[62] Andrew Ilyas, Shibani Santurkar, Dimitris Tsipras, Logan Engstrom, Brandon Tran, and Aleksander M ËŻ Adry Mit. Adversarial Examples are not Bugs, they are Features. In *Advances in Neural Information Processing Systems.*, pages 125–136, 2019.

[63] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *32nd International Conference on Machine Learning, ICML 2015*, volume 1, pages 448–456. International Machine Learning Society (IMLS), feb 2015. ISBN 9781510810587.

[64] Ali Jahanian, Lucy Chai, and Phillip Isola. On the "steerability" of generative adversarial networks. In *International Conference on Learning Representations*, jul 2019. URL http://arxiv.org/abs/1907.07171.

[65] Sarthak Jain and Byron C. Wallace. Attention is not Explanation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, feb 2019. ISBN 1902.10186v3.

[66] Jay Alammar. Visualizing A Neural Machine Translation Model (Mechanics of Seq2seq Models With Attention). URL https://jalammar.github.io/visualizing-neural-machine-translation-mechanics-of-seq2seq-models-with-attention/.

[67] Shagun Jhaver, Amy Bruckman, and Eric Gilbert. Does transparency in moderation really matter?: User behavior after content removal explanations on reddit. In *Proceedings of the ACM on Human-Computer Interaction*, volume 3, pages 1–27. Association for Computing Machinery, nov 2019. doi: 10.1145/3359252.

[68] Jia Deng, Wei Dong, R. Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *CVPR09*, pages 248–255, 2009. doi: 10.1109/cvprw.2009.5206848.

[69] Jordan Weismann. Amazon's AI hiring tool discriminated against women., 2018. URL https://slate.com/business/2018/10/amazon-artificial-intelligence-hiring-discrimination-women.html.

[70] Harmanpreet Kaur, Harsha Nori, Samuel Jenkins, Rich Caruana, Hanna Wallach, and Jennifer Wortman Vaughan. Interpreting Interpretability: Understanding Data Scientists' Use of Interpretability Tools for Machine Learning. In *Conference on Human Factors in Computing Systems - Proceedings*, pages 1–14, New York, NY, USA, 2020. ACM. ISBN 9781450367080. doi: 10.1145/3313831.3376219.

[71] Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, and Rory Sayres. Interpretability beyond feature attribution: Quantitative Testing with Concept Activation Vectors (TCAV). *35th International Conference on Machine Learning, ICML 2018*, 6:4186–4195, nov 2018.

[72] Pieter Jan Kindermans, Sara Hooker, Julius Adebayo, Maximilian Alber, Kristof T. Schütt, Sven Dähne, Dumitru Erhan, and Been Kim. The (Un)reliability of Saliency Methods. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in*

Bioinformatics), 11700 LNCS:267–280, 2019. ISSN 16113349. doi: 10.1007/978-3-030-28954-6_14.

[73] Diederik P. Kingma and Jimmy Lei Ba. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings*. International Conference on Learning Representations, ICLR, dec 2015.

[74] Lauren Kirchner, Surya Mattu, Jeff Larson, and Julia Angwin. Machine Bias, 2016. URL https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

[75] Pang Wei Koh and Percy Liang. Understanding black-box predictions via influence functions. In *34th International Conference on Machine Learning, ICML 2017*, volume 4, pages 2976–2987, 2017. ISBN 9781510855144.

[76] Pang Wei Koh, Thao Nguyen, Yew Siang Tang, Stephen Mussmann, Emma Pierson, Been Kim, and Percy Liang. Concept Bottleneck Models. Technical report, jul 2020. URL http://arxiv.org/abs/2007.04612.

[77] Johannes Kunkel, Tim Donkers, Lisa Michael, Catalin Mihai Barbu, and Jürgen Ziegler. Let me explain: Impact of personal and impersonal explanations on trust in recommender systems. In *Conference on Human Factors in Computing Systems - Proceedings*, page 12, New York, New York, USA, may 2019. ACM Press. ISBN 9781450359702. doi: 10.1145/3290605.3300717.

[78] Carmen Lacave and Francisco Diez. A Review of Explanation Methods for Bayesian networks. In *Knowledge Engineering Review*, pages 107—-127, 2002.

[79] Vivian Lai and Chenhao Tan. On human predictions with explanations and predictions of machine learning models: A case study on deception detection. In *FAT* 2019 - Proceedings of the 2019 Conference on Fairness, Accountability, and Transparency*, pages 29–38, New York, New York, USA, jan 2019. Association for Computing Machinery, Inc. ISBN 9781450361255. doi: 10.1145/3287560.3287590.

[80] Himabindu Lakkaraju and Osbert Bastani. "how do i fool you?": Manipulating user trust via misleading black box explanations. In *AIES 2020 - Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, pages 79–85. Association for Computing Machinery, Inc, nov 2020. ISBN 9781450371100. doi: 10.1145/3375627.3375833.

[81] Nicholas Lange, Stephen C. Strother, Jon R. Anderson, Finn Å Nielsen, Andrew P. Holmes, Thomas Kolenda, Robert Savoy, and Lars Kai Hansen. Plurality and resemblance in fMRI data analysis. *NeuroImage*, 10(3 I):282–303, sep 1999. ISSN 10538119. doi: 10.1006/nimg.1999.0472.

[82] Sebastian Lapuschkin, Stephan Wäldchen, Alexander Binder, Grégoire Montavon, Wojciech Samek, and Klaus Robert Müller. Unmasking Clever Hans predictors and assessing what machines really learn. *Nature Communications*, 10(1), dec 2019. ISSN 20411723. doi: 10.1038/s41467-019-08987-4.

[83] Zachary C. Lipton. The mythos of model interpretability. *Communications of the ACM*, 61(10):35–43, jun 2018. ISSN 15577317. doi: 10.1145/3233231.

[84] Weibo Liu, Zidong Wang, Xiaohui Liu, Nianyin Zeng, Yurong Liu, and Fuad E. Alsaadi. A survey of deep neural network architectures and their applications. *Neurocomputing*, 234:11–26, apr 2017. ISSN 18728286. doi: 10.1016/j.neucom.2016.12.038.

[85] Francesco Locatello, Stefan Bauer, Mario Lucie, Gunnar Rätsch, Sylvain Gelly, Bernhard Schölkopf, and Olivier Bachem. Challenging common assumptions in the unsupervised learning of disentangled representations. In *36th International Conference on Machine Learning, ICML 2019*, volume 2019-June, pages 7247–7283, nov 2019. ISBN 9781510886988.

[86] Scott M. Lundberg and Su In Lee. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*, volume 2017-Decem, pages 4766–4775, may 2017.

[87] Piyushi Manupriya, J Saketha Nath, and Vineeth N Balasubramanian. SEA-NN: Submodular Ensembled Attribution for Neural Networks. Technical report, 2020.

[88] Brent Mittelstadt, Chris Russell, and Sandra Wachter. Explaining explanations in AI. In *FAT* 2019 - Proceedings of the 2019 Conference on Fairness, Accountability, and Transparency*, pages 279–288. Association for Computing Machinery, Inc, jan 2019. ISBN 9781450361255. doi: 10.1145/3287560.3287574.

[89] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A. Rusu, Joel Veness, Marc G. Bellemare, Alex Graves, Martin Riedmiller, Andreas K. Fidjeland, Georg Ostrovski, Stig Petersen, Charles Beattie, Amir Sadik, Ioannis Antonoglou, Helen King, Dharshan Kumaran, Daan Wierstra, Shane Legg, and Demis Hassabis. Human-level control through deep reinforcement learning. *Nature*, 518(7540):529–533, feb 2015. ISSN 14764687. doi: 10.1038/nature14236. URL https://www.nature.com/articles/nature14236.

[90] Christoph Molnar. Interpretable Machine Learning. A Guide for Making Black Box Models Explainable. *Book*, page 247, 2019.

[91] W. James Murdoch, Peter J. Liu, and Bin Yu. Beyond word importance: Contextual decomposition to extract interactions from LSTMs. In *6th International Conference on Learning Representations, ICLR 2018 - Conference Track Proceedings*, jan 2018.

[92] W. James Murdoch, Chandan Singh, Karl Kumbier, Reza Abbasi-Asl, and Bin Yu. Definitions, methods, and applications in interpretable machine learning. In *Proceedings of the National Academy of Sciences of the United States of America*, volume 116, pages 22071–22080. National Academy of Sciences, jan 2019. doi: 10.1073/pnas.1900654116.

[93] Menaka Narayanan, Emily Chen, Jeffrey He, Been Kim, Sam Gershman, and Finale Doshi-Velez. How do Humans Understand Explanations from Machine Learning Systems? An Evaluation of the Human-Interpretability of Explanation. feb 2018.

[94] Adam Noack, Isaac Ahern, Dejing Dou, and Boyang Li. Does Interpretability of Neural Networks Imply Adversarial Robustness? Technical report, 2019.

[95] Center for Data Science of Chicago and Public Policy University. Aequitas - The Bias Report. URL http://aequitas.dssg.io/.

[96] Ahmed Osman, Leila Arras, and Wojciech Samek. Towards Ground Truth Evaluation of Visual Explanations. mar 2020.

[97] Dong Huk Park, Lisa Anne Hendricks, Zeynep Akata, Anna Rohrbach, Bernt Schiele, Trevor Darrell, and Marcus Rohrbach. Multimodal Explanations: Justifying Decisions and Pointing to the Evidence. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 8779–8788, 2018. ISBN 9781538664209. doi: 10.1109/CVPR.2018.00915.

[98] Kevin Peachey. Sexist and biased? How credit firms make decisions - BBC News, 2019. URL https://www.bbc.co.uk/news/business-50432634.

[99] Danish Pruthi, Mansi Gupta, Bhuwan Dhingra, Graham Neubig, and Zachary C. Lipton. Learning to Deceive with Attention-Based Explanations. Technical report, sep 2019.

[100] Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do ImageNet classifiers generalize to ImageNet? In *36th International Conference on Machine Learning, ICML 2019*, volume 2019-June, pages 9413–9424, 2019. ISBN 9781510886988.

[101] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "Why should i trust you?" Explaining the predictions of any classifier. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, volume 13-17-Augu, pages 1135–1144, 2016. ISBN 9781450342322. doi: 10.1145/2939672.2939778.

[102] Laura Rieger. Separable explanations of neural network decisions. In *Proceedings Workshop on Interpreting, Explaining and Visualizing Deep Learning (at NIPS)*. Proceedings Workshop on Interpreting, Explaining and Visualizing Deep Learning (at NIPS), 2017.

[103] Laura Rieger and Lars Kai Hansen. IROF: a low resource evaluation metric for explanation methods. In *Workshop AI for Affordable Healthcare at ICLR 2020, Addis Ababa, Ethiopia*, 2020.

[104] Laura Rieger and Lars Kai Hansen. A simple defense against adversarial attacks on heatmap explanations. In *Workshop on Human Interpretability in Machine Learning (WHI) at ICML*, 2020.

[105] Laura Rieger, Pattarawat Chormai, Grégoire Montavon, Lars Kai Hansen, and Klaus-Robert Müller. Structuring Neural Networks for More Explainable Predictions. In *Explainable and Interpretable Models in Computer Vision and Machine Learning*, pages 115–131. Springer, Cham, 2018.

[106] Laura Rieger, Rasmus M. Th. Høegh, and Lars K. Hansen. Client Adaptation improves Federated Learning with Simulated Non-IID Clients. *International Workshop on Federated Learning for User Privacy and Data Confidentiality in Conjunction with ICML 2020 (FL-ICML'20)*, 2020.

[107] Laura Rieger, Chandan Singh, W James Murdoch, and Bin Yu. Interpretations are Useful: Penalizing Explanations to Align Neural Networks with Prior Knowledge. In *Proceedings of the International Conference on Machine Learning*, 2020.

[108] Andrew Slavin Ros and Finale Doshi-Velez. Improving the adversarial robustness and interpretability of deep neural networks by regularizing their input gradients. In *32nd AAAI Conference on Artificial Intelligence, AAAI 2018*, pages 1660–1669, apr 2018. ISBN 9781577358008.

[109] Ribana Roscher, Bastian Bohn, Marco F. Duarte, and Jochen Garcke. Explainable Machine Learning for Scientific Insights and Discoveries. *IEEE Access*, 8: 42200–42216, 2020. ISSN 21693536. doi: 10.1109/ACCESS.2020.2976199.

[110] Andrew Slavin Ross, Michael C. Hughes, and Finale Doshi-Velez. Right for the right reasons: Training differentiable models by constraining their explanations. *IJCAI International Joint Conference on Artificial Intelligence*, 0:2662–2670, mar 2017. ISSN 10450823. doi: 10.24963/ijcai.2017/371.

[111] JK Rowling. *Harry Potter and the chamber of secrets.* Bloomsbury, 1998.

[112] Sebastian Ruder. An overview of gradient descent optimization algorithms. Technical report, sep 2016.

[113] Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5):206–215, may 2019. ISSN 25225839. doi: 10.1038/s42256-019-0048-x.

[114] Wojciech Samek, Alexander Binder, Grégoire Montavon, Sebastian Lapuschkin, and Klaus Robert Müller. Evaluating the visualization of what a deep neural network has learned. *IEEE Transactions on Neural Networks and Learning Systems*, 28(11):2660–2673, nov 2017. ISSN 21622388. doi: 10.1109/TNNLS.2016.2599820.

[115] Philipp Schmidt and Felix Biessmann. Quantifying Interpretability and Trust in Machine Learning Systems. Technical report, jan 2019.

[116] Lukas Schott, Jonas Rauber, Matthias Bethge, and Wieland Brendel. Towards the first adversarially robust neural network model on MNIST. In *7th International Conference on Learning Representations, ICLR 2019*, 2019.

[117] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. *International Journal of Computer Vision*, 128(2):336–359, 2020. ISSN 15731405. doi: 10.1007/s11263-019-01228-7.

[118] Sofia Serrano and Noah A. Smith. Is Attention Interpretable? In *ACL 2019 - 57th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference*, pages 2931–2951. Association for Computational Linguistics (ACL), jun 2019.

[119] Yujun Shen, Jinjin Gu, Xiaoou Tang, and Bolei Zhou. Interpreting the Latent Space of GANs for Semantic Face Editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9243—-9252, 2020. doi: 10.1109/cvpr42600.2020.00926.

[120] Edward H. Shortliffe, Stanton G. Axline, Bruce G. Buchanan, Thomas C. Merigan, and Stanley N. Cohen. An Artificial Intelligence program to advise physicians regarding antimicrobial therapy. *Computers and Biomedical Research*, 6(6):544–560, dec 1973. ISSN 00104809. doi: 10.1016/0010-4809(73)90029-3.

[121] Edward H. Shortliffe, Randall Davis, Stanton G. Axline, Bruce G. Buchanan, C. Cordell Green, and Stanley N. Cohen. Computer-based consultations in clinical therapeutics: Explanation and rule acquisition capabilities of the MYCIN system. *Computers and Biomedical Research*, 8(4):303–320, aug 1975. ISSN 00104809. doi: 10.1016/0010-4809(75)90009-9.

[122] Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. Learning important features through propagating activation differences. In *34th International Conference on Machine Learning, ICML 2017*, volume 7, pages 4844–4866, apr 2017. ISBN 9781510855144.

[123] David Silver, Aja Huang, Chris J. Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda

Panneershelvam, Marc Lanctot, Sander Dieleman, Dominik Grewe, John Nham, Nal Kalchbrenner, Ilya Sutskever, Timothy Lillicrap, Madeleine Leach, Koray Kavukcuoglu, Thore Graepel, and Demis Hassabis. Mastering the game of Go with deep neural networks and tree search. *Nature*, 529(7587):484–489, jan 2016. ISSN 14764687. doi: 10.1038/nature16961. URL https://www.nature.com/articles/nature16961.

[124] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. In *2nd International Conference on Learning Representations, ICLR 2014 - Workshop Track Proceedings*, dec 2014.

[125] Chandan Singh, Bin Yu, and W. James Murdoch. Hierarchical interpretations for neural network predictions. In *7th International Conference on Learning Representations, ICLR 2019*, 2019.

[126] Sumedha Singla, Brian Pollack, Junxiang Chen, and Kayhan Batmanghelich. Explanation by Progressive Exaggeration. In *International Conference on Learning Representations*, pages 1–13, 2019. URL http://arxiv.org/abs/1911.00483.

[127] Leon Sixt, Maximilian Granz, and Tim Landgraf. When Explanations Lie: Why Many Modified BP Attributions Fail. Technical report, 2019.

[128] Dylan Slack, Sophie Hilgard, Emily Jia, Sameer Singh, and Himabindu Lakkaraju. Fooling LIME and SHAP: Adversarial attacks on post hoc explanation methods. In *AIES 2020 - Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, pages 180–186, nov 2020. ISBN 9781450371100. doi: 10.1145/3375627.3375830.

[129] Zachary Small. 600,000 Images Removed from AI Database After Art Project Exposes Racist Bias. URL https://hyperallergic.com/518822/600000-images-removed-from-ai-database-after-art-project-exposes-racist-bias/.

[130] Daniel Smilkov, Nikhil Thorat, Been Kim, Fernanda Viégas, and Martin Wattenberg. SmoothGrad: removing noise by adding noise. Technical report, jun 2017.

[131] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, and Ruslan Salakhutdinov. Dropout: A Simple Way to Prevent Neural Networks from Overfitting. Technical report, 2014.

[132] Stephen C. Strother, Jon Anderson, Lars Kai Hansen, Ulrik Kjems, Rafal Kustra, John Sidtis, Sally Frutiger, Suraj Muley, Stephen LaConte, and David Rottenberg. The quantitative evaluation of functional neuroimaging experiments: The NPAIRS data analysis framework. *NeuroImage*, 15(4):747–771, apr 2002. ISSN 10538119. doi: 10.1006/nimg.2001.1034.

[133] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *34th International Conference on Machine Learning, ICML 2017*, volume 7, pages 5109–5118, jul 2017. ISBN 9781510855144.

[134] William R. Swartout. XPLAIN: a system for creating and explaining expert consulting programs. *Artificial Intelligence*, 21(3):285–325, sep 1983. ISSN 00043702. doi: 10.1016/S0004-3702(83)80014-9.

[135] William R. Swartout and Johanna D. Moore. Explanation in Second Generation Expert Systems. In *Second Generation Expert Systems*, pages 543–585. Springer Berlin Heidelberg, 1993. doi: 10.1007/978-3-642-77927-5_24.

[136] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In *2nd International Conference on Learning Representations, ICLR 2014 - Conference Track Proceedings*, dec 2014.

[137] Richard Tomsett, Dave Braines, Dan Harborne, Alun Preece, and Supriyo Chakraborty. Interpretable to Whom? A Role-based Model for Analyzing Interpretable Machine Learning Systems. Technical report, jun 2018.

[138] Richard Tomsett, Dan Harborne, Supriyo Chakraborty, Prudhvi Gurram, and Alun Preece. Sanity Checks for Saliency Metrics. Technical report, 2019.

[139] Dimitris Tsipras, Shibani Santurkar, Logan Engstrom, Alexander Turner, and Aleksander Madry. Robustness may be at odds with accuracy. In *7th International Conference on Learning Representations, ICLR 2019*, may 2019.

[140] Sjoerd van Steenkiste, Francesco Locatello, Jürgen Schmidhuber, and Olivier Bachem. Are disentangled representations helpful for abstract visual reasoning? In *Advances in Neural Information Processing Systems*, volume 32, may 2019. ISBN 1905.12506v2.

[141] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 2017-Decem, pages 5999–6009, 2017.

[142] Tom Viering, Ziqi Wang, Marco Loog, and Elmar Eisemann. How to Manipulate CNNs to Make Them Lie: the GradCAM Case. Technical report, 2019. URL http://arxiv.org/abs/1907.10901.

[143] Andrey Voynov and Artem Babenko. Unsupervised Discovery of Interpretable Directions in the GAN Latent Space. Technical report, feb 2020.

[144] Sandra Wachter, Brent Mittelstadt, and Luciano Floridi. Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation. *SSRN Electronic Journal*, 7(2):76–99, 2017. ISSN 07384602. doi: 10.2139/ssrn.2903469.

[145] Zifan Wang, Piotr Mardziel, Anupam Datta, and Matt Fredrikson. Interpreting interpretations: Organizing attribution methods by criteria. Technical report, 2020.

[146] Ethan Weinberger, Joseph Janizek, and Su-In Lee. Learning Deep Attribution Priors Based On Prior Knowledge. Technical report, 2019.

[147] M R Wick. The 1988 AAAI Workshop on Explanation. *AI Magazine*, 10(3): 22–26, 1989.

[148] Sarah Wiegreffe and Yuval Pinter. Attention is not not Explanation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 11–20, 2019. ISBN 9781950737901. doi: 10.18653/v1/d19-1002.

[149] Chih Kuan Yeh, Cheng Yu Hsieh, Arun Sai Suggala, David I. Inouye, and Pradeep Ravikumar. On the (In)fidelity and sensitivity of explanations. In *Advances in Neural Information Processing Systems*, volume 32, jan 2019.

[150] John R. Zech, Marcus A. Badgeley, Manway Liu, Anthony B. Costa, Joseph J. Titano, and Eric K. Oermann. Confounding variables can degrade generalization performance of radiological deep learning models. Technical report, 2018.

[151] Xinyang Zhang, Ningfei Wang, Hua Shen, Shouling Ji, Xiapu Luo, and Ting Wang. *Interpretable Deep Learning under Fire*. 2018. ISBN 978-1-939133-17-5.

[152] Yujia Zhang, Kuangyan Song, Yiming Sun, Sarah Tan, and Madeleine Udell. "Why Should You Trust My Explanation?" Understanding Uncertainty in LIME Explanations. Technical report, apr 2019.

[153] Yunfeng Zhang, Q. Vera Liao, and Rachel K.E. Bellamy. Efect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making. In *FAT* 2020 - Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 295–305, New York, NY, USA, jan 2020. Association for Computing Machinery, Inc. ISBN 9781450369367. doi: 10.1145/3351095.3372852.

[154] Luisa M. Zintgraf, Taco S. Cohen, Tameem Adel, and Max Welling. Visualizing deep neural network decisions: Prediction difference analysis. In *5th International Conference on Learning Representations, ICLR 2017 - Conference Track Proceedings*, feb 2017.

CONTRIBUTION A

Interpretability in Intelligent Systems - A New Concept?

Lars Kai Hansen and Laura Rieger. Interpretability in Intelligent Systems - A New Concept? In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 11700 LNCS, pages 41–49. Springer, Cham, 2019

Interpretability in intelligent systems - a new concept?

Lars Kai Hansen [*†] Laura Rieger[*]

Abstract

The very active community for interpretable machine learning can learn from the rich 50+ year history of explainable AI. We here give two specific examples from this legacy that could enrich current interpretability work: First, *Explanation desiderata* we were point to the rich set of ideas developed in the 'explainable expert systems' field and, second, tools for *quantification of uncertainty* of high-dimensional feature importance maps which have been developed in the field of computational neuroimaging.

1 Neural network interpretability

High activity research fields often develop to be somewhat myopic, simply because the large body of published work leaves little time to follow progress in other areas or even to look back at previous research in the field. Deep learning interpretability is such an area. One could easily get the impression that the interpretability issue surfaced with the new wave of deep learning, however, this is not the case. While end-to-end learning has hugely accentuated the need for explanations, interpretability is an active research topic with an over 50-year history. In fact, since the early days of intelligent systems the importance and focus on interpretability has only increased [29]. From scientific contexts, where interpretability methods can assist formulation of causal hypotheses, see e.g., work in bio-medicine [41] and computational chemistry [38], to recent societal importance in the European Union's General Data Protection Regulatory, establishing the so-called *Right to explanation* [14].

Here we make two dives into the rich history of explainability in intelligent systems and ask what can modern work learn?

First a semantic note. The terms interpretability and explainability are often used interchangeably in the literature. However, in a recent review [12] a useful distinction is made. The more general concept is explainability which

[*]Technical University of Denmark, DTU Compute, Kgs. Lyngby, Denmark
[†]lkai@dtu.dk
The final authenticated version is available online at https://doi.org/10.1007/978-3-030-28954-6_3

covers interpretability, i.e., to communicate machine learning function to user, and completeness, i.e., that the explanation is a close enough approximation that it can be audited. The distinction is described: '...interpretability alone is insufficient. In order for humans to trust black-box methods, we need explainability models that are able to summarize the reasons for neural network behavior, gain the trust of users, or produce insights about the causes of their decisions. While interpretability is a substantial first step, these mechanisms need to also be complete, with the capacity to defend their actions, provide relevant responses to questions, and be audited' [12]. In the tutorial [29] a related distinction is made: '... interpreting the concepts learned by a model by building prototypes, and explaining the model's decisions by identifying the relevant input variables'.

Explainability, in this broader sense, has been a key component in several intelligent systems communities and the central tenet of this paper is that future work can learn from looking back at this history. We will focus on two specific lines of research, the first concerns the broader foundation of explainability: What are the desiderata, i.e., the salient dimensions and issues that should be addressed? Our second focus area concerns the important specific challenge of understanding the dimensions of uncertainty in machine learning models and their explanations.

Going back in time prior to the new wave of deep learning, many have stressed the importance of interpretability. The classic paper *Statistical Modeling: The Two Cultures* has a strong focus on interpretability [3]. Breiman notes: 'Occams Razor, long admired, is usually interpreted to mean that simpler is better. Unfortunately, in prediction, accuracy and simplicity(interpretability) are in conflict. For instance, linear regression gives a fairly interpretable picture of the y, x relation. But its accuracy is usually less than that of the less interpretable neural nets'. As we will see below, this dilemma has been acknowledged by the explainable expert systems community many years earlier. Breiman clearly expressed his preferences: 'On interpretability, trees rate an A+', however, it was already known that trees and rule based systems have severe limitations when it comes to both implementing and comprehending function, see e.g. [28].

The interest in intelligent systems' interpretability has earlier roots. In a position paper [27] discussed how AI would pass different criteria from weak to ultra-strong: 'The ultra-strong criterion demands that the system be capable not only of explaining how it has structured its acquired skills: it should also be able to teach them'. We are not quite there yet.

Going further back in early expert system history, explanation and human interaction were key issues. Expert systems in the late 60's - like 'SCHOLAR' developed for instructional support - were designed for interaction [5], such as explaining why a student's answer was wrong in a mixed initiative dialogue. Stanford's widely discussed 'MYCIN' expert system for antimicrobial selection was designed with three components: A rule based decision support component that combined MYCIN and physicians judgment, an explanation module and a learning module [39, 40, 4]. This rule based system had about 200 rules in 1975.

MYCIN developers held it self-evident that AI could get medical acceptance only with convincing explanations [40]. Thus, MYCIN was equipped to map its internal rules to natural language and answer both 'why' and 'how' questions. In the 1984 book summarizing experiences with MYCIN [4] no less than four chapters are devoted to MYCIN's explanation mechanisms. Prior to Breiman's comments, earlier work on explainability in statistics includes Good's discussion of evidence in context of belief networks [13]. Good considered three dimensions of explanations: 'What', concerning semantic explanations as in a dictionary, 'How' as in natural or manufacturing process descriptions, and finally the 'Why' type explanations - hypothesizing causal mechanisms behind an event.

2 Desiderata of explainable AI

By 1983 the MYCIN system had expanded to 500 rules and the state of the art was summarized in a review in Science [9]. Yet, expert systems moved on and important principles can be learned from the 1993 review of 'second generation explainable expert systems'. [45]'s review lists five general desiderata for useful explanations of AI, adding significant perspective to recent work in the field:

1. $\mathcal{D}_1$ Fidelity: the explanation must be a reasonable representation of what the system actually does.
2. $\mathcal{D}_2$ Understandability: Involves multiple usability factors including terminology, user competencies, levels of abstraction and interactivity.
3. $\mathcal{D}_3$ Sufficiency: Should be able to explain function and terminology and be detailed enough to justify decision.
4. $\mathcal{D}_4$ Low Construction overhead: The explanation should not dominate the cost of designing AI.
5. $\mathcal{D}_5$ Efficiency: The explanation system should not slow down the AI significantly.

Expert systems have developed through several generations. The notion of second versus first generation AI was based on the modes of explanation. First generation systems were characterized by explanations based directly on rules applied by the AI to reach decisions. This leads to high fidelity ($\mathcal{D}_1$), but often conflicts with understandability ($\mathcal{D}_2$) because the rules used for inference may be incomprehensive for the user [45]. So-called *Explainable Expert Systems* (EES) addressed this dilemma. The XPLAIN system [44] is an example. XPLAIN was based on two key principles to enhance understandability: 'explicitly distinguishing different forms of domain knowledge present in the knowledge base and formal recording of the system development process' [33]. The evaluation of XPLAIN is anecdotal, yet quite convincing. Cases are presented in which the system is able to answer 'why' questions - and even at times resorting to 'white lies' to create a smoother learning experience [44]. Computational complexity both in construction and execution (desiderata $\mathcal{D}_4 - \mathcal{D}_5$) are not so prominent in current literature, although the most widely used methods differ significant

in complexity. The so-called Local Interpretable Model-agnostic Explanation (LIME) scheme, for example, is based on image segmentation, random sampling and multiple linear model fittings, hence rather complex at explanation time [35],hence a challenge to $\mathcal{D}_5$. An approach such a 'Testing with Concept Activation Vectors' (TCAV) comes at a significant initial cost [18], hence may pose a challenge to $\mathcal{D}_4$.

Much of the EES progress was produced in the context of rule based expert systems, while AI based on machine learning - so-called connectionists' methods - more often was considered 'black box'. Interest in connectionists' methods was primarily based on performance and not interpretability, c.f., the quote from [2] '...symbolic learning techniques produce more understandable outputs but they are not as good as connectionist learning techniques in generalization' . We already noted that this view was propagated by Breiman, hence, the sparking interest in converting existing neural networks to decision tree form [46, 47, 1] or even learn neural networks that more readily are converted to trees see for example work by Gallant [11] and by Craven and Shavlik [6]. But trees may not deliver on $\mathcal{D}_2$, in particular, as discussed above and noted by [44] - the intuitive appeal of trees fails in practice when trees get to be complex in structure or operate in high dimensional feature spaces. These challenges were also recently noted in [32]. For domains where modern neural networks excel such as image, audio and text data, tree based explanations are challenged.

Returning to the list of desiderata, several recent papers have aimed at framing the discourse of interpretability. Unaware of [45], Lipton notes that interpretability is not a well-defined concept and goes on to discuss multiple dimensions of interpretability and formulates a set of desiderata [24, 25] closely related to $\mathcal{D}_1 - \mathcal{D}_3$. Lipton's desiderata read i) 'Trust', ii) 'Causality', iii) 'Transferability', iv) 'Informativeness', and v) 'Fair and Ethical Decision-Making'. Here Lipton discusses several dimensions of i) 'Trust' mostly covered in desiderata $\mathcal{D}_1 - \mathcal{D}_2$, ii) 'Causality' is roughly equivalent to [45]'s $\mathcal{D}_3$, while the notion of 'Transferability' and 'Informativeness' both refer to the user's ability to gain abstract 'knowledge' from explanations. This idea also appeared in the original paper's discussion of usability $\mathcal{D}_2$, viz. the need to explain a system at different levels of abstraction. 'Fair and Ethical Decision-Making' is noted by Lipton as an area that specifically requires interpretability. In [45] such considerations are framed in a general discussion of usability ($\mathcal{D}_2$). It is also noted that an explanation systems must be able to explain from different perspectives '.. e.g., form versus function in the biological domain or safety vs. profitability in the financial domain'.

The usability dimension ($\mathcal{D}_2$) remains an important issue in contemporary interpretability papers. The question 'Interpretable to Whom?' was raised in [7, 48, 32] focusing on the user and addressed by human factors evaluation. In fact, [7] open their paper with the more general statement 'Unfortunately, there is little consensus on what interpretability in machine learning is and how to evaluate it for benchmarking'. Their key contribution is to point out that current machine learning workflows are incomplete in the sense that they have unspecified objectives from the application domain. This can be important issues that

were not included in the machine learning objective function: '...incompleteness that produces some kind of unquantified bias, e.g. the effect of including domain knowledge in a model selection process'[7]. Unaware of the results of the EES community they focus on usability, and in case there are human users involved, a fully specified application somehow will entail human factors evaluation, immediately making [45]'s discussion of desideratum $\mathcal{D}_2$ relevant. When evaluating AI explanation systems with human subjects, we should be aware of the users' potential cognitive biases [37]. In the context of explainability, it is interesting to note that users may suffer from biases, for example the interesting phenomenon 'choice blindness' discovered by [17]. Choice blindness shows up in failure to make and explain consistent decisions. In the work of [17] magicians 'fool' users to explain decisions users did or did not make with similar strengths. Also, the importance of actual usability evaluation of explanation methods have appeared early. A 1990 AAAI workshop featured work on user scenarios [8] and later work was reported in [43].

Breiman equated simplicity and interpretability. However, it is well-known that seemingly simple models can be hard to interpret. Even simple linear classification models need careful tuning to optimize stability of feature importance maps [34]. The 'filters vs. pattern' discussion that first emerged in the context of neuroimaging is another example of unexpected complexity. In this context, there is an important difference between visualizing the classification model and the corresponding (causal) generative model. The difference is induced by correlated input noise and can lead to wrong conclusions if not handled appropriately [16]. Similar challenges appear in deep networks [19]. Further examples of the dissociation of simplicity and interpretability are discussed in [24, 25], citing the work on 'Interpretable Boosted Nave Bayes Classification' by [36]. This paper opens with a statement aligned with the Breiman's dilemma: 'Efforts to develop classifiers with strong discrimination power using voting methods have marginalized the importance of comprehensibility'. The objective of the paper is to demonstrate that the interpretation problem for voting systems can be mitigated. Specifically, [26]'s tools for interpretation in Naive Bayes classifiers is shown to be useful for complex boosting ensembles.

3 Quantify similarity and uncertainty of feature importance maps using resampling

In certain application domains of neural networks, including for example scientific computing and bio-medicine, interpretation plays and important role and tools have been developed for explanation of neural networks' function.

In early work on mind reading based on brain scanning interpretability was naturally in focus [23, 30]. The dominating analysis paradigm at the time was SPM 'statistical parametric mapping' [10]. This approach produced intuitively appealing three dimensional brain maps (SPMs) of voxel-wise significant activation. These maps have had significant impact in the field and it was a strong

aim of neural networks visualizations to produce matching SPMs. The tools developed included 3D mapping of voxel-wise saliency [31] and sensitivity [31, 20]. The usability ($\mathcal{D}_2$) for neuroscientists were enhanced by embedding the maps in 3D anatomically informed navigation tools, see e.g., [31] for examples.

Concerning the first desideratum ($\mathcal{D}_1$)- fidelity of explanations - we need to consider the two logical fundamental properties: 'Existence' and 'uniqueness'. Considering the many constraints imposed by the desiderata, the very existence of a satisfactory interpretability scheme is a non-trivial issue. Finding such schemes is the concern of current interpretability engineering literature. Given existence, we face an equally important issue of uniqueness. Note that at least two mechanisms of uncertainty can contribute to non-uniqueness: Firstly, epistemic uncertainty, i.e., uncertainty in the explainability model, typically induced by a combination of limited data and knowledge. Epistemic uncertainty gives rise to multiple competing paradigm of explainability. The second source of non-uniqueness is the inherent randomness of a given problem domain for which noise and finite samples can conspire to create large fluctuations in solutions ('aleatory uncertainty').

Epistemic uncertainty was discussed in detail in [22]. Nine different interpretation schemes were evaluated to explore the diversity in model space and learn similarities. The idiosyncratic scales employed by different mapping procedures is a significant challenge for quantitative comparisons of visualizations. This problem was addressed in [15] proposing a simple nonparametric approach to standardization of maps, hence, allowing maps to be meaningfully combined, e.g., by simple averaging. Such consensus based methods allow reduction of model uncertainty and quantification of inter-map differences (epistemic uncertainty).

Aleatory uncertainty in brain maps was addressed by the so-called NPAIRS framework [42]. Statistical re-sampling techniques such as split-half, can provide unbiased estimates of variance of interpretability heat maps. This allows for mapping of the local visualization 'effect size', by scaling heat maps by their standard deviation. Application of these tools include imaging pipeline optimization [21]. Outside the original domain of these methods, they have been applied for skin cancer diagnosis support [41].

4 Concluding remarks

Explainability is at the core of modern machine learning. The transparency made possible by effective tools for explainability can improve design and debugging for the machine learning engineer and even more importantly, our users' trust and usability in the tools we develop. It would be productive if the very active community of scientist working in this field made an even bigger effort to embrace the rich 50+ year history of explainable AI. Here we focused on two specific topics from this legacy that could enrich current interpretability work: 1) *Explanation desiderata*, were we pointed to a rich set of ideas developed in the 'explainable expert systems' field and 2) tools for *quantification of uncertainty*

of high-dimensional feature importance maps, originally developed in the field of computational neuroimaging.

References

[1] Andrews, R., Diederich, J., Tickle, A.B.: Survey and critique of techniques for extracting rules from trained artificial neural networks. Knowledge-based systems **8**(6), 373–389 (1995)

[2] Boz, O.: Converting a trained neural network to a decision tree dectext-decision tree extractor (2000)

[3] Breiman, L.: Statistical modeling: The two cultures (with comments and a rejoinder by the author). Statist. Sci. **16**(3), 199–231 (08 2001). https://doi.org/10.1214/ss/1009213726, `https://doi.org/10.1214/ss/1009213726`

[4] Bruce, G., Buchanan, B., Shortliffe, E.: Rule-based expert systems: the mycin experiments of the stanford heuristic programming project (1984)

[5] Carbonell, J.R.: Ai in cai: An artificial-intelligence approach to computer-assisted instruction. IEEE transactions on man-machine systems **11**(4), 190–202 (1970)

[6] Craven, M.W., Shavlik, J.W.: Using sampling and queries to extract rules from trained neural networks. In: Machine Learning Proceedings 1994, pp. 37–45. Elsevier (1994)

[7] Doshi-Velez, F., Kim, B.: Towards a rigorous science of interpretable machine learning. arXiv preprint arXiv:1702.08608 (2017)

[8] Druzdzel, M.J., Henrion, M.: Using scenarios to explain probabilistic inference. In: Working notes of the AAAI-90 Workshop on Explanation. pp. 133–141. Citeseer (1990)

[9] Duda, R.O., Shortliffe, E.H.: Expert systems research. Science **220**(4594), 261–268 (1983)

[10] Friston, K.J., Holmes, A.P., Worsley, K.J., Poline, J.P., Frith, C.D., Frackowiak, R.S.: Statistical parametric maps in functional imaging: a general linear approach. Human brain mapping **2**(4), 189–210 (1994)

[11] Gallant, S.I.: Connectionist expert systems. Communications of the ACM **31**(2), 152–169 (1988)

[12] Gilpin, L.H., Bau, D., Yuan, B.Z., Bajwa, A., Specter, M., Kagal, L.: Explaining explanations: An approach to evaluating interpretability of machine learning. arXiv preprint arXiv:1806.00069 (2018)

[13] Good, I.: Explicativity: a mathematical theory of explanation with statistical applications. Proc. R. Soc. Lond. A **354**(1678), 303–330 (1977)

[14] Goodman, B., Flaxman, S.: European union regulations on algorithmic decision-making and a" right to explanation". arXiv preprint arXiv:1606.08813 (2016)

[15] Hansen, L.K., Nielsen, F.Å., Strother, S.C., Lange, N.: Consensus inference in neuroimaging. NeuroImage **13**(6), 1212–1218 (2001)

[16] Haufe, S., Meinecke, F., Görgen, K., Dähne, S., Haynes, J.D., Blankertz, B., Bießmann, F.: On the interpretation of weight vectors of linear models in multivariate neuroimaging. Neuroimage **87**, 96–110 (2014)

[17] Johansson, P., Hall, L., Sikström, S., Olsson, A.: Failure to detect mismatches between intention and outcome in a simple decision task. Science **310**(5745), 116–119 (2005)

[18] Kim, B., Wattenberg, M., Gilmer, J., Cai, C., Wexler, J., Viegas, F., et al.: Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). In: International Conference on Machine Learning. pp. 2673–2682 (2018)

[19] Kindermans, P.J., Schütt, K.T., Alber, M., Müller, K.R., Erhan, D., Kim, B., Dähne, S.: Learning how to explain neural networks: Patternnet and patternattribution. arXiv preprint arXiv:1705.05598 (2017)

[20] Kjems, U., Hansen, L.K., Anderson, J., Frutiger, S., Muley, S., Sidtis, J., Rottenberg, D., Strother, S.: The quantitative evaluation of functional neuroimaging experiments: mutual information learning curves. NeuroImage **15**(4), 772–786 (2002)

[21] LaConte, S., Anderson, J., Muley, S., Ashe, J., Frutiger, S., Rehm, K., Hansen, L.K., Yacoub, E., Hu, X., Rottenberg, D., et al.: The evaluation of preprocessing choices in single-subject bold fmri using npairs performance metrics. NeuroImage **18**(1), 10–27 (2003)

[22] Lange, N., Strother, S.C., Anderson, J.R., Nielsen, F.Å., Holmes, A.P., Kolenda, T., Savoy, R., Hansen, L.K.: Plurality and resemblance in fmri data analysis. NeuroImage **10**(3), 282–303 (1999)

[23] Lautrup, B., Hansen, L.K., Law, I., Mørch, N., Svarer, C., Strother, S.C.: Massive weight sharing: a cure for extremely ill-posed problems. In: Workshop on supercomputing in brain research: From tomography to neural networks. pp. 137–144. Citeseer (1994)

[24] Lipton, Z.C.: The mythos of model interpretability (2016). arXiv preprint arXiv:1606.03490 (2016)

[25] Lipton, Z.C.: The mythos of model interpretability. Queue **16**(3), 30 (2018)

[26] Madigan, D., Mosurski, K., Almond, R.G.: Graphical explanation in belief networks. Journal of Computational and Graphical Statistics **6**(2), 160–181 (1997)

[27] Michie, D.: Machine learning in the next five years. In: Proceedings of the 3rd European Conference on European Working Session on Learning. pp. 107–122. Pitman Publishing (1988)

[28] Minsky, M.L.: Logical versus analogical or symbolic versus connectionist or neat versus scruffy. AI magazine **12**(2), 34 (1991)

[29] Montavon, G., Samek, W., Müller, K.R.: Methods for interpreting and understanding deep neural networks. Digital Signal Processing **73**, 1–15 (2018)

[30] Mørch, N., Hansen, L.K., Strother, S.C., Svarer, C., Rottenberg, D.A., Lautrup, B., Savoy, R., Paulson, O.B.: Nonlinear versus linear models in functional neuroimaging: Learning curves and generalization crossover. In: Biennial International Conference on Information Processing in Medical Imaging. pp. 259–270. Springer (1997)

[31] Mørch, N.J., Kjems, U., Hansen, L.K., Svarer, C., Law, I., Lautrup, B., Strother, S., Rehm, K.: Visualization of neural networks using saliency maps. In: 1995 IEEE International Conference on Neural Networks. IEEE (1995)

[32] Narayanan, M., Chen, E., He, J., Kim, B., Gershman, S., Doshi-Velez, F.: How do humans understand explanations from machine learning systems? an evaluation of the human-interpretability of explanation. arXiv preprint arXiv:1802.00682 (2018)

[33] Neches, R., Swartout, W.R., Moore, J.D.: Enhanced maintenance and explanation of expert systems through explicit models of their development. IEEE Transactions on Software Engineering (11), 1337–1351 (1985)

[34] Rasmussen, P.M., Hansen, L.K., Madsen, K.H., Churchill, N.W., Strother, S.C.: Model sparsity and brain pattern interpretation of classification models in neuroimaging. Pattern Recognition **45**(6), 2085–2100 (2012)

[35] Ribeiro, M.T., Singh, S., Guestrin, C.: Why should i trust you?: Explaining the predictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. pp. 1135–1144. ACM (2016)

[36] Ridgeway, G., Madigan, D., Richardson, T., O'Kane, J.: Interpretable boosted naïve bayes classification. In: KDD. pp. 101–104 (1998)

[37] Saposnik, G., Redelmeier, D., Ruff, C.C., Tobler, P.N.: Cognitive biases associated with medical decisions: a systematic review. BMC medical informatics and decision making **16**(1), 138 (2016)

[38] Schütt, K.T., Arbabzadah, F., Chmiela, S., Müller, K.R., Tkatchenko, A.: Quantum-chemical insights from deep tensor neural networks. Nature communications **8**, 13890 (2017)

[39] Shortliffe, E.H., Axline, S.G., Buchanan, B.G., Merigan, T.C., Cohen, S.N.: An artificial intelligence program to advise physicians regarding antimicrobial therapy. Computers and Biomedical Research **6**(6), 544–560 (1973)

[40] Shortliffe, E., Davis, R., Axline, S., Buchanan, B., Green, C., Cohen, S.: Computer-based consultations in clinical therapeutics: explanation and rule acquisition capabilities of the mycin system. Computers and biomedical research, an international journal **8**(4), 303–320 (1975)

[41] Sigurdsson, S., Philipsen, P.A., Hansen, L.K., Larsen, J., Gniadecka, M., Wulf, H.C.: Detection of skin cancer by classification of raman spectra. IEEE transactions on biomedical engineering **51**(10), 1784–1793 (2004)

[42] Strother, S.C., Anderson, J., Hansen, L.K., Kjems, U., Kustra, R., Sidtis, J., Frutiger, S., Muley, S., LaConte, S., Rottenberg, D.: The quantitative evaluation of functional neuroimaging experiments: the npairs data analysis framework. NeuroImage **15**(4), 747–771 (2002)

[43] Suermondt, H.J., Cooper, G.F.: An evaluation of explanations of probabilistic inference. In: Proceedings of the annual symposium on computer application in medical care. p. 579. American Medical Informatics Association (1992)

[44] Swartout, W.R.: Xplain: A system for creating and explaining expert consulting programs. Tech. rep., UNIVERSITY OF SOUTHERN CALIFORNIA MARINA DEL REY INFORMATION SCIENCES INST (1983)

[45] Swartout, W.R., Moore, J.D.: Explanation in second generation expert systems. In: Second generation expert systems, pp. 543–585. Springer (1993)

[46] Thrun, S.: Extracting provably correct rules from artificial neural networks. Technical Report IAI-TR-93-5, Institut for Informatik III Universitat Bonn, Germany (1994)

[47] Thrun, S.: Extracting rules from artificial neural networks with distributed representations. In: Advances in neural information processing systems. pp. 505–512 (1995)

[48] Tomsett, R., Braines, D., Harborne, D., Preece, A., Chakraborty, S.: Interpretable to whom? a role-based model for analyzing interpretable machine learning systems. arXiv preprint arXiv:1806.07552 (2018)

CONTRIBUTION B

IROF: a low resource evaluation metric for explanation methods

Laura Rieger and Lars Kai Hansen. IROF: a low resource evaluation metric for explanation methods. In *Workshop AI for Affordable Healthcare at ICLR 2020, Addis Ababa, Ethiopia*, 2020

IROF: A LOW RESOURCE EVALUATION METRIC FOR EXPLANATION METHODS

Laura Rieger & Lars Kai Hansen
DTU Compute
Technical University Denmark
Lyngby, Denmark
{lauri, lkai}@dtu.dk

ABSTRACT

The adoption of machine learning in health care hinges on the transparency of the used algorithms, necessitating the need for explanation methods. However, despite a growing literature on explaining neural networks, no consensus has been reached on how to evaluate those explanation methods. We propose IROF, a new approach to evaluating explanation methods that circumvents the need for manual evaluation. Compared to other recent work, our approach requires several orders of magnitude less computational resources and no human input, making it accessible to lower resource groups and robust to human bias.

1 INTRODUCTION

Health AI has already started to show value in high-income countries and there is broad consensus that the potential in low and middle income countries is even more significant Wahl et al. (2018). Health AI will play important roles, e.g., for cost reduction, scaling of treatments and for training of new medical professionals. Explainability of Health AI is a critical factor for all of these dimensions. Explainability methods have matured considerably and numerous schemes have appeared for e.g. explaining image based diagnostics Samek et al. (2019), however, the evaluation of visual explanation methods remains an unsolved problem. Hence it remains a challenge to select among the many proposed explainability schemes and understand the value for a particular application. When designing an evaluation metric for explanation methods, certain desiderata apply, e.g. a metric should be quantitative in order to assist decision making, it should be sensitive to the important features of the give problem, and it should not be too costly to evaluate.

We introduce IROF (**I**terative **R**emoval **O**f **F**eatures) as a new approach to quantitatively evaluate explanation methods without relying on human evaluation. The new evaluation metric is relevant to the features of medical decision problems as it is based on diagnostic accuracy and circumvents the problem of high correlation between neighbour pixels as well as the human bias that are present in current evaluation methods. The new method is computationally 'lightweight' compared to other proposed metrics.

2 PREVIOUS WORK ON EVALUATING EXPLANATION METHODS

The evaluation of explanation methods is a relatively recent topic with few systematic approaches (Ancona et al., 2018; Hooker et al., 2018; Adebayo et al., 2018; Fong & Vedaldi, 2017). To our knowledge, Bach et al. (2015) proposed the first quantitative approach to evaluate an explanation method by flipping pixels to their opposite and comparing the decrease in output with the relevance attributed to the pixel by the explanation method. As the authors note, this only works for low-dimensional input. This work was followed up upon in Samek et al. (2016). By dividing high-dimensional images into squares, they make the method feasible for high-dimensional inputs. Squares with high relevance (as measured by the explanation method) consecutively get replaced with noise sampled from the uniform distribution. The difference between the original output and the output for the degraded images indicates the quality of the explanation method.

Ancona et al. (2018) proposed a different approach to evaluate explanation methods, called Sensitivity-*n*, based on the notion that the decrease in output when several inputs are cancelled out should be equal to the sum of their relevances. As this is based on the assumption that input dimensions are not correlated, it is questionable for high resolution medical images.

Hooker et al. (2018) proposes a quantitative approach to evaluate explanation methods, ROAR. For each explanation method, they extract the relevance maps over the entire training set. They degrade the training set by setting different percentages of the pixels with the highest relevance to the mean and retrain the network. Each retrained network is evaluated on the test set. The accuracy on the test set decreases dependent on the percentage of pixels set to the mean. Compared to our method, ROAR requires more computational resources in the order of magnitudes. Following the authors suggestions requires retraining ResNet50 25 times for each method. As a rough estimate, evaluating all methods we considered with ROAR would take 241 days using eight GPUs [1], making it infeasible for most research groups. Due to this we will not consider ROAR further in this report.

3 Evaluating explanation methods quantitatively with IROF

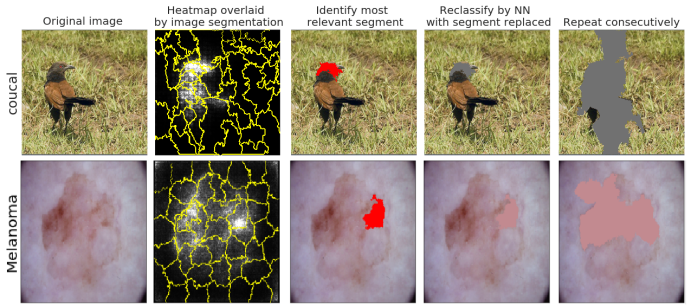


Figure 1: Quantitative evaluation with IROF: Relevant segments as identified by an explanation method get consecutively replaced by the mean colour over the entire dataset. The IROF score of an explanation method is the integrated decrease in the class score over the number of removed segments.

Quantitative evaluation is a recurring problem with explainability methods. This is especially true for high-dimensional input, such as images, where important features consist of locally highly correlated pixels. If the information in one pixel is lost, this loss will not change the overall feature and should not result in a changed output score. The relevance values of single pixels are not indicative of the feature's importance as a whole. We circumvent this problem by utilizing conventional image segmentation. First dividing the image into coherent segments bypasses the interdependency between the inputs.

IROF methodology We assume a neural network $F : X \mapsto y$ with $X \in \mathcal{R}^{m \times m}$ and a set of explanation methods $\{e_j\}_{j=1}^{J}$ with $e_j : X, y, F \mapsto E$ with $E \in \mathcal{R}^{m \times m}$. Without loss of generality we assume a quadratic image with width and height m. Furthermore we partition each image X_n into a set of segments $\{S_n^l\}_{l=1}^{L}$ using a given segmentation method with $s_{n,i,j}^l = 1$ indicating that pixel $x_{n,i,j}$ belongs to segment l. Computing the mean importance $\frac{||E_{j,n} S_n^l||_1}{||S_n^l||_1}$ of each segment according to a given explanation method j, two segments can be compared with each other. We then sort the segments in decreasing order of importance according to the explanation method.

We use $X_n^{'l}$ to indicate X_n with the l segments with highest mean relevance replaced with the mean value. Computing $F(X_n^{'l})_y$ repeatedly with increasing $l \in 0, ..., L$ results in a curve of the class score dependent on how many segments of the image are removed. Dividing this curve by $F(X_n^{'0})_y$

[1] Training ResNet50 with 8 Tesla P100 GPUs takes 29 hours according to Goyal et al. (2017)

normalizes the scores to be within $[0, 1]$ and makes curves comparable between input samples and networks. A good explanation method will attribute high relevance to segments important for classification. As segments with high relevance are removed first, the score for the target class will decrease faster. By computing the area over the curve (AOC) for the class score curve and averaging over many input samples, we can score the methods according to how reliably they identify relevant areas of the image input. For a good explanation method, the AOC will be higher. We refer to this evaluation method as the **iterative removal of features (IROF)**. The IROF score for a given explanation method e_j is expressed as:

$$\text{IROF}(e_j) = \frac{1}{N}\sum_{n=1}^{N} \text{AOC}\left(\frac{F(X_n'^{l})_y}{F(X_n'^{0})_y}\right)_{l=0}^{L} \tag{1}$$

This approach is a quantitative comparison of two or more explainability methods that does not rely on human evaluation or alignment between human and NN reasoning. For each explanation method the work-flow produces a single value, enabling convenient comparison between multiple explanation methods. A higher IROF score indicates that more information about the classification was captured.

IROF is dependent on having meaningful segments in the input, as natural images do. Dividing up text or non-natural images such as EEG into meaningful and independent segments does not have a natural solution and is left for future research.

4 Experiments

We tested our method on five neural network architectures trained on ImageNet. Details are in the appendix. In addition we present results for a medical task, diagnosing skin diseases, in section 4.3. We compared Saliency (SM), Guided Backpropagation (GB), SmoothGrad (SG), Grad-CAM (GC) and Integrated Gradients (IG) to have a selection of attribution-based methods Simonyan et al. (2013); Selvaraju et al. (2017); Smilkov et al. (2017); Springenberg et al. (2014); Sundararajan et al. (2017). We use SLIC for image segmentation due to availability and quick run time Achanta et al. (2012). Preliminary experiments with Quickshift showed similar results (Vedaldi & Soatto, 2008). We therefore hypothesize that IROF is not sensitive to the choice of segmentation algorithm, although a more thorough study with more segmentation algorithms would be needed to confirm this.

4.1 Evaluating IROF for validity as an evaluation method

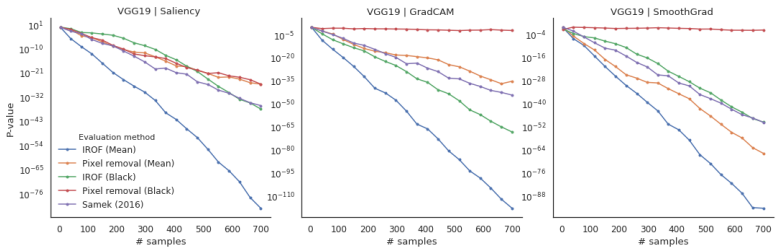


Figure 2: P-values for the rejection of the random removal null-hypothesis. Lower is better. P-values are on a logarithmic scale. IROF performs best in all scenarios.

A good evaluation method should be able to reject the null hypothesis (a given explanation method is no better than random choice) with high confidence. Motivated by this, we compare IROF by calculating the paired t-test of an explanation method versus random guessing. This comparison is done with multiple explanation methods and networks, to reduce the impact of the explanation method.

We compare IROF and pixel removal with mean value and black as a replacement value respectively and additionally against Samek et al. (2016). For IROF and Samek et al. (2016) we set the 10%

most relevant segments to the mean value over the dataset. For pixel removal, we set the equivalent number of pixels to the mean value. The percentage of segments or pixels being removed was chosen ad-hoc. If the difference in degradation between random choice and the explanation method is high, the explanation method reports meaningful information. Since we compare the same explanation method on the same neural network with different evaluation methods, the p-values only contain information about how meaningful the evaluation method is.

Results are shown in fig. 2 (extended in section 6.2). In table 1 we provide results for forty images in tabular form (other methods in section 6.2). On forty images, all evaluation methods produce p-values below 0.05. However, IROF can reject the null hypothesis (this explanation method does not contain any information), with much higher confidence with the same number of samples for any configuration. Thus, IROF is more sensitive to the explanation method than pixel removal or Samek et al. (2016), making it the better choice to quantitatively evaluate an explanation method.

Table 1: t-test: p-values of Random choice vs Saliency Mapping on forty images. All p-values < 0.05.

EVALUATION METHOD	T-STAT	P-VAL
IROF (BLACK)	1.58	1.22E-01
IROF (MEAN)	5.44	3.60E-06
PIXEL FLIPPING (BLACK)	1.92	6.22E-02
PIXEL FLIPPING (MEAN)	2.10	4.25E-02
SAMEK (2016)	2.77	8.65E-03

Especially compared to Samek et al. (2016), IROF has several advantages. By taking natural image features into account via segmentation, it circumvents the problem of high local correlation between pixels belonging to the same feature. By normalizing the degraded outputs with the original output we decrease the dependency of IROF on the original outputs and consequently on the quality of the trained neural network and the clarity of the images. Furthermore, replacement with the mean value does not move an image as far from the input distribution as replacement with uniform noise, meaning that we do not measure how sensitive a neural network is to out-of-distribution input. Thus, IROF is a more sensitive way to evaluate explanations for neural networks than pixel removal.

4.2 EVALUATING EXPLANATION METHODS WITH IROF

We apply IROF on multiple neural network architectures trained on ImageNet with a hundred randomly chosen correctly classified images from the test set and show the results in fig. 3. To check we are not simply measuring vulnerability of the neural network to degradation of the input we include two non-informative baselines, *Random* (randomly chooses segments to remove) and *Sobel* (applying Sobel edge detection on the image and using the extracted edges as the relevance heatmap).

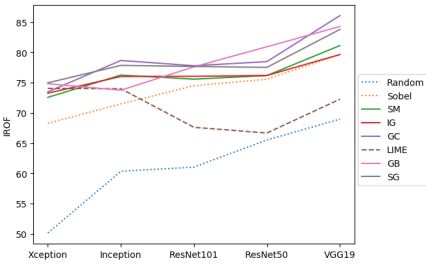


Figure 3: IROF scores for different methods and networks. Higher is better. Dotted are baselines without any information about weights. The networks are ordered according to accuracy from high to low.

Additionally we compared against LIME as a method that is not based on attribution but on local approximation of the neural network Ribeiro et al. (2016). The results in fig. 3 (numbers in table 6) show several important insights:

Explanation methods capture relevant information: All explanation methods have a higher IROF than the random baseline on all architectures tested. Except for LIME, all methods also surpass the stronger baseline, *Sobel*, indicating that all backpropagation-based methods capture meaningful information on the classification.

Explainability seems to be inversely correlated to accuracy: The IROF scores for all methods except LIME for a particular network architectures are strongly correlated. The networks in fig. 3 are ordered according to accuracy from low to high. Higher IROF scores for a network, indicating easier interpretability, are correlated with lower accuracies. Especially in a medical context where

transparency is important, this implies a trade-off between performance and interpretability. Measuring the interpretability of a neural network with IROF could provide an objective measure alongside accuracy.

Explanation methods are relatively constant in their ordering: Though the individual IROF values vary, the ranking of explainability methods between networks is relatively constant, with GradCAM and SmoothGrad outperforming Integrated Gradients and Saliency Maps. We attribute the strong variance of LIME to the relatively high uncertainty due to random sampling that was already noted in Zhang et al. (2019). As shown in Adebayo et al. (2018) the other exception, Guided Backprop (GP), is resistant to randomization of the neural network weights, supporting our finding here. In section 4.3 we investigate whether this constant ordering is due to testing all methods on the same dataset or due to inherent properties of the explanation method.

4.3 Evaluating explanation methods on ISIC skin cancer

Table 2: IROF scores on a neural network trained for skin disease classification. Higher is better.

Method	Random	Sobel	LIME	SM	IG	GB	GC	SG
Xception	19.2	30.3	23.2	25.7	29.2	33.0	35.2	**35.3**

To investigate the validity of IROF in a medical context, we consider the classification task of the ISIC skin lesion challenge Codella et al. (2019). An Xception architecture is fine-tuned to classify seven skin diseases based on dermoscopic images (details in section 6.4). Results are shown in table 2. Despite the task being very different from ImageNet, the ranking of the respective methods is identical to the ImageNet task. SmoothGrad and GradCAM perform best, indicating that they would be most useful for aiding medical professionals. In contrast to the previous experiment, the Sobel edge detector outperforms both Saliency Mapping and Integrated Gradients. Given that Saliency Mapping is often used to check the validity of neural networks for medical tasks, this is an interesting result Esteva et al. (2017). Especially in a low-resource environment, more reliable explanation methods can be used to aid cooperation between doctor and machine.

However, since we only finetuned the network weights with a small dataset, more experiments are needed to verify this result. Additionally, IROF scores are substantially lower on the ISIC task than on the Imagenet task. This is not surprising, as it implies that diagnosis of skin lesions is harder to interpret than the classification of natural images.

5 Conclusion

The adoptions of health AI and realizing the huge potential in low and middle income countries is contingent on performance, usability and trust. In this context explainability is of paramount importance. However, the proposed schemes for explaining, e.g., visual diagnostics based on deep learning, currently come without a satisfying evaluation metric.

In this work we propose a novel way of evaluation for explanation methods that circumvents the problem of high correlation between pixels and does not rely on visual inspection by humans. To our knowledge, this is the first work that systematically evaluates the evaluation metric at hand. We found that IROF is more reliable than previous metrics while needing less resources. We evaluate several explanation methods on multiple architectures with IROF. The results imply a trade-off between accuracy and interpretability. We show evidence supporting the use of SmoothGrad (an aggregated version of Saliency Maps) over Saliency Maps. This was particularly pronounced for medical diagnosis.

Finally, we suggest the use of IROF for evaluating the inherent interpretability of a network architecture, giving practitioners an objective and intuitive metric to weigh against accuracy when deciding on neural network architectures for high-stakes tasks.

REFERENCES

Radhakrishna Achanta, Appu Shaji, Kevin Smith, Aurelien Lucchi, Pascal Fua, and Sabine Süsstrunk. Slic superpixels compared to state-of-the-art superpixel methods. *IEEE transactions on pattern analysis and machine intelligence*, 34(11):2274–2282, 2012.

Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. Sanity checks for saliency maps. In *Advances in Neural Information Processing Systems*, pp. 9525–9536, 2018.

Marco Ancona, Enea Ceolini, Cengiz Oztireli, and Markus Gross. Towards better understanding of gradient-based attribution methods for deep neural networks. In *6th International Conference on Learning Representations (ICLR 2018)*, 2018.

Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller, and Wojciech Samek. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PloS one*, 10(7):e0130140, 2015.

Francois Chollet. Xception: Deep learning with depthwise separable convolutions. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1800–1807. IEEE, 2017.

Noel Codella, Veronica Rotemberg, Philipp Tschandl, M Emre Celebi, Stephen Dusza, David Gutman, Brian Helba, Aadi Kalloo, Konstantinos Liopyris, Michael Marchetti, et al. Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (isic). *arXiv preprint arXiv:1902.03368*, 2019.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pp. 248–255. Ieee, 2009.

Andre Esteva, Brett Kuprel, Roberto A Novoa, Justin Ko, Susan M Swetter, Helen M Blau, and Sebastian Thrun. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, 542(7639):115, 2017.

Ruth C Fong and Andrea Vedaldi. Interpretable explanations of black boxes by meaningful perturbation. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 3429–3437, 2017.

Priya Goyal, Piotr Dollár, Ross Girshick, Pieter Noordhuis, Lukasz Wesolowski, Aapo Kyrola, Andrew Tulloch, Yangqing Jia, and Kaiming He. Accurate, large minibatch sgd: Training imagenet in 1 hour. *arXiv preprint arXiv:1706.02677*, 2017.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Sara Hooker, Dumitru Erhan, Pieter-Jan Kindermans, and Been Kim. Evaluating feature importance estimates. *arXiv preprint arXiv:1806.10758*, 2018.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. Why should i trust you?: Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 1135–1144. ACM, 2016.

Wojciech Samek, Alexander Binder, Grégoire Montavon, Sebastian Lapuschkin, and Klaus-Robert Müller. Evaluating the visualization of what a deep neural network has learned. *IEEE transactions on neural networks and learning systems*, 28(11):2660–2673, 2016.

Wojciech Samek, Grégoire Montavon, Andrea Vedaldi, Lars Kai Hansen, and Klaus-Robert Muller. *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*. Springer Nature, 2019.

Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 618–626. IEEE, 2017.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps. dec 2013. URL `http://arxiv.org/abs/1312.6034`.

Daniel Smilkov, Nikhil Thorat, Been Kim, Fernanda Viégas, and Martin Wattenberg. Smoothgrad: removing noise by adding noise. *arXiv preprint arXiv:1706.03825*, 2017.

JT Springenberg, A Dosovitskiy, T Brox, and M Riedmiller. Striving for simplicity: The all convolutional net. In *ICLR (workshop track)*, 2014.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *International Conference on Machine Learning*, pp. 3319–3328, 2017.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2818–2826, 2016.

Philipp Tschandl, Cliff Rosendahl, and Harald Kittler. The ham10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Scientific data*, 5:180161, 2018.

Andrea Vedaldi and Stefano Soatto. Quick shift and kernel methods for mode seeking. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 5305 LNCS, pp. 705–718, Berlin, Heidelberg, oct 2008. Springer Berlin Heidelberg. ISBN 3540886923. doi: 10.1007/978-3-540-88693-8-52.

Brian Wahl, Aline Cossy-Gantner, Stefan Germann, and Nina R Schwalbe. Artificial intelligence (ai) and global health: how can ai contribute to health in resource-poor settings? *BMJ global health*, 3 (4):e000798, 2018.

Yujia Zhang, Kuangyan Song, Madeilene Udell, and Yiming Sun. "why should you trust my explanation?" understanding uncertainty in lime explanations. 2019.

6 APPENDIX

6.1 EXPERIMENTAL SETUP

6.1.1 GENERAL

We use SLIC for image segmentation due to availability and quick run time Achanta et al. (2012). Preliminary experiments with Quickshift showed similar results (Vedaldi & Soatto, 2008). SLIC was chosen over Quickshift due to the quicker run time. The number of segments was set to 300 ad hoc. fig. 1 shows the same procedure with 100 segments for the sake of clarity.

Some of the methods result in positive and negative evidence. We only conssdered positive evidence for the ImageNet tasks to compare methods against each other. To check that this does not corrupt the methods, we compared the methods that do contain negative results against their filtered version and found negligible difference between the two versions of a method in the used metrics.

6.1.2 IMAGENET

We tested our method on five network architectures that were pre-trained on ImageNet: VGG19, Xception, Inception, ResNet50 and ResNet101 Deng et al. (2009); Simonyan & Zisserman (2014); He et al. (2016); Chollet (2017); Szegedy et al. (2016)[2]. Due to the softmax non-linearity commonly used in the last layer of neural networks, the output for all classes sum up to one, i.e. there is always at least one class with output greater than zero.

We downloaded the data from the ImageNet Large Scale Visual Recognition Challenge website and used the validation set only. No images were excluded. The images were preprocessed to be within $[-1, 1]$ unless a custom range was used for training (indicated by the preprocess function of keras).

The dataset consists of images from 1000 non-overlapping categories such as 'slug', 'pelican' or 'soccer ball'. Each image contains one and only one class.

[2]Models retrieved from https://github.com/keras-team/keras.

6.2 EVALUATING THE EVALUATION

We report p-values for evaluating with 50 images on ResNet101 in the manner described in section 4.1 in tabular form to provide a clear overview.

Table 3: t-test p-values of explanation methods for SmoothGrad.

EVALUATION METHOD	T STATISTIC	P-VALUE
IROF (BLACK)	3.97	3.22E-04
IROF (MEAN)	5.99	6.42E-07
PIXEL FLIPPING (BLACK)	0.55	5.86E-01
PIXEL FLIPPING (MEAN)	5.00	1.39E-05
SAMEK (2016)	2.97	5.17E-03

Table 4: t-test p-values of explanation methods for Saliency.

EVALUATION METHOD	T STATISTIC	P-VALUE
IROF (BLACK)	3.97	3.22E-04
IROF (MEAN)	5.99	6.42E-07
PIXEL FLIPPING (BLACK)	0.55	5.86E-01
PIXEL FLIPPING (MEAN)	5.00	1.39E-05
SAMEK (2015)	2.97	5.17E-03

Table 5: t-test p-values of explanation methods for GradCAM.

EVALUATION METHOD	T STATISTIC	P-VALUE
IROF (BLACK)	4.73	3.23E-05
IROF (MEAN)	7.81	2.42E-09
PIXEL FLIPPING (BLACK)	-1.75	8.85E-02
PIXEL FLIPPING (MEAN)	3.26	2.38E-03
SAMEK (2016)	3.32	2.04E-03

We provide an extended version of fig. 2.

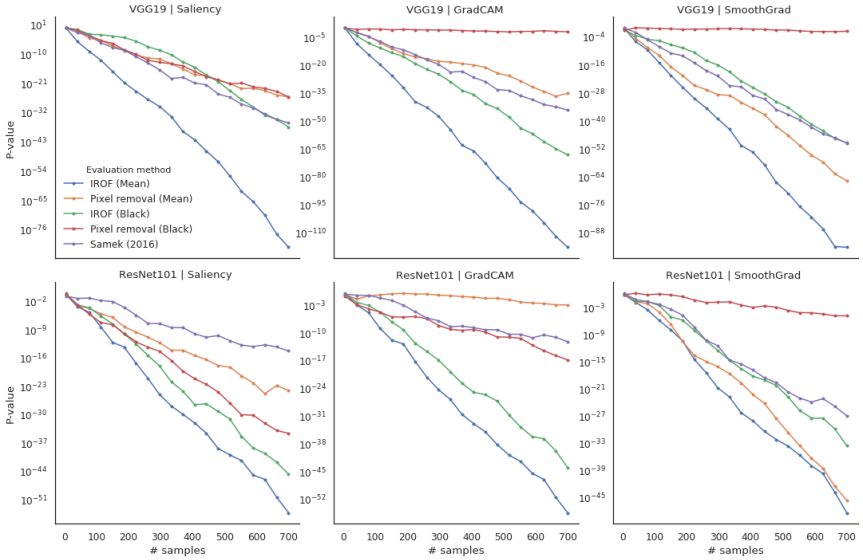


Figure 4: P-values for the rejection of the random removal null-hypothesis. Extended graphs

6.3 IROF SCORES FOR EVALUATION METHODS

In the main script we visualize the IROF scores in a plot, since the comparison of eight methods over five networks is hard to comprehend in a table. We supply the numerical values in table 6.

Table 6: IROF scores across methods and architectures. All SE < 0.05.

METHOD	INCEPTION	RESNET101	RESNET50	VGG19	XCEPTION
RANDOM	60.3	61.0	65.5	68.4	50.1
SOBEL	71.5	74.5	75.6	79.6	68.3
LIME	74.0	67.6	66.7	73.4	74.0
SM	76.2	75.6	76.2	81.4	72.6
IG	76.0	76.0	76.2	79.8	73.3
SG	77.9	77.7	77.5	83.9	**75.0**
GC	**78.7**	**77.8**	**78.5**	**86.1**	73.5

6.4 ISIC 2018: SKIN LESION ANALYSIS TOWARDS MELANOMA DETECTION

6.4.1 EXPERIMENTAL DETAILS

We trained an Xception architecture, using the pretrained weights and retraining the last layer. Since our goal was not to reach state of the art, we did not use any fine-tuning. On the test set, we reached 66.4% accuracy (Macro ROC AUC: 0.81). The network is trained to distinguish seven disease categories: Melanoma, Melanocytic nevus, Basal cell carcinoma, Actinic keratosis / Bowen's disease (intraepithelial carcinoma), Benign keratosis (solar lentigo / seborrheic keratosis / lichen planus-like keratosis), Dermatofibromam, and Vascular lesion.

6.4.2 EXAMPLE IMAGES

We show example images from the test set of Codella et al. (2019); Tschandl et al. (2018)

fs

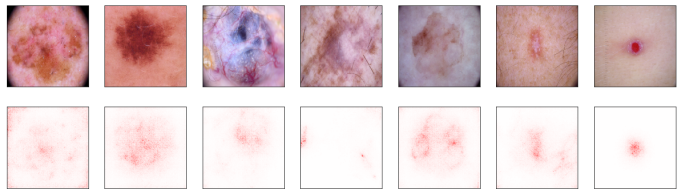


Figure 5: Upper row: Example images for the skin diseases. Lower row: SmoothGrad visualizations (best performing) for respective images. Lesions are marked as relevant. From left to right: Melanoma, Melanocytic nevus, Basal cell carcinoma, Actinic keratosis / Bowen's disease (intraepithelial carcinoma), Benign keratosis (solar lentigo / seborrheic keratosis / lichen planus-like keratosis), Dermatofibromam, and Vascular lesion.

CONTRIBUTION C

A simple defense against adversarial attacks on heatmap explanations

Laura Rieger and Lars Kai Hansen. A simple defense against adversarial attacks on heatmap explanations. In *Workshop on Human Interpretability in Machine Learning (WHI) at ICML*, 2020

A simple defense against adversarial attacks on heatmap explanations

Laura Rieger [1] **Lars Kai Hansen** [1]

Abstract

With machine learning models being used for more sensitive applications, we rely on interpretability methods to prove that no discriminating attributes were used for classification. A potential concern is the so-called "fair-washing" - manipulating a model such that the features used in reality are hidden and more innocuous features are shown to be important instead.

In our work we present an effective defence against such adversarial attacks on neural networks. By a simple aggregation of multiple explanation methods, the network becomes robust against manipulation. This holds even when the attacker has exact knowledge of the model weights and the explanation methods used.

1. Introduction

In recent years machine learning algorithms have become more complex and are used for more important decisions. Since models, especially neural networks, are trained with large amounts of data, it is hard to oversee just what is hidden in the data and what correlations the model picks up on. Explainability methods present a solution for this (Hansen & Rieger, 2019). By looking at what features of the input were important for a classification, we can make sure that the classification is aligned with our ethical convictions and understanding of the task.

It follows that there are many reasons why someone might want to manipulate an explanation, referred to as "fairwashing" (Aïvodji et al., 2019). For example, a company may want to hide that they use discriminatory practices in their hiring or someone may want to hide adversarial attacks on machine learning algorithms. Before explainability methods can be used and relied on in practice, we need to evaluate the the risk for this and find effective defences. Previous works have shown that explainability methods are remarkably brittle against adversarial attacks. In practice, an attacker can effectively manipulate the explanation at will without any visual changes to the input that a human would pick up on (Dombrowski et al., 2019; Ghorbani et al., 2019).

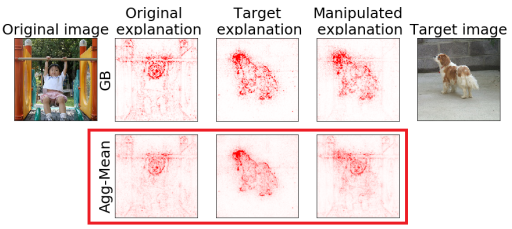


Figure 1. Explanation methods (here Guided Backpropagation) are very vulnerable to adversarial attacks. Our method, AGG-Mean, presents a simple but effective defence.

We propose a simple way to ward against this potential security risk and make explainability methods more viable in deployment. Our approach is motivated by a key insight in machine learning: Ensemble models can reduce both bias and variance compared to applying a single model. A related approach was pursued for *functional* visualization in neuroimaging (Hansen et al., 2001). Based on this insight, we propose a way to aggregate explanation methods, *AGG-Mean*. This approach is analysed theoretically and evaluated empirically. In experiments on Imagenet, the aggregate is more robust to adversarial attacks than any single method. Even when the attacker has complete knowledge of the model weights and the explanation methods to be used as well as complete control over the input, the explanation stays robust as shown in Fig. 1.

2. Related Work

2.1. Explanation methods

The open problem of explainability is reflected in a lot of recent work (Kindermans et al., 2017; Selvaraju et al., 2017; Bach et al., 2015; Zhang et al., 2018a; Zhou et al., 2016; Ancona et al., 2018; Ribeiro et al., 2016; Rieger et al., 2018; Kim et al., 2018; Lundberg & Lee, 2017; Zintgraf et al.,

[1]DTU Compute, Technical University Denmark, 2800 Kgs. Lyngby, Denmark. Correspondence to: Laura Rieger <lauri@dtu.dk>.

Accepted at 2020 Workshop on Human Interpretability in Machine Learning (WHI), Vienna, Austria, PMLR 108, 2020.

2017; Simonyan et al., 2013; Zeiler & Fergus, 2014; Selvaraju et al., 2017; Smilkov et al., 2017; Sundararajan et al., 2017; Shrikumar et al., 2017; Montavon et al., 2017; Chang et al., 2018). We focus on generating visual explanations for single samples. To our knowledge the first work in this direction was Simonyan et al. (2013) with *Saliency Maps (SM)* that proposed backpropagating the output onto the input to gain an understanding of a neural network decision. The relevance for each input dimension is extracted by taking the gradient of the output w. r. t. to the input. This idea was extended by Springenberg et al. (2014) into *Guided Backpropagation (GM)* by applying ReLU non-linearities after each layer during the backpropagation. Compared to Saliency, this removes visual noise in the explanation. *Grad-CAM (GC)* from Selvaraju et al. (2017) is an explanation method, developed for use with convolutional neural networks. By backpropagating relevance through the dense layers and up-sampling the evidence for the convolutional part of the network, the method obtains coarse heatmaps that highlight relevant parts of the input image. *Integrated Gradients (IG)* Sundararajan et al. (2017) sums up the gradients from linearly interpolated pictures between a baseline, e.g. a black image, and the actual image. *SmoothGrad (SG)* filters out noise from a basic saliency map by creating many samples of the original input with Gaussian noise (Smilkov et al., 2017). The final saliency map is the average over all samples. In concurrent work, (Bhatt et al., 2020) also proposed aggregating explanation methods, albeit with the goal of decreasing complexity rather than vulnerability. Finally, (Yeh et al., 2019) showed that a combination of two popular explanation methods is optimal in terms of fidelity.

2.2. Adversarial attacks on explanation methods

While adversarial examples for classification are well-known, recently there has been growing interest in adversarial manipulation of explanations (Ghorbani et al., 2019; Heo et al., 2019; Dombrowski et al., 2019). Attacks on explanation can serve multiple purposes including "fairwashing" (Aïvodji et al., 2019). All of these methods exploit the fully differentiable nature of neural networks and iteratively update the input (or the model weights) to change the explanation while only minimally changing the input and output. The goal is to manipulate the explanation while keeping the input and output (visually) similar. It is assumed that the network architecture and weights are known and that either the input ((Ghorbani et al., 2019; Zhang et al., 2018b; Dombrowski et al., 2019)) or the network weights (Heo et al., 2019) can be changed by the attacker.

Focussing on changing the input, Ghorbani et al. (2019),Zhang et al. (2018b) and Dombrowski et al. (2019) attack the explanation by manipulating the image, not changing the network weights. Interestingly, Zhang et al. (2018b) discuss the transferability of attacks and conclude that attacks are not that transferable. If the attacker is allowed to modify networks weights, as in Heo et al. (2019), the attacks generalize to all the considered explanation methods. We are interested in the more realistic situation where the attacker can modify the input but not the network. We investigate the transferability, c.f., Zhang et al. (2018b), and hypothesise that the limited transferability leads to improved robustness of the ensemble explanation. While ensemble methods have been proposed earlier as a defense for attacks on the label (Tramèr et al., 2017), they have not previously been investigated as a defense mechanism against attacks on explanations.

3. Averaging explanation methods to reduce vulnerability

3.1. Averaging explanations

All currently available explanation methods have weaknesses that are inherent to the approach and include significant uncertainty in the resulting heatmap (Kindermans et al., 2017; Adebayo et al., 2018; Smilkov et al., 2017). A natural way to mitigate this issue and reduce noise is to combine multiple explanation methods. Ensemble methods have been used for a long time to reduce the variance and bias of machine learning models. We apply the same idea to explanation methods and build an ensemble of explanation methods. Ensemble methods have also been previously used to defend against adversarial attacks on neural network outputs (Pang et al., 2019; Liao et al., 2018), motivating the usage of an explanation ensemble to defend against attacks on the explanation.

We assume a neural network $F : X \mapsto y$ with $X \in \mathbb{R}^{m \times m}$ and a set of explanation methods $\{e_j\}_{j=1}^{J}$ with $e_j : X, y, F \mapsto E$ with $E \in \mathbb{R}^{m \times m}$. We write $E_{j,n}$ for the explanation obtained for X_n with method e_j and denote the mean aggregate explanation as $\bar{e}$ with $\bar{E}_n = \frac{1}{J} \sum_{j=1}^{J} E_{j,n}$. While we assume the input to be an image $\in \mathbb{R}^{m \times m}$, this method is generalizable to inputs of other dimensionalities as well.

To get a theoretical understanding of the aggregation, we hypothesize the existence of a 'true' explanation $\hat{E}_n$. This allows us to quantify the error of an explanation method as the mean squared difference between the 'true' explanation and an explanation procured by an explanation method, i.e. the MSE.

For clarity we subsequently omit the notation for the neural network. We write the error of explanation method j on image X_n as $\text{err}_{j,n} = ||E_{j,n} - \hat{E}_n||^2$ with

$$\text{MSE}(E_j) = \frac{1}{N} \sum_n \text{err}_{j,n}$$

and $\text{MSE}(\bar{E}) = \frac{1}{N}\sum_n ||\bar{E}_n - \hat{E}_n||^2$ is the MSE of the aggregate. The typical error of an explanation method is the mean error over all explanation methods

$$\overline{\text{MSE}} = \frac{1}{J}\sum_j \text{MSE}(E_j).$$

With these definitions we can do a standard bias-variance decomposition (Geman et al., 1992). Accordingly we can show the error of the aggregate will be less that the typical error of explanation methods,

$$\begin{aligned} \overline{\text{MSE}} = & \quad \tfrac{1}{N}\textstyle\sum_n \tfrac{1}{J}\sum_j ||\hat{E}_n - E_{j,n}||^2 & (1)\\ = & \quad \tfrac{1}{N}\textstyle\sum_n ||\hat{E}_n - \bar{E}_n||^2 & (2)\\ & + \tfrac{1}{NJ}\textstyle\sum_{n,j} ||\bar{E}_n - E_{j,n}||^2, \end{aligned}$$

hence,

$$\begin{aligned} \overline{\text{MSE}} = & \quad \tfrac{1}{J}\textstyle\sum_j \tfrac{1}{N}\sum_n ||\bar{E}_n - E_{j,n}||^2 + \text{MSE}(\bar{E}) & (3)\\ \geq & \quad \text{MSE}(\bar{E}). \end{aligned}$$

A detailed calculation is given in the appendix. The error of the aggregate $\text{MSE}(\bar{E})$ is less than the typical error of the participating methods. The difference - a 'variance' term - represents the epistemic uncertainty and only vanishes if all methods produce identical maps. By taking the average over all available explanation methods, we reduce the variance of the explanation compared to using a single method. To obtain this average, we normalize all input heatmaps such that the relevance over all pixels sum up to one. This reflects our initial assumption that all individual explanation methods are equally good estimators. We refer to this approach as *AGG-Mean*.

$$E_{\text{AGG-Mean},n} = \frac{1}{J}\sum_{j=1}^{J} E_{j,n} \quad (4)$$

3.2. Adversarial scenarios

With the increasing interest and practical importance of explainability of neural networks the interest in methods for manipulation and control of explanations is also increasing. A typical scenario is to make imperceptible changes to the input of the neural network such that the output/label is unchanged while the explanation changes according to a given goal. Such effort could, e.g., be used to hide bias or other fairness issues a given classifier might have.

Dombrowski et al. (2019) showed that explanations can be made more robust by replacing the ReLU nonlinearity with a Softplus function. However, this requires changing the network and using a different architecture for classification and explanation, which is highly undesirable as it defeats the purpose of the explanation. The analysis of Zhang et al. (2018b) showed that transferability of attacks is limited, hence, our ensemble of multiple explanations may offer robustness also towards certain types of adversarials.

In the following we will assume an attacker who has full knowledge of the neural network, including the architecture and weights. In contrast to Heo et al. (2019), however, we will assume that the attacker cannot *change* the neural network, following Dombrowski et al. (2019); Ghorbani et al. (2019). Furthermore the attacker has full control over the input to the neural network. The goal is to adversarially manipulate the image according to a predefined objective.

In the following, x will refer to the original image. $\hat{x}$ is the 'target' image. The objective is to produce an adversarial input x' with $x' \approx x$ but the explanation $E_{x'} \approx E_{\hat{x}}$. While we focus on assimilating the explanation map of another input as in Dombrowski et al. (2019), all techniques introduced can be readily adapted to other objectives, f.e. to move the mass center of the explanation.

Exploring the robustness of aggregates of multiple explanation methods we concentrate on the following two scenarios:

Arsenal of explanation methods In this scenario we have a pool of potential explanation methods. The attacker does not have knowledge of what explanation method is used and optimizes for a different explanation method than is used by the defender. The success of the attack depends on how readily an attack of one explanation method translates to another method.

We hypothesize that attacks on explanation methods are fragile and do not translate well across explanation methods, as they exploit locally high variances in the gradient landscape. This hypothesis is examined in Section 3.2.

Aggregation of explanation methods In this more challenging scenario we aggregate multiple explanation methods as described in Eq. (4). The attacker knows the exact explanation methods and ratio going into the mixture and attacks this aggregation.

Many attribution-based methods are utilizing the gradient $\frac{\delta y}{\delta x}(x)$ of the output to create an explanation. Due to the non-linearity of the neural network, the gradient can change rapidly with small distances in input space (Dombrowski et al., 2019; Ghorbani et al., 2019). Attacks on explanation methods exploit this vulnerability.

4. Experiments

We evaluate how robust aggregated methods are against adversarial attacks, compared to unaggregated methods. In all cases we assume that the attacker has full knowledge of the network architecture and weights (white box attack) but cannot change them. However, the attacker has full control

over the input.

Following (Dombrowski et al., 2019) we run experiments on a pretrained VGG16[1] We consider Layerwise Relevance Propagation (LRP), Saliency Mapping (SM), Guided Backprop (GB) and Integrated Gradients (IG) as explanation methods. The latter was not used in the aggregation.

Unless otherwise noted we followed (Dombrowski et al., 2019) in the choice of hyperparameters for attacking explanation methods. In the appendix we show that our defence also works against the attack as proposed in Ghorbani et al. (2019).

Since the ReLU function used in neural networks is not twice differentiable, we replace it with a differentiable approximation, SoftPlus for the iterative creation of the adversarial input. The final manipulated heatmaps are created with the ReLU non-linearity. Further details about the experiments are in the appendix.

We consider the two scenarios introduced in Section 3.2. In all cases, the objective of the attacker is to make the explanation of input $E_{x'} \approx E_{\hat{x}}$ while keeping $x' \approx x$. To do this, the attacker manipulates x'.

We visually confirmed that the adversarial images look similar to the input images and provide examples in the appendix. We measure the difference between the start explanation E_x, target explanation $E_{\hat{x}}$ and adversarial explanation $E_{x'}$ with the MSE (Mean Square Error), the PCC (Pearson Correlation Coefficient) and the top-k intersection with k being ad-hoc set to 10% (Dombrowski et al., 2019; Ghorbani et al., 2019).

In all metrics, explanations obtained with different methods have different 'base' values (similarity between the explanations of two randomly chosen images) due to structural differences between explanation methods. To account for this, we consider for each similarity metric m_{sim} the difference $m_{\text{sim}}(E_{\hat{x}}, E_{x'}) - m_{\text{sim}}(E_{\hat{x}}, E_x)$, i.e. how much *more* similar the attack makes $E_{x'}$ look to $E_{\hat{x}}$. For the MSE, this results in a negative score, since the difference between the target and the attack is less than between the target and the starting point. For all metrics, the ideal score is 0, i.e. the attack did not change the explanation at all. Thus, for MSE a high value is desirable, for PCC and top-k union a low value is desirable.

Transferability of attacks on explanation methods Visually comparing the success of an attack on AGG-Mean on the y-axis compared to unaggregated method (Guided Backprop) on the x-axis. Similarity metrics (*topK* and *PCC*) should be low, MSE should be high for less similarity between target and adversarial. Since for most samples *topK* and *PCC* are lower for AGG-Mean than for Guided Backprop, AGG-Mean is more robust than Guided Backprop. The red dot is the sample visualized in Fig. 4.

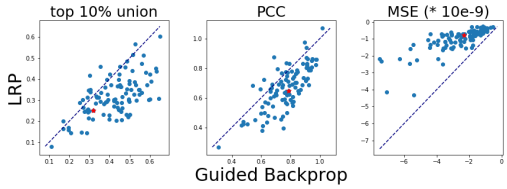


Figure 2. Attacking one method does not translate to attacks on the other methods. Similarity metrics (topK and PCC) should be low, MSE should be high. Since for most samples *topK* and *PCC* are higher for the attacked method (Guided Backprop) than for LRP, attacks on Guided Backprop do not translate well to LRP. Red dot is the single sample visualized in Fig. 3

The lack of transferability results are in line with the findings of (Zhang et al., 2018b).

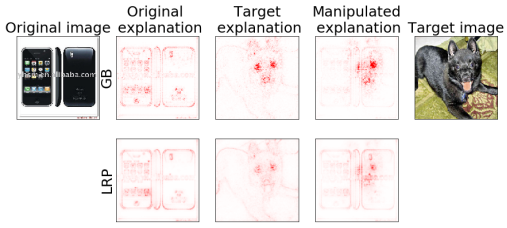


Figure 3. Adversarial attacks do not transfer well between explanation methods. The adversarial input was calculated to attack GB (upper row). We then extracted the explanation with LRP (lower row). The attack does not transfer well to LRP. To visualize details better we clipped values at the 99th percentile.

We consider a case where the attacker does not know what explanation method is used, i.e. we attack a different explanation method than the one that is used. This would be the case if the defender has not made the specific explanation method used public or is choosing one at random to ward off attacks. If the attack translates well, i.e. the image manipulation fools both methods, similarity metrics should be similar for both explanation methods.

In Fig. 2 we provide results for attacking Guided Backpropagation and extracting an explanation with LRP. For a hundred samples we visualize for each sample the respective similarity metrics for both explanation methods in Fig. 2. If the attack translates well, the points should lie on the identity line in Fig. 2. Samples below the identity line for PCC and topK and above for MSE indicate that the attack does not generalize to other explanation methods.

[1]Models retrieved from https://github.com/keras-team/keras.

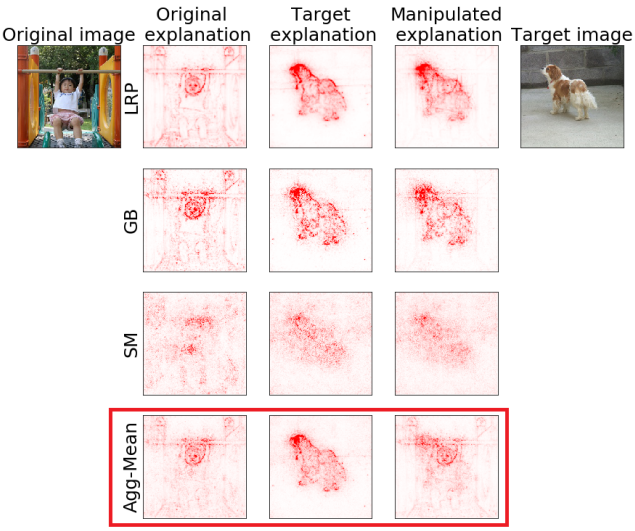


Figure 4. Attacking explanation methods. Each method is targeted individually. AGG-Mean visibly preserves original explanation best, thus being the most resistant to adversarial attacks. To visualize details better we clipped values at the 99 percentile. We provide the adversarial input images in the appendix.

As visible in Fig. 2 and anecdotally in Fig. 3 (red data point in Fig. 2), attacks perform much worse on other methods (here LRP) than the targeted one (here GB). We provide statistics for other combinations in the appendix.

Attacking aggregations of explanation methods In the second scenario the attacker knows that the explanations are aggregated and attacks the aggregation. We aggregate LRP, GB and SM and compare against those methods as well as Integrated Gradient. IG was not included in the aggregation as it requires sampling for each step, making it computationally much more expensive than the other methods.

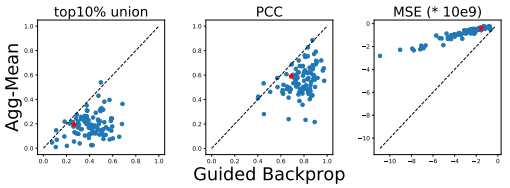


Figure 5. Visually comparing the success of an attack on AGG-Mean (y-axis) to the unaggregated method Guided Backprop (x-axis). Similarity metrics (*topK* and *PCC*) should be low, MSE should be high for less similarity between target and adversarial. Since *topK* and *PCC* are lower for AGG-Mean than for Guided Backprop, AGG-Mean is more robust than Guided Backprop. Red dot is sample visualized in Fig. 4.

In Table 1 we provide metrics averaged over a hundred samples. AGG-Mean outperforms unaggregated methods. We also provide a direct comparison to GuidedBackprop in Fig. 5. To give an intuition on what differences in the metrics look like, we visualize a sample (red dot in Fig. 5) in Fig. 4. We see that AGG-Mean opposed to the unaggregated methods largely preserves the original heatmap after the attack. More examples are provided in the appendix.

The resilience of the aggregate to attacks can be understood in terms of averaging induced smoothness. In (Dombrowski et al., 2019) the beneficial effects of averaging in the SmoothGrad method are described. As noted in (Dombrowski et al., 2019) SmoothGrad is computationally expensive. We conjecture that the diversity of the methods involved in the present aggregate implies that smoothing can be achieved at less computational effort.

5. Conclusion

In recent times, attacks on explanation method have received increased attention as the so-called "fairwashing", manipulating explanations to more innocuous ones, has become a concern.

We provide a simple and intuitive approach to defend against such attacks that does not require the model to be changed in any way and is computationally inexpensive. This approach is explored theoretically. We then provided experimental

Table 1. Evaluation scores across methods and architectures on a hundred samples. *AGG-Mean* surpasses all considered methods in all metrics. All SE ≤ 0.02.

METHOD	MSE (*10E-9)	PCC	TOP 10% UNION
SM	-0.92	0.74	0.40
GB	-3.25	0.77	0.42
LRP	-1.45	0.81	0.49
IG	-1.76	0.82	0.47
AGG-MEAN	**-0.89**	**0.54**	**0.24**

evidence that aggregations are a more robust to adversarial manipulation than individual explanation methods.

Perhaps surprisingly, a simple average with the originally attacked method included induces a more robust explanation than replacing the explanation method with a different one. In (Dombrowski et al., 2019) arguments are presented that the observed vulnerability is due to non-smoothness of contemporary networks. It is also argued that averaging as in SmoothGrad increases robustness. We theorize that the averaging of the diverse set of explanation methods involved in the aggregate creates similar smoothness. We noted that in contrast to (Dombrowski et al., 2019), the aggregate does not require modification (smoothing) of the network.

We hope that our approach will be useful to make neural networks more transparent and increase their credibility as they are applied in real-life scenarios.

References

Adebayo, J., Gilmer, J., Muelly, M., Goodfellow, I., Hardt, M., and Kim, B. Sanity checks for saliency maps. In *Advances in Neural Information Processing Systems*, pp. 9525–9536, 2018.

Aïvodji, U., Arai, H., Fortineau, O., Gambs, S., Hara, S., and Tapp, A. Fairwashing: the risk of rationalization. *arXiv preprint arXiv:1901.09749*, 2019.

Ancona, M., Ceolini, E., Oztireli, C., and Gross, M. Towards better understanding of gradient-based attribution methods for deep neural networks. In *6th International Conference on Learning Representations (ICLR 2018)*, 2018.

Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K.-R., and Samek, W. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PloS one*, 10(7):e0130140, 2015.

Bhatt, U., Weller, A., and Moura, J. M. Evaluating and aggregating feature-based model explanations. *arXiv preprint arXiv:2005.00631*, 2020.

Chang, C.-H., Creager, E., Goldenberg, A., and Duvenaud, D. Explaining image classifiers by counterfactual generation. 2018.

Dombrowski, A.-K., Alber, M., Anders, C. J., Ackermann, M., Müller, K.-R., and Kessel, P. Explanations can be manipulated and geometry is to blame. *arXiv preprint arXiv:1906.07983*, 2019.

Geman, S., Bienenstock, E., and Doursat, R. Neural networks and the bias/variance dilemma. *Neural computation*, 4(1):1–58, 1992.

Ghorbani, A., Abid, A., and Zou, J. Interpretation of neural networks is fragile. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 3681–3688, 2019.

Hansen, L. K. and Rieger, L. Interpretability in intelligent systems–a new concept? In *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*, pp. 41–49. Springer, 2019.

Hansen, L. K., Nielsen, F. Å., Strother, S. C., and Lange, N. Consensus inference in neuroimaging. *NeuroImage*, 13 (6):1212–1218, 2001.

Heo, J., Joo, S., and Moon, T. Fooling neural network interpretations via adversarial model manipulation. *arXiv preprint arXiv:1902.02041*, 2019.

Kim, B., Wattenberg, M., Gilmer, J., Cai, C., Wexler, J., Viegas, F., et al. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). In *International Conference on Machine Learning*, pp. 2673–2682, 2018.

Kindermans, P.-J., Hooker, S., Adebayo, J., Brain, G., Alber, M., Schütt, K. T., Dähne, S., Erhan, D., and Kim, B. The (un)reliability of saliency methods. In *Proceedings Workshop on Interpreting, Explaining and Visualizing Deep Learning (at NIPS)*, 2017.

Liao, F., Liang, M., Dong, Y., Pang, T., Hu, X., and Zhu, J. Defense against adversarial attacks using high-level representation guided denoiser. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1778–1787, 2018.

Lundberg, S. M. and Lee, S.-I. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*, pp. 4765–4774, 2017.

Mohseni, S. and Ragan, E. D. A human-grounded evaluation benchmark for local explanations of machine learning. *arXiv preprint arXiv:1801.05075*, 2018.

Montavon, G., Lapuschkin, S., Binder, A., Samek, W., and Müller, K.-R. Explaining nonlinear classification decisions with deep taylor decomposition. *Pattern Recognition*, 65:211–222, 2017.

Pang, T., Xu, K., Du, C., Chen, N., and Zhu, J. Improving adversarial robustness via promoting ensemble diversity. *arXiv preprint arXiv:1901.08846*, 2019.

Ribeiro, M. T., Singh, S., and Guestrin, C. Why should i trust you?: Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 1135–1144. ACM, 2016.

Rieger, L., Chormai, P., Montavon, G., Hansen, L. K., and Müller, K.-R. Structuring Neural Networks for More Explainable Predictions. pp. 115–131. Springer, Cham, 2018.

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., and Batra, D. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 618–626. IEEE, 2017.

Shrikumar, A., Greenside, P., and Kundaje, A. Learning important features through propagating activation differences. In *International Conference on Machine Learning*, pp. 3145–3153, 2017.

Simonyan, K. and Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

Simonyan, K., Vedaldi, A., and Zisserman, A. Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps. dec 2013. URL http://arxiv.org/abs/1312.6034.

Smilkov, D., Thorat, N., Kim, B., Viégas, F., and Wattenberg, M. Smoothgrad: removing noise by adding noise. 06 2017.

Springenberg, J., Dosovitskiy, A., Brox, T., and Riedmiller, M. Striving for simplicity: The all convolutional net. In *ICLR (workshop track)*, 2014.

Sundararajan, M., Taly, A., and Yan, Q. Axiomatic attribution for deep networks. In *International Conference on Machine Learning*, pp. 3319–3328, 2017.

Tramèr, F., Kurakin, A., Papernot, N., Goodfellow, I., Boneh, D., and McDaniel, P. Ensemble adversarial training: Attacks and defenses. *arXiv preprint arXiv:1705.07204*, 2017.

Yeh, C.-K., Hsieh, C.-Y., Suggala, A., Inouye, D. I., and Ravikumar, P. K. On the (in) fidelity and sensitivity of explanations. In *Advances in Neural Information Processing Systems*, pp. 10967–10978, 2019.

Zeiler, M. D. and Fergus, R. Visualizing and understanding convolutional networks. In *European Conference on Computer Vision*, pp. 818–833. Springer, 2014.

Zhang, Q., Nian Wu, Y., and Zhu, S.-C. Interpretable convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 8827–8836, 2018a.

Zhang, X., Wang, N., Shen, H., Ji, S., Luo, X., and Wang, T. Interpretable deep learning under fire. *arXiv preprint arXiv:1812.00891*, 2, 2018b.

Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., and Torralba, A. Learning deep features for discriminative localization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2921–2929, 2016.

Zintgraf, L. M., Cohen, T. S., Adel, T., and Welling, M. Visualizing deep neural network decisions: Prediction difference analysis. In *ICLR*, pp. 3, 2017.

A. Appendix

A.1. Aggregating explanation methods to reduce variance - detailed derivation

All currently available explanation methods have weaknesses that are inherent to the approach and include significant noise in the heatmap (Kindermans et al., 2017; Adebayo et al., 2018; Smilkov et al., 2017). A natural way to mitigate this issue and reduce noise is to combine multiple explanation methods. Ensemble methods have been used for a long time to reduce the variance and bias of machine learning models. We apply the same idea to explanation methods and build an ensemble of explanation methods.

We assume a neural network $F : X \mapsto y$ with $X \in \mathbb{R}^{mxm}$ and a set of explanation methods $\{e_j\}_{j=1}^{J}$ with $e_j : X, y, F \mapsto E$ with $E \in \mathbb{R}^{mxm}$. We write $E_{j,n}$ for the explanation obtained for X_n with method e_j and denote the mean aggregate explanation as $\bar{e}$ with $\bar{E}_n = \frac{1}{J}\sum_{j=1}^{J} E_{j,n}$. While we assume the input to be an image $\in R^{mxm}$, this method is generalizable to inputs of other dimensions as well.

We define the error of an explanation method as the mean squared difference between a hypothetical 'true' explanation and an explanation procured by the explanation method, i.e. the MSE. For this definition we assume the existence of the hypothetical 'true' explanation $\hat{E}_n$ for image X_n.

For clarity we subsequently omit the notation for the neural network.

We write the error of explanation method j on image X_n as $err_{j,n} = ||E_{j,n} - \hat{E}_n||^2$ with

$$\text{MSE}(E_j) = \frac{1}{N}\sum_n err_{j,n}$$

and $\text{MSE}(\bar{E}) = \frac{1}{N}\sum_n ||\bar{E}_n - \hat{E}_n||^2$ is the MSE of the aggregate. The typical error of an explanation method is represented by the mean

$$\begin{aligned}
\overline{\text{MSE}} &= \frac{1}{N}\sum_n \frac{1}{J}\sum_j ||\hat{E}_n - E_{j,n}||^2 \\
&= \frac{1}{NJ}\sum_{n,j} ||\hat{E}_n - E_{j,n} + \bar{E}_n - \bar{E}_n||^2 \\
&= \frac{1}{NJ}\sum_{n,j} ||(\hat{E}_n - \bar{E}_n) + (\bar{E}_n - E_{j,n})||^2 \\
&= \frac{1}{NJ}\sum_{n,j} ||\hat{E}_n - \bar{E}_n||^2 + ||\bar{E}_n - E_{j,n}||^2 + \frac{1}{NJ}\sum_{n,j}\left(2\text{Tr}\left[(\hat{E}_n - \bar{E}_n)(\bar{E}_n - E_{j,n})\right]\right) \\
&= \frac{1}{N}\sum_n ||\hat{E}_n - \bar{E}_n||^2 + \frac{1}{NJ}\sum_{n,j} ||\bar{E}_n - E_{j,n}||^2 + 2\frac{1}{N}\sum_n \text{Tr}\left[(\hat{E}_n - \bar{E}_n)\left(\frac{1}{J}\sum_j (\bar{E}_n - E_{j,n})\right)\right] \\
&= \frac{1}{N}\sum_n ||\hat{E}_n - \bar{E}_n||^2 + \frac{1}{NJ}\sum_{n,j} ||\bar{E}_n - E_{j,n}||^2 + 2\frac{1}{N}\sum_n \text{Tr}\left[(\hat{E}_n - \bar{E}_n)\underbrace{\frac{1}{J}\sum_j (\bar{E}_n - E_{j,n})}_{=0}\right] \\
&= \frac{1}{N}\sum_n ||\hat{E}_n - \bar{E}_n||^2 + \frac{1}{NJ}\sum_{n,j} ||\bar{E}_n - E_{j,n}||^2,
\end{aligned}$$

hence,

$$\overline{\text{MSE}} = \text{MSE}(\bar{E}) + \frac{1}{NJ}\sum_{n,j}\underbrace{||\bar{E}_n - E_{j,n}||^2}_{\text{epistemic uncertainty}} \geq \text{MSE}(\bar{E})$$

The error of the aggregate MSE($\bar{E}$) is less than the typical error of the participating methods. The difference - a 'variance' term - represents the epistemic uncertainty and only vanishes if all methods produce identical maps.

A.2. Experimental setup

A.2.1. General

For AGG-Var, we add a constant to the denominator. We set this constant to 10 times the mean std, a value chosen empirically after trying values in the range of $[1, 10, 100]$ times the mean.
Evaluations were run with a set random seed for reproducibility. SE were reported either for each individual result or if they were non-significant in the caption to avoid cluttering the results.
All experiments were done on a Titan X.

A.2.2. ImageNet

We downloaded the data from the ImageNet Large Scale Visual Recognition Challenge website and used the validation set only. No images were excluded. The images were preprocessed to be within $[-1, 1]$ unless a custom range was used for training (indicated by the preprocess function of keras).

A.2.3. Details about attacking explanation methods

For a range of explanation methods we chose to compare against LRP, Gradient, Guided Backpropagation and Integrated Gradients, a range of well-known and well-established explanation methods (Sundararajan et al., 2017; Bach et al., 2015; Springenberg et al., 2014; Simonyan et al., 2013). Since Integrated Gradients is thirty times more computationally expensive than other methods, we did not include it in the aggregation as it would have slowed down experiments considerably.

Unless otherwise noted, all metrics are computed as the average of a hundred data samples with mean and SE. Informally, we also found that the MSE does not align well with perceived changes in the explanations, likely due to it being susceptible to outliers.

We used a pretrained VGG16 for all experiments attacking explanation methods (Simonyan & Zisserman, 2014).

A.3. Alignment between human attribution and explanation methods

We want to quantify whether an explanation method agrees with human judgement on which parts of an image should be important. While human annotation is expensive, there exists a benchmark for human evaluation introduced in (Mohseni & Ragan, 2018). The benchmark includes ninety images of categories in the ImageNet Challenge (ten images were excluded

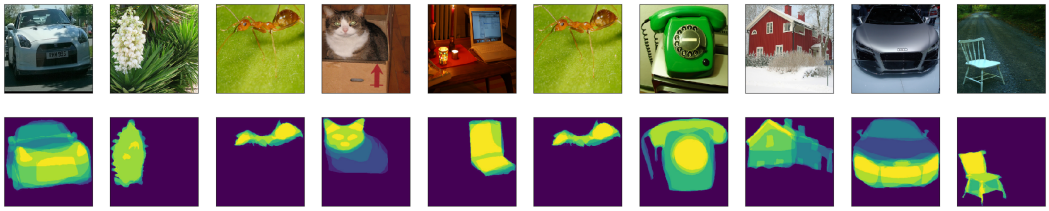


Figure 6. Example images from (Mohseni & Ragan, 2018) with human-annotated overlays.

due to the category not being in the ImageNet challenge) and provides annotations of relevant segments that ten human test objects found important. Example images are shown in Fig. 6.

While human evaluation is not a precise measure, we still expect some correlation between neural network and human judgement.

To test the alignment, we calculate the cosine similarity,

$$\text{similarity}(e_j) = \frac{\sum_{n=1}^{N} A_n E_{j,n}}{\sqrt{\sum_{n=1}^{N} A_n^2}\sqrt{\sum_{n=1}^{N} E_{j,n}^2}}$$

between the human annotation and the explanations produced by the respective explanation methods. A_n is the human annotation of what is important for image X_n

Since the images in this dataset are 224x224 pixel large, we only compute the cosine similarity for the network architectures where pretrained networks with this input size were available.

We see that *AGG-Mean* and *AGG-Var* perform on-par with the best methods (SmoothGrad and GradCAM). While the aggregated methods perform better than the average explanation method, they do not surpass the best method.

When we combine the two best-performing single methods, SmoothGrad and GradCAM, we surpass each individual method. We hypothesize that this is because the epistemic uncertainty is reduced by the aggregate.

Table 2. Cosine similarity between heatmap and human annotated benchmark. All SE below 0.05

METHOD	RESNET101	RESNET50	VGG19
AGG-MEAN	0.63	0.66	0.64
AGG-VAR	0.66	0.68	0.67
GB	0.42	0.49	0.47
GC	0.60	0.62	0.60
IG	0.45	0.45	0.47
MEAN(SG+GC)	**0.69**	**0.70**	**0.65**
SG	0.63	0.65	0.59
SM	0.45	0.45	0.47

A.4. Details about attacking explanation methods

Choice of explanation methods We focused on explanation methods that have previously been shown to be susceptible to adversarial attacks. As such, we did not include GradCAM in the experiments, neither as a comparison or in the aggregation.

Different explanation methods have different computational loads. Notably, SmoothGrad and IntegratedGradients involve the sampling of many explanations for a single pass, increasing computation times by the number of samples () and were not included in the aggregation but as a comparison.

Choice of hyperparameters We followed (Dombrowski et al., 2019) for the choice of hyperparameters in learning rate and beta growth. For AGG-Mean we chose a learning rate of 10^{-3} and 1500 iterations for the attack.

hyperparameter choice when attacking explanation methods, using a learning rate of 10^{-3}.

A.4.1. More examples

We provide more examples showing different explanation methods being attacked in Figs. 7 to 10. An abridged version of Fig. 7 is also shown in the main text.

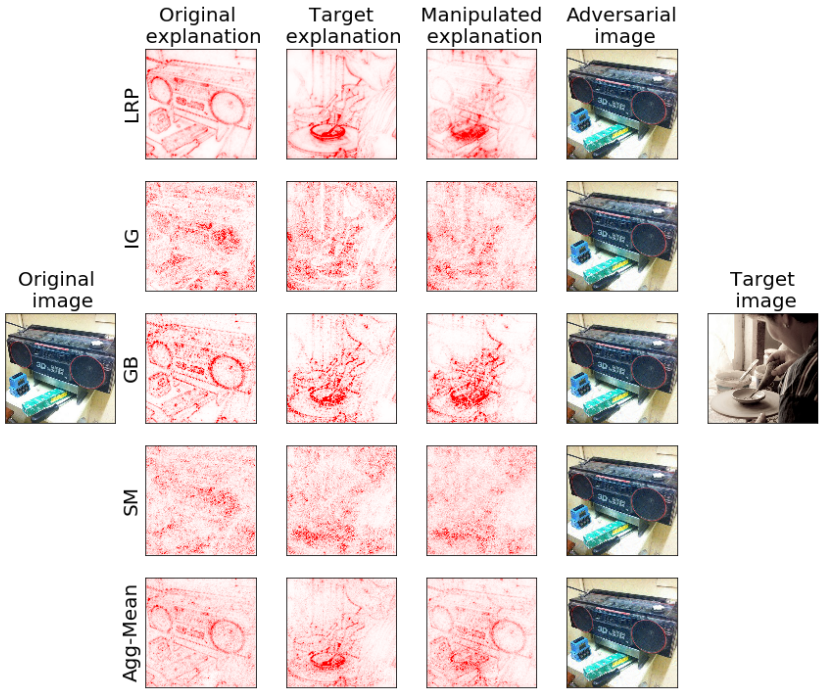


Figure 7. Attack shown in the main text, including the adversarial input images. There are no visual differences for any of the adversarial inputs.

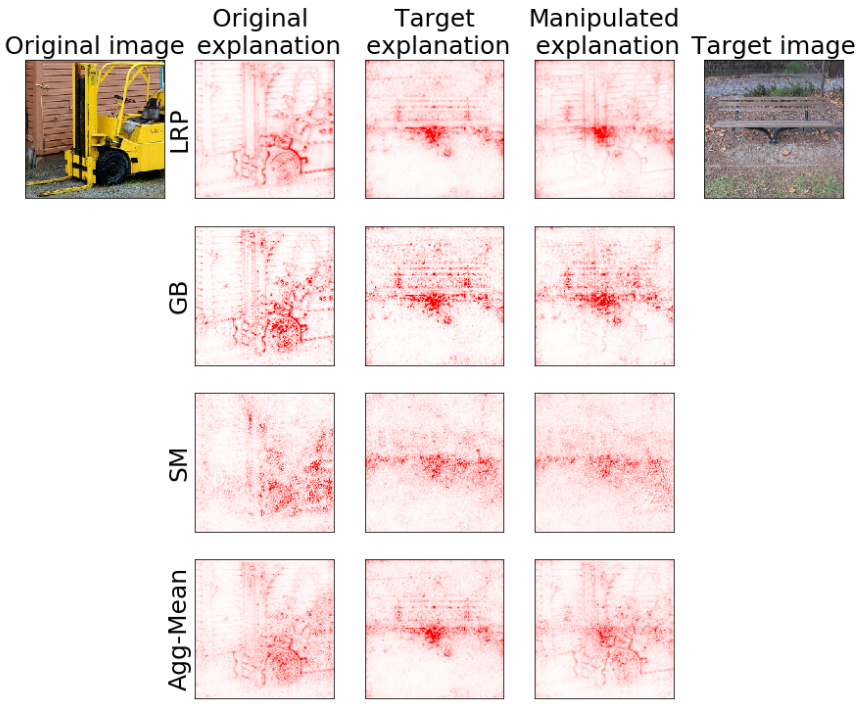


Figure 8. Appendix example 1

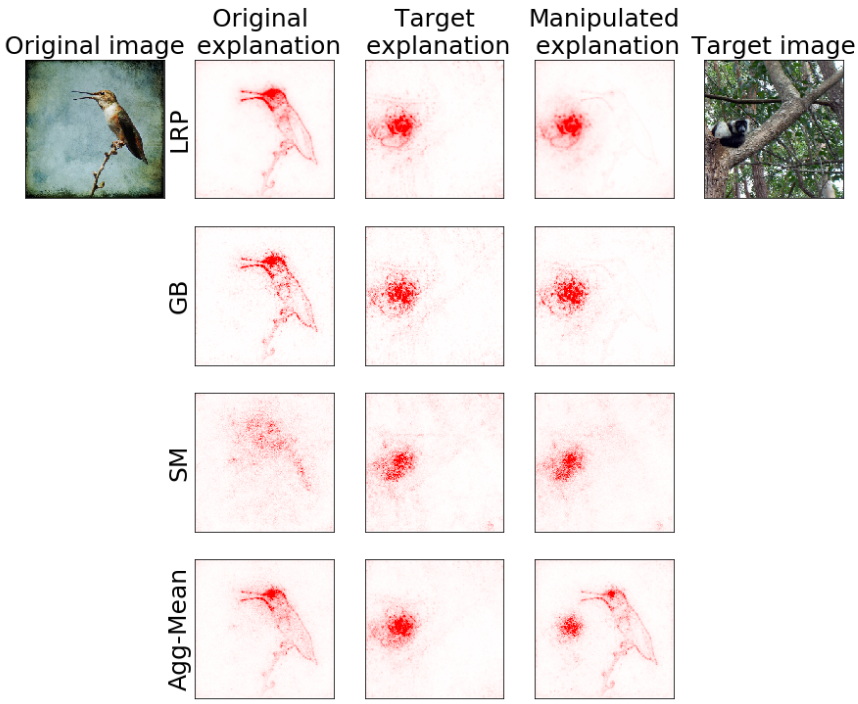


Figure 9. Appendix example 2

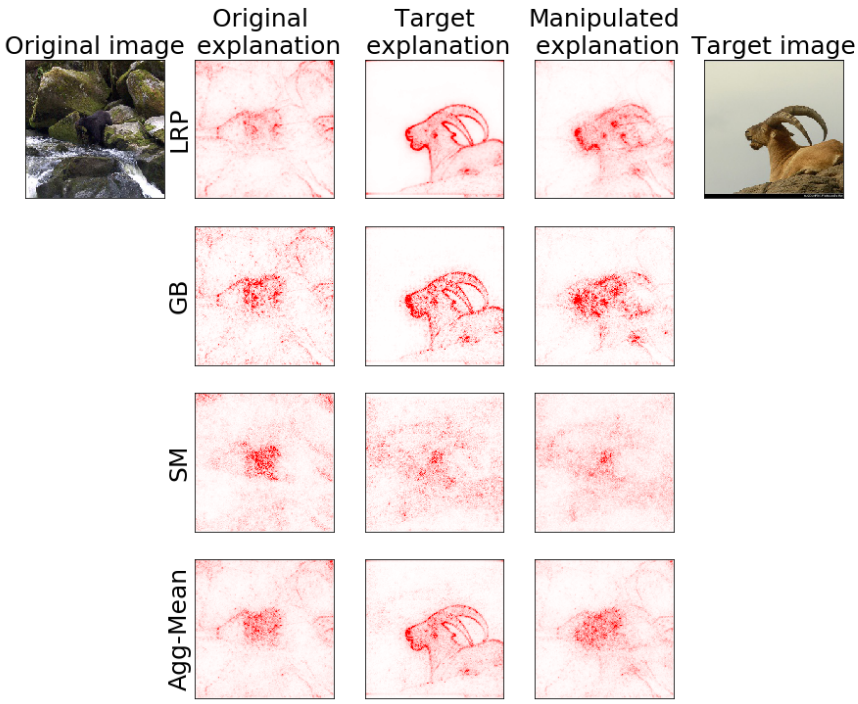


Figure 10. Appendix example 3

A.4.2. TRANSFERABILITY OF ATTACKS

In the main text we show similarity metrics differences between the method being attacked and not being attacked for Guided Backprop and LRP. Here we provide scatter plots for the rest of the considered methods in Figs. 11 to 14:

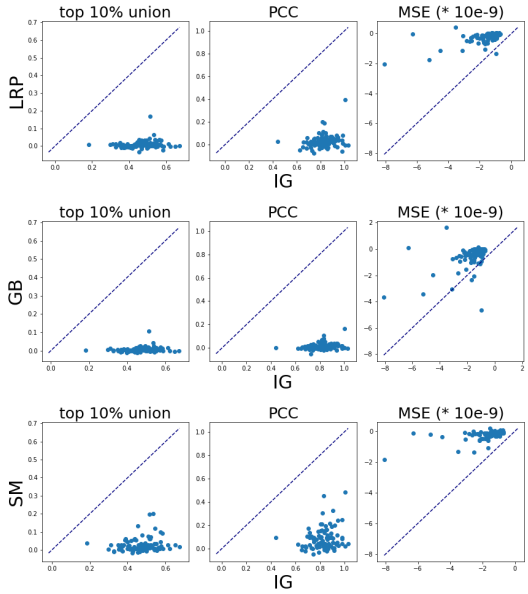


Figure 11. Integrated Gradient as starting method

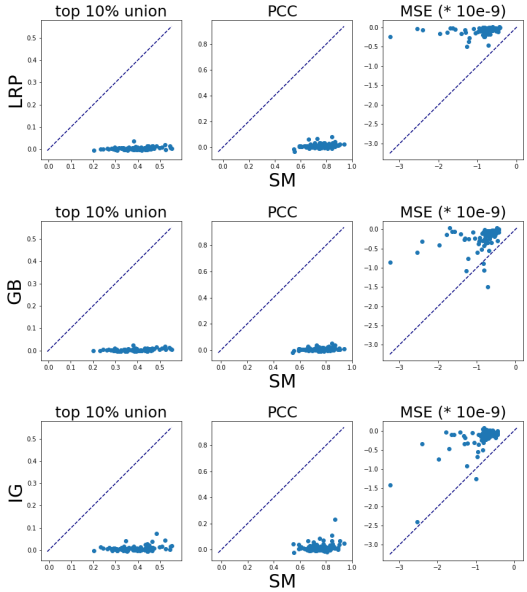


Figure 12. Gradient as starting method

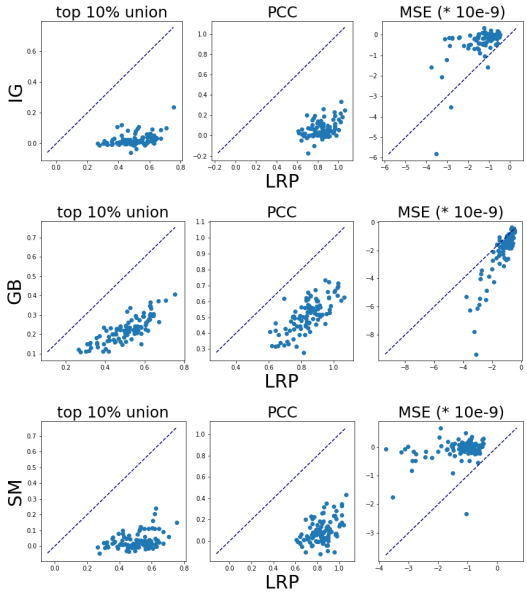


Figure 13. LRP as starting method

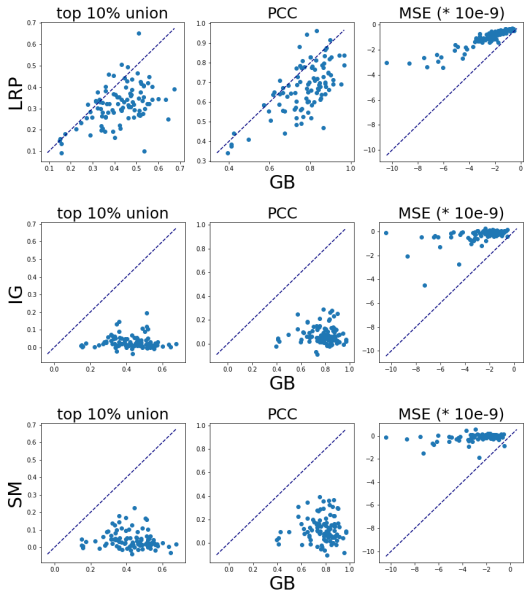


Figure 14. GuidedBackprop as starting method

A.4.3. SIMILARITY OF THE ATTACKED IMAGES TO THE STARTING IMAGES

We provide the average distance of the adversarial images to the original images in Table 3 (calculated in RGB space, average over all pixels). As can be seen in Figs. 7 to 9 and 10, there is no visual difference to the input images for any of the attacks. Attacking *AGG-Mean* has the smallest distance to the input image, supporting our hypothesis that aggregating explanation methods removes vulnerabilities to adversarial manipulation.

Table 3. Evaluation scores across methods and architectures on a hundred samples, including similarity of the resulting image to the starting image. Deviation is SE.

METHOD	MSE Δ (*10E-9)	PCC	TOP 10% UNION	MSE ON IMAGES
SM	-0.92 $\pm$ 0.00	0.74 $\pm$ 0.01	0.40 $\pm$ 0.01	0.0027 $\pm$ 0.0002
GB	-3.25 $\pm$ 0.02	0.77 $\pm$ 0.01	0.42 $\pm$ 0.01	0.0110 $\pm$ 0.0025
LRP	-1.45 $\pm$ 0.01	0.81 $\pm$ 0.01	0.49 $\pm$ 0.01	0.0047 $\pm$ 0.0006
IG	-1.76 $\pm$ 0.01	0.82 $\pm$ 0.01	0.47 $\pm$ 0.01	0.0102 $\pm$ 0.0022
AGG-MEAN	**-0.89** $\pm$ 0.01	**0.54** $\pm$ **0.01**	**0.24** $\pm$ **0.01**	0.0013 $\pm$ 0.0001

A.4.4. Other attacks

In the main text we mainly concern ourselves with making the explanation of one image look like a pre-specified target explanation, as this is a use case where the motivation of an attacker is apparent. However, as introduced in (Ghorbani et al., 2019) other attack objectives are also conceivable.

We show results when following the objective of making a specified area of the explanation not relevant, i.e. a blank space in the explanation as introduced in (Ghorbani et al., 2019). A square (in size a quarter of the image) centered on the middle should not contain any relevance for the explanation. Size and position of the blank space were chosen ad-hoc, we assume that the center of the image generally contains useful information for the classification. We show quantitative results in Table 4, computing how much percentage of the original relevance is preserved and qualitative results in Figs. 15 and 16. While an aggregation is not completely robust to the attack, far more of the original explanation is preserved.

Table 4. Manipulating explanations to show a blank (irrelevant) square. Aggregating explanation methods preserves far more of the original explanation than any single method.

Method	Start	After attack	Percentage
SM	0.34	2.05e-02	0.06
GB	0.41	7.57e-03	0.02
IG	0.38	1.48e-02	0.04
LRP	0.37	4.11e-03	0.01
AGG-Mean	0.37	**5.10e-02**	**0.14**

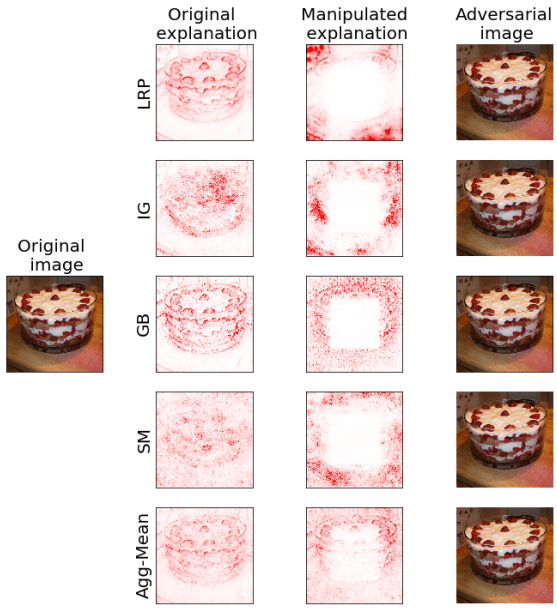


Figure 15. Attacking explanation methods to make an area irrelevant as in (Ghorbani et al., 2019). AGG-Mean is most robust.

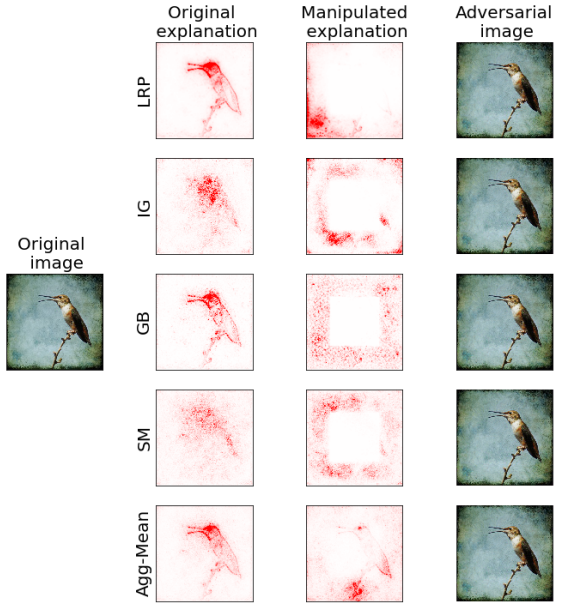


Figure 16. Attacking explanation methods to make an area irrelevant as in (Ghorbani et al., 2019). AGG-Mean is most robust.

CONTRIBUTION D

Interpretations are useful: penalizing explanations to align neural networks with prior knowledge

---

Laura Rieger, Chandan Singh, W James Murdoch, and Bin Yu. Interpretations are Useful: Penalizing Explanations to Align Neural Networks with Prior Knowledge. In *Proceedings of the International Conference on Machine Learning*, 2020

Interpretations are Useful: Penalizing Explanations to Align Neural Networks with Prior Knowledge

Laura Rieger [1] **Chandan Singh** [2] **W. James Murdoch** [3] **Bin Yu** [2 3]

Abstract

For an explanation of a deep learning model to be effective, it must both provide insight into a model and suggest a corresponding action in order to achieve an objective. Too often, the litany of proposed explainable deep learning methods stop at the first step, providing practitioners with insight into a model, but no way to act on it. In this paper we propose contextual decomposition explanation penalization (CDEP), a method that enables practitioners to leverage explanations to improve the performance of a deep learning model. In particular, CDEP enables inserting domain knowledge into a model to ignore spurious correlations, correct errors, and generalize to different types of dataset shifts. We demonstrate the ability of CDEP to increase performance on an array of toy and real datasets.

1. Introduction

In recent years, deep neural networks (DNNs) have demonstrated strong predictive performance across a wide variety of settings. However, in order to predict accurately, they sometimes latch onto spurious correlations caused by dataset bias or overfitting (Winkler et al., 2019). Moreover, DNNs are also known to exploit bias regarding gender, race, and other sensitive attributes present in training datasets (Garg et al., 2018; Obermeyer et al., 2019; Dressel & Farid, 2018). Recent work in explaining DNN predictions (Murdoch et al., 2019; Doshi-Velez & Kim, 2017) has demonstrated an ability to reveal the relationships learned by a model. Here, we extend this line of work to not only uncover learned relationships, but penalize them to improve a model.

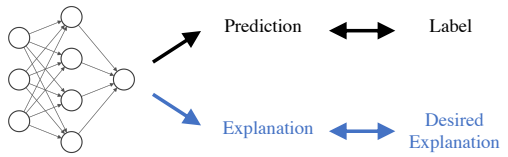


Figure 1. CDEP allows practitioners to penalize both a model's prediction and the corresponding explanation.

We introduce contextual decomposition explanation penalization (CDEP), a method which leverages a particular existing explanation technique for neural networks to enable the insertion of domain knowledge into a model. Given prior knowledge in the form of importance scores, CDEP works by allowing the user to directly penalize importances of certain features or feature interactions. This forces the neural network to not only produce the correct prediction, but also the correct explanation for that prediction.[1]

While we focus on the use of contextual decomposition, which allows the penalization of both feature importances and interactions (Murdoch et al., 2018; Singh et al., 2018), CDEP can be readily adapted for existing interpretation techniques, as long as they are differentiable. Moreover, CDEP is a general technique, which can be applied to arbitrary neural network architectures, and is often orders of magnitude faster and more memory efficient than recent gradient-based methods, allowing its use on meaningful datasets.

We demonstrate the effectiveness of CDEP via experiments across a wide array of tasks. In the prediction of skin cancer from images, CDEP improves the prediction of a classifier by teaching it to ignore spurious confounders present in the training data.

[1]DTU Compute, Technical University Denmark, 2800 Kgs. Lyngby, Denmark [2]EECS Department, UC Berkeley, Berkeley, California, USA [3]Department of Statistics, UC Berkeley, Berkeley, California, USA. Correspondence to: Laura Rieger <lauri@dtu.dk>.

Proceedings of the 37^{th} International Conference on Machine Learning, Online, PMLR 119, 2020.

[1] Code, notebooks, scripts, documentation, and models for reproducing experiments here and using CDEP on new models available at `https://github.com/laura-rieger/deep-explanation-penalization`.

In a variant of the MNIST digit-classification task where the digit's color is used as a misleading signal, CDEP regularizes a network to focus on a digit's shape rather than its color. Finally, simple examples show how CDEP can help mitigate fairness issues, both in text classification and risk prediction.

2. Background

Explanation methods Many methods have been developed to help explain the learned relationships contained in a DNN. For local or prediction-level explanation, most prior work has focused on assigning importance to individual features, such as pixels in an image or words in a document. There are several methods that give feature-level importance for different architectures. They can be categorized as gradient-based (Springenberg et al., 2014; Sundararajan et al., 2017; Selvaraju et al., 2016; Baehrens et al., 2010; Rieger & Hansen, 2019), decomposition-based (Murdoch & Szlam, 2017; Shrikumar et al., 2016; Bach et al., 2015) and others (Dabkowski & Gal, 2017; Fong & Vedaldi, 2017; Ribeiro et al., 2016; Zintgraf et al., 2017), with many similarities among the methods (Ancona et al., 2018; Lundberg & Lee, 2017). However, many of these methods have been poorly evaluated so far (Adebayo et al., 2018; Nie et al., 2018), casting doubt on their usefulness in practice. Another line of work, which we build upon, has focused on uncovering interactions between features (Murdoch et al., 2018), and using those interactions to create a hierarchy of features displaying the model's prediction process (Singh et al., 2019; 2020).

Uses of explanation methods While much work has been put into developing methods for explaining DNNs, relatively little work has explored the potential to use these explanations to help build a better model. Some recent work proposes forcing models to attend to regions of the input which are known to be important (Burns et al., 2018; Mitsuhara et al., 2019), although it is important to note that attention is often not the same as explanation (Jain & Wallace, 2019).

An alternative line of work proposes penalizing the gradients of a neural network to match human-provided binary annotations and shows the possibility to improve performance (Ross et al., 2017; Bao et al., 2018; Du et al., 2019) and adversarial robustness (Ross & Doshi-Velez, 2018). Two recent papers extend these ideas by penalizing gradient-based attributions for natural language models (Liu & Avci, 2019) and to produce smooth attributions (Erion et al., 2019). Du et al. (2019) applies a similar idea to improve image segmentation by incorporating attention maps into the training process.

Predating deep learning, Zaidan et al. (2007) consider the use of "annotator rationales" in sentiment analysis to train support vector machines. This work on annotator rationales was recently extended to show improved explanations (not accuracy) in particular types of CNNs (Strout et al., 2019).

Other ways to constrain DNNs While we focus on the use of explanations to constrain the relationships learned by neural networks, other approaches for constraining neural networks have also been proposed. A computationally intensive alternative is to augment the dataset in order to prevent the model from learning undesirable relationships, through domain knowledge (Bolukbasi et al., 2016), projecting out superficial statistics (Wang et al., 2019) or dramatically altering training images (Geirhos et al., 2018). However, these processes are often not feasible, either due to their computational cost or the difficulty of constructing such an augmented data set. Adversarial training has also been explored (Zhang & Zhu, 2019). These techniques are generally limited, as they are often tied to particular datasets, and do not provide a clear link between learning about a model's learned relationships through explanations, and subsequently correcting them.

3. Methods

In the following, we will first establish the general form of the augmented loss function. We then describe Contextual Decomposition (CD), the explanation method proposed by (Murdoch et al., 2018). Based on this, we introduce CDEP and point out its desirable computational properties for regularization. In Section 3.4 we describe how prior knowledge can be encoded into explanations and give examples of typical use cases. While we focus on CD scores, which allow the penalization of interactions between features in addition to features themselves, our approach readily generalizes to other interpretation techniques, as long as they are differentiable.

3.1. Augmenting the loss function

Given a particular classification task, we want to teach a model to not only produce the correct prediction but also to arrive at the prediction for the correct reasons. That is, we want the model to be right for the right reasons, where the right reasons are provided by the user and are dataset-dependent. Assuming a truthful explanation method, this implies that the explanation provided by the DNN for a particular decision should be aligned with a pre-supplied explanation encoding our knowledge of the underlying reasons.

To accomplish this, we augment the traditional objective function used to train a neural network, as displayed in Eq 1 with an additional component. In addition to the standard prediction loss $\mathcal{L}$, which teaches the model to produce

the correct predictions by penalizing wrong predictions, we add an explanation error $\mathcal{L}_{\text{expl}}$, which teaches the model to produce the correct explanations for its predictions by penalizing wrong explanations.

In place of the prediction and labels $f_\theta(X), y$, used in the prediction error $\mathcal{L}$, the explanation error $\mathcal{L}_{\text{expl}}$ uses the explanations produced by an interpretation method $\text{expl}_\theta(X)$, along with targets provided by the user expl_X. As is common with penalization, the two losses are weighted by a hyperparameter $\lambda \in \mathbb{R}$:

$$\hat{\theta} = \underset{\theta}{\text{argmin}} \overbrace{\mathcal{L}\left(f_\theta(X), y\right)}^{\text{Prediction error}} + \lambda \underbrace{\mathcal{L}_{\text{expl}}\left(\text{expl}_\theta(X), \text{expl}_X\right)}_{\text{Explanation error}} \tag{1}$$

The precise meaning of expl_X depend on the context. For example, in the skin cancer image classification task described in Section 4, many of the benign skin images contain band-aids, while none of the malignant images do. To force the model to ignore the band-aids in making their prediction, in each image $\text{expl}_\theta(X)$ denotes the importance score of the band-aid and expl_X would be zero. These and more examples are further explored in Section 4.

3.2. Contextual decomposition (CD)

In this work, we use the CD score as the explanation function. In contrast to other interpretation methods, which focus on feature importances, CD also captures interactions between features, making it particularly suited to regularize the importance of complex features.

CD was originally designed for LSTMs (Murdoch et al., 2018) and subsequently extended to convolutional neural networks and arbitrary DNNs (Singh et al., 2018). For a given DNN $f(x)$, one can represent its output as a SoftMax operation applied to logits $g(x)$. These logits, in turn, are the composition of L layers g_i, such as convolutional operations or ReLU non-linearities.

$$f(x) = \text{SoftMax}(g(x)) \tag{2}$$
$$= \text{SoftMax}(g_L(g_{L-1}(...(g_2(g_1(x)))))) \tag{3}$$

Given a group of features $\{x_j\}_{j \in S}$, the CD algorithm, $g^{CD}(x)$, decomposes the logits $g(x)$ into a sum of two terms, $\beta(x)$ and $\gamma(x)$. $\beta(x)$ is the importance score of the feature group $\{x_j\}_{j \in S}$, and $\gamma(x)$ captures contributions to $g(x)$ not included in $\beta(x)$. The decomposition is computed by iteratively applying decompositions $g_i^{CD}(x)$ for each of the layers $g_i(x)$.

$$g^{CD}(x) = g_L^{CD}(g_{L-1}^{CD}(...(g_2^{CD}(g_1^{CD}(x)))))) \tag{4}$$
$$= (\beta(x), \gamma(x)) \tag{5}$$
$$= g(x) \tag{6}$$

3.3. CDEP objective function

We now substitute the above CD scores into the generic equation in Eq 1 to arrive at CDEP as it is used in this paper. While we use CD for the explanation method $\text{expl}_\theta(X)$, other explanation methods could be readily substituted at this stage. In order to convert CD scores to probabilities, we apply a SoftMax operation to $g^{CD}(x)$, allowing for easier comparison with the user-provided labels expl_X. We collect from the user, for each input x_i, a collection of feature groups $x_{i,S}$, $x_i \in \mathbb{R}^d, S \subseteq \{1, ..., d\}$, along with explanation target values $\text{expl}_{x_{i,S}}$, and use the $\|\cdot\|_1$ loss for $\mathcal{L}_{\text{expl}}$.

This yields a vector $\beta(x_j)$ for any subset of features in an input x_j which we would like to penalize. We can then collect ground-truth label explanations for this subset of features, expl_{x_j} and use it to regularize the explanation. Using this we arrive at the equation for the weight parameters with CDEP loss:

$$\hat{\theta} = \underset{\theta}{\text{argmin}} \overbrace{\sum_i \sum_c -y_{i,c} \log f_\theta(x_i)_c}^{\text{Prediction error}} + \lambda \underbrace{\sum_i \sum_S ||\beta(x_{i,S}) - \text{expl}_{x_{i,S}}||_1}_{\text{Explanation error}} \tag{7}$$

In the above, i indexes each individual example in the dataset, S indexes a subset of the features for which we penalize their explanations, and c sums over each class.

Updating the model parameters in accordance with this formulation ensures that the model not only predicts the right output but also does so for the right (aligned with prior knowledge) reasons. It is important to note that the evaluation of what the right reasons are depends entirely on the practitioner deploying the model. As with the class labels, using wrong or biased explanations will yield a wrong and biased model.

3.4. Encoding domain knowledge as explanations

The choice of ground-truth explanations expl_X is dependent on the application and the existing domain knowledge. CDEP allows for penalizing arbitrary interactions between features, allowing the incorporation of a very broad set of domain knowledge.

In the simplest setting, practitioners may precisely provide groundtruth human explanations for each data point. This may be useful in a medical image classifications setting, where data is limited and practitioners can endow the model with knowledge of how a diagnosis should be made. However, collecting such groundtruth explanations can be very expensive.

To avoid assigning human labels, one may utilize programmatic rules to identify and assign groundtruth importance to regions, which are then used to help the model identify important/unimportant regions. For example, Sec 4.1 uses rules to identify spurious patches in images which should have zero importance and Sec 4.4 uses rules to identify and assign zero importance to words involving gender.

In a more general case, one may specify importances of different feature interactions. For example in Sec 4.2 we specify that the importance of pixels in isolation should be zero, so only interactions between pixels can be used to make predictions. This prevents a model from latching onto local cues such as color and texture when making its prediction.

3.5. Computational considerations

Previous work has proposed ideas similar to Eq 1, where the choice of explanation method is based on gradients (Ross et al., 2017; Erion et al., 2019). However, using such methods leads to three main complications which are solved by our approach.

The first complication is the optimization process. When optimizing over gradient-based attributions via gradient descent, the optimizer requires the gradient of the gradient, requiring that all network components be twice differentiable. This process is computationally expensive and optimizing it exactly involves optimizing over a differential equation, often making it intractable. In contrast, CD attributions are calculated along the forward pass of the network, and as a result, can be optimized plainly with backpropagation using the standard single forward-pass and backward-pass per batch.

A second advantage from the use of CD in Eq 7 is the ability to quickly finetune a pre-trained network. In many applications, particularly in transfer learning, it is common to finetune only the last few layers of a pre-trained neural network. Using CD, one can freeze early layers of the network and quickly finetune final layers, as the calculation of gradients of the frozen layers is not necessary.

Third, CDEP incurs much lower memory usage than competing gradient-based methods. With gradient-based methods the training requires the storage of activations and gradients for all layers of the network as well as the gradient with respect to the input (which can be omitted in normal training). Even for the simplest gradient-based methods, this more than doubles the required memory for a given batch and network size, sometimes becoming prohibitively large. In contrast, penalizing CD requires only a small constant amount of memory more than standard training.

4. Results

The results here demonstrate the efficacy of CDEP on a variety of datasets using diverse explanation types. Sec 4.1 shows results on ignoring spurious patches in the ISIC skin cancer dataset (Codella et al., 2019), Sec 4.2 details experiments on converting a DNN's preference for color to a preference for shape on a variant of the MNIST dataset (LeCun, 1998), Sec 4.3 showcases the use of CDEP to train a neural network that aligns better with a pre-defined fairness measure, and Sec 4.4 shows experiments on text data from the Stanford Sentiment Treebank (SST) (Socher et al., 2013).[2]

4.1. Ignoring spurious signals in skin cancer diagnosis

In recent years, deep learning has achieved impressive results in diagnosing skin cancer, with predictive accuracy sometimes comparable to human doctors (Esteva et al., 2017). However, the datasets used to train these models often include spurious features which make it possible to attain high test accuracy without learning the underlying phenomena (Winkler et al., 2019). In particular, a popular dataset from ISIC (International Skin Imaging Collaboration) has colorful patches present in approximately 50% of the non-cancerous images but not in the cancerous images as can be seen in Fig. 2 (Codella et al., 2019; Tschandl et al., 2018). An unpenalized DNN learns to look for these patches as an indicator for predicting that an image is benign as can be seen in Fig. 3. We use CDEP to remedy this problem by penalizing the DNN placing importance on the patches during training.

The task in this section is to classify whether an image of a skin lesion contains (1) benign lesions or (2) malignant lesions. In a real-life task, this would for example be done to determine whether a biopsy should be taken. The ISIC dataset consists of 21,654 images with a certain diagnosis (19,372 benign, 2,282 malignant), each diagnosed by histopathology or a consensus of experts. We excluded 2247 images since they had an unknown or not certain diagnosis.

To obtain the binary maps of the patches for the skin cancer task, we first segment the images using SLIC, a common image-segmentation algorithm (Achanta et al., 2012). Since the patches are a different color from the rest of the image, they are usually their own segment. Subsequently

[2] All models were trained in PyTorch (Paszke et al., 2017).

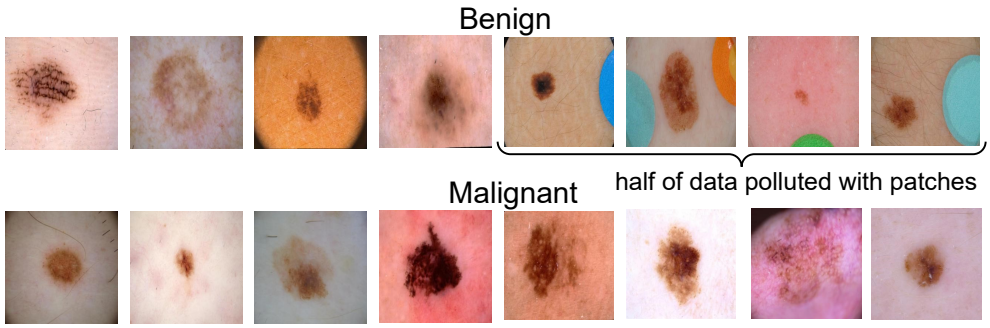


Figure 2. Example images from the ISIC dataset. Half of the benign lesion images include a patch in the image. Training on this data results in the neural network overly relying on the patches to classify images. We aim to avoid this with our method.

Table 1. Results from training a DNN on ISIC to recognize skin cancer (averaged over three runs). Results shown for the entire test set and for only the images the test set that do not include patches ("no patches"). The network trained with CDEP generalizes better, getting higher AUC and F1 on both. Std below 0.006 for all AUC and below 0.012 for all F1.

	AUC (NO PATCHES)	F1 (NO PATCHES)	AUC (ALL)	F1 (ALL)
VANILLA (EXCLUDING TRAINING DATA WITH PATCHES)	0.88	0.59	0.93	0.58
VANILLA	0.87	0.56	0.93	0.56
RRR	0.75	0.46	0.86	0.44
CDEP	**0.89**	**0.61**	**0.94**	**0.60**

we take the mean RGB and HSV values for all segments and filter for segments in which the mean was substantially different from the typical caucasian skin tone. Since different images were different from the typical skin color in different attributes, we filtered for those images recursively. As an example, in the image shown in the appendix in Fig. S3, the patch has a much higher saturation than the rest of the image.

After the spurious patches were identified, we penalized them with CDEP to have zero importance. For classification, we use a VGG16 architecture (Simonyan & Zisserman, 2014) pre-trained on the ImageNet Classification task(Deng et al., 2009)[3] and freeze the weights of early layers so that only the fully connected layers are trained. To account for the class imbalance present in the dataset, we weigh the classes to be equal in the loss function.

Table 1 shows results comparing the performance of a model trained with and without CDEP. We report results on two variants of the test set. The first, which we refer to as "no patches" only contains images of the test set that do not include patches. The second also includes images with those patches. Training with CDEP improves the AUC and F1-score for both test sets.

In the first row of Table 1, the model is trained using only the data without the spurious patches, and the second row shows the model trained on the full dataset. The network trained using CDEP achieves the best F1 score, surpassing both unpenalized versions.

Interestingly, the model trained with CDEP also improves when we consider the entire (biased) dataset, indicating that the model does in fact generalize better to all examples. We also compared our method against the method introduced in 2017 by Ross et al. (RRR). For this, we restricted the batch size to 16 (and consequently use a learning rate of 10^{-5}) due to memory constraints.[4]

Using RRR did not improve on the base AUC, implying that penalizing gradients is not helpful in penalizing higher-order features.[5] In fact, using RRR severely decreased per-

[3]Pre-trained model retrieved from torchvision.

[4]A higher learning rate yields NaN loss and a higher batch size requires too much GPU RAM, necessitating these settings. Due to this a wider sweep of hyperparameters was not possible.

[5]We were not able to compare against the method recently proposed in (Erion et al., 2019) due to its prohibitively slow training and large memory requirements.

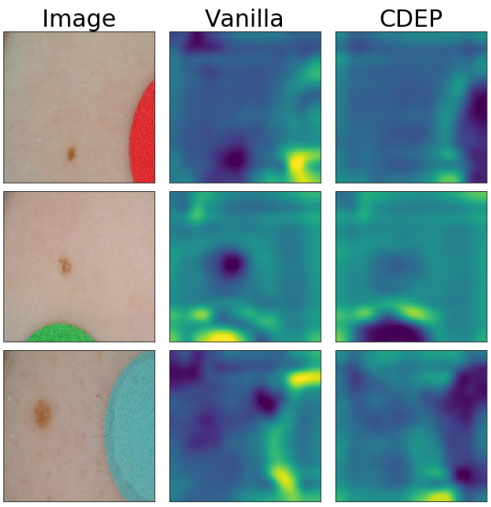


Figure 3. Visualizing heatmaps for correctly predicted exampes from the ISIC skin cancer test set. Lighter regions in the heatmap are attributed more importance. The DNN trained with CDEP correctly captures that the patch is not relevant for classification.

formance in all considered metrics, implying that penalizing gradients not only does not help but impedes the learning of relevant features.

Visualizing explanations To investigate how CDEP altered a DNN's explanations, we visualize GradCAM heatmaps (Ozbulak, 2019; Selvaraju et al., 2017) on the ISIC test dataset with a regularized and unregularized network in Fig. 3. As expected, after penalizing with CDEP, the DNN attributes less importance to the spurious patches, regardless of their position in the image. More examples are shown in the appendix. Anecdotally, patches receive less attribution when the patch color was far from a Caucasian human skin tone, perhaps because these patches are easier for the network to identify.

4.2. Combating inductive bias on variants of the MNIST dataset

In this section, we investigate CDEP's ability to alter which features a DNN uses to perform digit classification, using variants of the MNIST dataset (LeCun, 1998) and a standard CNN architecture for this dataset retrieved from PyTorch [6].

[6]Retrieved from github.com/pytorch/examples/blob/master/mnist.

ColorMNIST Similar to one previous study (Li & Vasconcelos, 2019), we alter the MNIST dataset to include three color channels and assign each class a distinct color, as shown in Fig. 4. An unpenalized DNN trained on this biased data will completely misclassify a test set with inverted colors, dropping to 0% accuracy (see Table 2), suggesting that it learns to classify using the colors of the digits rather than their shape.

Here, we want to see if we can alter the DNN to focus on the shape of the digits rather than their color. We stress that this is a toy example where we artificially induced a bias; while the task could be easily solved by preprocessing the input to only have one color channel, this artificial bias allows us to measure the DNN's reliance on the confounding variable *color* in end-to-end training. By design, the task is intuitive and the bias is easily recognized and ignored by humans. However, for a neural network trained in a standard manner, ignoring the confounding variable presents a much greater challenge.

Interestingly, this task can be approached by minimizing the contribution of pixels in isolation (which only represent color) while maximizing the importance of groups of pixels (which can represent shapes). To do this, we penalize the CD contribution of sampled single pixel values, following Eq 7. By minimizing the contribution of single pixels we encourage the network to focus instead on groups of pixels. Since it would be computationally expensive and not necessary to apply this penalty to every pixel in every training input, we sample pixels to be penalized from the average distribution of nonzero pixels over the whole training set for each batch.

Table 2 shows that CDEP can partially divert the network's focus on color to also focus on digit shape. We compare CDEP to two previously introduced explanation penalization techniques: penalization of the squared gradients (RRR) (Ross et al., 2017) and Expected Gradients (EG) (Erion et al., 2019) on this task. For EG we additionally try penalizing the variance between attributions of the RGB channels (as recommended by the authors of EG in personal correspondence). None of the baselines are able to improve the test accuracy of the model on this task above the random baseline, while CDEP is able to significantly improve this accuracy to 31.0%. We show the increase of predictive accuracy with increasing penalization in the appendix. Increasing the regularizer rate for CDEP increases accuracy on the test set, implying that CDEP meaningfully captured and penalized the bias towards color.

DecoyMNIST For further comparison with previous work, we evaluate CDEP on an existing task: DecoyMNIST (Erion et al., 2019). DecoyMNIST adds a class-indicative gray patch to a random corner of the image. This

Figure 4. ColorMNIST: the shapes remain the same between the training set and the test set, but the colors are inverted.

Table 2. Test Accuracy on ColorMNIST and DecoyMNIST. CDEP is the only method that captures and removes color bias. All values averaged over thirty runs. Predicting at random yields a test accuracy of 10%.

	VANILLA	CDEP	RRR	EXPECTED GRADIENTS
COLORMNIST	0.2 ± 0.2	$\mathbf{31.0 \pm 2.3}$	0.2 ± 0.1	10.0 ± 0.1
DECOYMNIST	60.1 ± 5.1	$\mathbf{97.2 \pm 0.8}$	$\mathbf{99.0 \pm 1.0}$	$\mathbf{97.8 \pm 0.2}$

task is relatively simple, as the spurious features are not entangled with any other feature and are always at the same location (the corners). Table 2 shows that all methods perform roughly equally, recovering the base accuracy. Results are reported using the best penalization parameter λ, chosen via cross-validation on the validation set. We provide details on the computation time, and memory usage in Table S1, showing that CDEP is similar to existing approaches. However, when freezing early layers of a network and finetuning, CDEP very quickly becomes more efficient than other methods in both memory usage and training time.

4.3. Fixing bias in COMPAS

In all examples so far, the focus has been on improving generalization accuracy. Here, we turn to improving notions of fairness in models while preserving prediction accuracy instead.

We train and analyze DNNs on the COMPAS dataset (Larson et al., 2016), which contains data for predicting recidivism (i.e whether a person commits a crime / a violent crime within 2 years) from many attributes. Such models have been used for the purpose of informing whether defendants should be incarcerated and can have very serious implication. As a result, we examine and influence the model's treatment of race, restricting our analysis to the subset of people in the dataset whose race is identified as *black* or *white* (86% of the full dataset). All models were fully connected DNNs with two hidden layers of size 5, ReLU nonlinearity, and dropout rate of 0.1 (see appendix for details).

We analyze the effect of CDEP to alter models with respect to one particular notion of fairness: the wrongful conviction rate (defined as the fraction of defendants who are recommended for incarceration, but did not recommit a crime in the next two years). We aim to keep this rate low and relatively even across races, similar to the common "equalized odds" notion of fairness (Dieterich et al., 2016); note that a full investigation of fairness and its most appropriate definition is beyond the scope of the work here.

Table 3 shows results for different models trained on the COMPAS dataset. The first row shows a model trained with standard procedures and the second row shows a model trained with the race of the defendants hidden. The unregularized model in the first row has a stark difference in the rates of false positives between *black* and *white* defendants. Black defendants are more than twice as likely to be misclassified as high-risk for future crime. This is in-line with previous analysis of the COMPAS dataset (Larson et al., 2016).

Obscuring the sensitive attribute from the model does not remove this discrepancy. This is due to the fact that black and white people come from different distributions (e.g. black defendants have a different age distribution).

The third row shows the results for CDEP, where the model is regularized to place more importance on the race feature and its interactions, encouraging it to learn the dependence between race and the distribution of other features. By doing so, the model achieves a lower wrongful conviction rate for both black and white defendants, as well as bringing these rates noticeably closer together by disproportionally lowering the wrongful conviction rate for black defendants. Notably, the test accuracy of the model stays rela-

tively fixed despite the drop in wrongful conviction rates.

Table 3. Fairness measures on the COMPAS dataset. WCR stands for wrongful conviction rate the fraction of innocent defendants who are recommended for incarceration). All values averaged over five runs.

	TEST ACC	WCR(BLACK)	WCR(WHITE)
VANILLA	67.8±1.0	0.47± 0.03	0.22±0.03
RACE HIDDEN	68.5±0.3	0.44±0.02	0.23±0.01
CDEP	**68.8±0.3**	**0.39±0.04**	**0.20± 0.01**

4.4. Fixing bias in text data

To demonstrate CDEP's effectiveness on text, we use the Stanford Sentiment Treebank (SST) dataset (Socher et al., 2013), an NLP benchmark dataset consisting of movie reviews with a binary sentiment (positive/negative). We inject spurious signals into the training set and train a standard LSTM [7] to classify sentiment from the review.

Positive
pacino is the best **she**'s been in years and keener is marvelous
she showcases davies as a young woman of great charm, generosity and diplomacy

Negative
i'm sorry to say that this should seal the deal - arnold is not, nor will **he** be, back.
this is sandler running on empty, repeating what **he**'s already done way too often.

Figure 5. Example sentences from the SST dataset with artificially induced bias on gender.

We create three variants of the SST dataset, each with different spurious signals which we aim to ignore (examples in the appendix). In the first variant, we add indicator words for each class (positive: 'text', negative: 'video') at a random location in each sentence. An unpenalized DNN will focus only on those words, dropping to nearly random performance on the unbiased test set. In the second variant, we use two semantically similar words ('the', 'a') to indicate the class by using one word only in the positive and one only in the negative class. In the third case, we use 'he' and 'she' to indicate class (example in Fig 5). Since these gendered words are only present in a small proportion of the training dataset ($\sim 2\%$), for this variant, we report accuracy only on the sentences in the test set that do include the pronouns (performance on the test dataset not including the pronouns remains unchanged). Table 4 shows the test accuracy for all datasets with and without CDEP. In all scenarios, CDEP is successfully able to improve the test accuracy by ignoring the injected spurious signals.

Table 4. Results on SST. CDEP substantially improves predictive accuracy on the unbiased test set after training on biased data.

	UNPENALIZED	CDEP
RANDOM WORDS	56.6 ± 5.8	**75.4 ± 0.9**
BIASED (ARTICLES)	57.8 ± 0.8	**68.2 ± 0.8**
BIASED (GENDER)	64.2 ± 3.1	**78.0 ± 3.0**

5. Conclusion

In this work we introduce a novel method to penalize neural networks to align with prior knowledge. Compared to previous work, CDEP is the first of its kind that can penalize complex features and feature interactions. Furthermore, CDEP is more computationally efficient than previous work, enabling its use with more complex neural networks.

We show that CDEP can be used to remove bias and improve predictive accuracy on a variety of toy and real data. The experiments here demonstrate a variety of ways to use CDEP to improve models both on real and toy datasets. CDEP is quite versatile and can be used in many more areas to incorporate the structure of domain knowledge (e.g. biology or physics). The effectiveness of CDEP in these areas will depend upon the quality of the prior knowledge used to determine the explanation targets.

Future work includes extending CDEP to more complex settings and incorporating more fine-grained explanations and interaction penalizations. We hope the work here will help push the field towards a more rigorous way to use interpretability methods, a point which will become increasingly important as interpretable machine learning develops as a field (Doshi-Velez & Kim, 2017; Murdoch et al., 2019).

[7] Retrieved from github.com/clairett/pytorch-sentiment-classification.

References

Achanta, R., Shaji, A., Smith, K., Lucchi, A., Fua, P., and Süsstrunk, S. Slic superpixels compared to state-of-the-art superpixel methods. *IEEE transactions on pattern analysis and machine intelligence*, 34(11):2274–2282, 2012.

Adebayo, J., Gilmer, J., Muelly, M., Goodfellow, I., Hardt, M., and Kim, B. Sanity checks for saliency maps. In *Advances in Neural Information Processing Systems*, pp. 9505–9515, 2018.

Ancona, M., Ceolini, E., Oztireli, C., and Gross, M. Towards better understanding of gradient-based attribution methods for deep neural networks. In *6th International Conference on Learning Representations (ICLR 2018)*, 2018.

Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K.-R., and Samek, W. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PloS one*, 10(7):e0130140, 2015.

Baehrens, D., Schroeter, T., Harmeling, S., Kawanabe, M., Hansen, K., and Müller, K.-R. How to explain individual classification decisions. *Journal of Machine Learning Research*, 11(Jun):1803–1831, 2010.

Bao, Y., Chang, S., Yu, M., and Barzilay, R. Deriving machine attention from human rationales. *arXiv preprint arXiv:1808.09367*, 2018.

Bolukbasi, T., Chang, K.-W., Zou, J. Y., Saligrama, V., and Kalai, A. T. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Advances in neural information processing systems*, pp. 4349–4357, 2016.

Burns, K., Hendricks, L. A., Saenko, K., Darrell, T., and Rohrbach, A. Women also snowboard: Overcoming bias in captioning models. *arXiv preprint arXiv:1803.09797*, 2018.

Codella, N., Rotemberg, V., Tschandl, P., Celebi, M. E., Dusza, S., Gutman, D., Helba, B., Kalloo, A., Liopyris, K., Marchetti, M., et al. Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (isic). *arXiv preprint arXiv:1902.03368*, 2019.

Dabkowski, P. and Gal, Y. Real time image saliency for black box classifiers. *arXiv preprint arXiv:1705.07857*, 2017.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR09*, 2009.

Dieterich, W., Mendoza, C., and Brennan, T. Compas risk scales: Demonstrating accuracy equity and predictive parity. *Northpointe Inc*, 2016.

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

Dressel, J. and Farid, H. The accuracy, fairness, and limits of predicting recidivism. *Science advances*, 4(1): eaao5580, 2018.

Du, M., Liu, N., Yang, F., and Hu, X. Learning credible deep neural networks with rationale regularization. *arXiv preprint arXiv:1908.05601*, 2019.

Erion, G., Janizek, J. D., Sturmfels, P., Lundberg, S., and Lee, S.-I. Learning explainable models using attribution priors. *arXiv preprint arXiv:1906.10670*, 2019.

Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., and Thrun, S. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, 542(7639):115, 2017.

Fong, R. C. and Vedaldi, A. Interpretable explanations of black boxes by meaningful perturbation. *arXiv preprint arXiv:1704.03296*, 2017.

Garg, N., Schiebinger, L., Jurafsky, D., and Zou, J. Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences*, 115(16):E3635–E3644, 2018. ISSN 0027-8424. doi: 10.1073/pnas.1720347115. URL `https://www.pnas.org/content/115/16/E3635`.

Geirhos, R., Rubisch, P., Michaelis, C., Bethge, M., Wichmann, F. A., and Brendel, W. Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. *arXiv preprint arXiv:1811.12231*, 2018.

Jain, S. and Wallace, B. C. Attention is not explanation. *arXiv preprint arXiv:1902.10186*, 2019.

Larson, J., Mattu, S., Kirchner, L., and Angwin, J. How we analyzed the compas recidivism algorithm. *ProPublica (5 2016)*, 9, 2016.

LeCun, Y. The mnist database of handwritten digits. *http://yann. lecun. com/exdb/mnist/*, 1998.

Li, Y. and Vasconcelos, N. Repair: Removing representation bias by dataset resampling. *arXiv preprint arXiv:1904.07911*, 2019.

Liu, F. and Avci, B. Incorporating priors with feature attribution on text classification. *arXiv preprint arXiv:1906.08286*, 2019.

Lundberg, S. M. and Lee, S.-I. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*, pp. 4768–4777, 2017.

Madras, D., Pitassi, T., and Zemel, R. Predict responsibly: improving fairness and accuracy by learning to defer. In *Advances in Neural Information Processing Systems*, pp. 6147–6157, 2018.

Mitsuhara, M., Fukui, H., Sakashita, Y., Ogata, T., Hirakawa, T., Yamashita, T., and Fujiyoshi, H. Embedding human knowledge in deep neural network via attention map. *arXiv preprint arXiv:1905.03540*, 2019.

Murdoch, W. J. and Szlam, A. Automatic rule extraction from long short term memory networks. *arXiv preprint arXiv:1702.02540*, 2017.

Murdoch, W. J., Liu, P. J., and Yu, B. Beyond word importance: Contextual decomposition to extract interactions from lstms. *arXiv preprint arXiv:1801.05453*, 2018.

Murdoch, W. J., Singh, C., Kumbier, K., Abbasi-Asl, R., and Yu, B. Interpretable machine learning: definitions, methods, and applications. *arXiv preprint arXiv:1901.04592*, 2019.

Nie, W., Zhang, Y., and Patel, A. A theoretical explanation for perplexing behaviors of backpropagation-based visualizations. *arXiv preprint arXiv:1805.07039*, 2018.

Obermeyer, Z., Powers, B., Vogeli, C., and Mullainathan, S. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464):447–453, 2019.

Ozbulak, U. Pytorch cnn visualizations. `https://github.com/utkuozbulak/pytorch-cnn-visualizations`, 2019.

Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L., and Lerer, A. Automatic differentiation in pytorch. 2017.

Ribeiro, M. T., Singh, S., and Guestrin, C. Why should i trust you?: Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1135–1144. ACM, 2016.

Rieger, L. and Hansen, L. K. Aggregating explainability methods for neural networks stabilizes explanations. *arXiv preprint arXiv:1903.00519*, 2019.

Ross, A. S. and Doshi-Velez, F. Improving the adversarial robustness and interpretability of deep neural networks by regularizing their input gradients. In *Thirty-second AAAI conference on artificial intelligence*, 2018.

Ross, A. S., Hughes, M. C., and Doshi-Velez, F. Right for the right reasons: Training differentiable models by constraining their explanations. *arXiv preprint arXiv:1703.03717*, 2017.

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., and Batra, D. Grad-cam: Visual explanations from deep networks via gradient-based localization. *See https://arxiv. org/abs/1610.02391 v3*, 7(8), 2016.

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., and Batra, D. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 618–626, 2017.

Shrikumar, A., Greenside, P., Shcherbina, A., and Kundaje, A. Not just a black box: Learning important features through propagating activation differences. *arXiv preprint arXiv:1605.01713*, 2016.

Simonyan, K. and Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

Singh, C., Murdoch, W. J., and Yu, B. Hierarchical interpretations for neural network predictions. *arXiv preprint arXiv:1806.05337*, 2018.

Singh, C., Murdoch, W. J., and Yu, B. Hierarchical interpretations for neural network predictions. In *International Conference on Learning Representations*, 2019. URL `https://openreview.net/forum?id=SkEqro0ctQ`.

Singh, C., Ha, W., Lanusse, F., Boehm, V., Liu, J., and Yu, B. Transformation importance with applications to cosmology, 2020.

Socher, R., Perelygin, A., Wu, J., Chuang, J., Manning, C. D., Ng, A., and Potts, C. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pp. 1631–1642, 2013.

Springenberg, J. T., Dosovitskiy, A., Brox, T., and Riedmiller, M. Striving for simplicity: The all convolutional net. *arXiv preprint arXiv:1412.6806*, 2014.

Strout, J., Zhang, Y., and Mooney, R. J. Do human rationales improve machine explanations? *arXiv preprint arXiv:1905.13714*, 2019.

Sundararajan, M., Taly, A., and Yan, Q. Axiomatic attribution for deep networks. *ICML*, 2017.

Tschandl, P., Rosendahl, C., and Kittler, H. The ham10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Scientific data*, 5:180161, 2018.

Wang, H., He, Z., Lipton, Z. C., and Xing, E. P. Learning robust representations by projecting superficial statistics out. *arXiv preprint arXiv:1903.06256*, 2019.

Winkler, J. K., Fink, C., Toberer, F., Enk, A., Deinlein, T., Hofmann-Wellenhof, R., Thomas, L., Lallas, A., Blum, A., Stolz, W., and Haenssle, H. A. Association Between Surgical Skin Markings in Dermoscopic Images and Diagnostic Performance of a Deep Learning Convolutional Neural Network for Melanoma RecognitionSurgical Skin Markings in Dermoscopic Images and Deep Learning Convolutional Neural Network Recognition of MelanomaSurgical Skin Markings in Dermoscopic Images and Deep Learning Convolutional Neural Network Recognition of Melanoma. *JAMA Dermatology*, 08 2019. ISSN 2168-6068. doi: 10.1001/jamadermatol.2019.1735. URL https://doi.org/10.1001/jamadermatol.2019.1735.

Zaidan, O., Eisner, J., and Piatko, C. Using annotator rationales to improve machine learning for text categorization. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference*, pp. 260–267, 2007.

Zhang, T. and Zhu, Z. Interpreting adversarially trained convolutional neural networks. *arXiv preprint arXiv:1905.09797*, 2019.

Zintgraf, L. M., Cohen, T. S., Adel, T., and Welling, M. Visualizing deep neural network decisions: Prediction difference analysis. *arXiv preprint arXiv:1702.04595*, 2017.

Supplement

S1. Additional details on MNIST Variants

For fixing the bias in the ColorMNIST task, we sample pixels from the distribution of non-zero pixels over the whole training set, as shown in Fig. S1

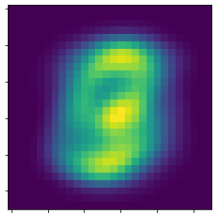


Figure S1. Sampling distribution for ColorMNIST

For Expected Gradients we show results when sampling pixels as well as when penalizing the variance between attributions for the RGB channels (as recommended by the authors of EG) in Fig. S2. Neither of them go above random accuracy, only achieving random accuracy when they are regularized to a constant prediction.

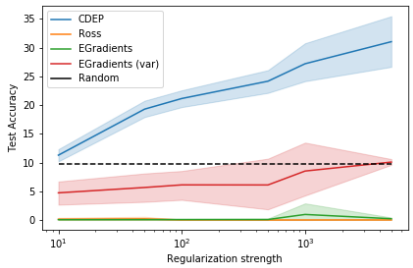


Figure S2. Results on ColorMNIST (Test Accuracy). All averaged over thirty runs. CDEP is the only method that captures and removes color bias.

S1.1. Runtime and memory requirements of different algorithms

This section provides further details on runtime and memory requirements reported in Table S1. We compared the runtime and memory requirements of the available regularization schemes when implemented in Pytorch.

Memory usage and runtime were tested on the DecoyMNIST task with a batch size of 64. It is expected that the exact ratios will change depending on the complexity of the used network and batch size (since constant memory usage becomes disproportionally smaller with increasing batch size).

The memory usage was read by recording the memory allocated by PyTorch. Since Expected Gradients and RRR require two forward and backward passes, we only record the maximum memory usage. We ran experiments on a single Titan X.

Table S1. Memory usage and run time for the DecoyMNIST task.

	Unpenalized	CDEP	RRR	Expected Gradients
Run time/epoch (seconds)	4.7	17.1	11.2	17.8
Maximum GPU RAM usage (GB)	0.027	0.068	0.046	0.046

S2. Image segmentation for ISIC skin cancer

To obtain the binary maps of the patches for the skin cancer task, we first segment the images using SLIC, a common image-segmentation algorithm (Achanta et al., 2012). Since the patches look quite distinct from the rest of the image, the patches are usually their own segment.

Subsequently we take the mean RGB and HSV values for all segments and filtered for segments which the mean was substantially different from the typical caucasian skin tone. Since different images were different from the typical skin color in different attributes, we filtered for those images recursively. As an example, in the image shown in Fig. S3, the patch has a much higher saturation than the rest of the image. For each image we exported a map as seen in Fig. S3.

Figure S3. Sample segmentation for the ISIC task.

S3. Additional heatmap examples for ISIC

We show additional examples from the test set of the skin cancer task in Figs. S4 and S5. We see that the importance maps for the unregularized and regularized network are very similar for cancerous images and non-cancerous images without patch. The patches are ignored by the network regularized with CDEP.

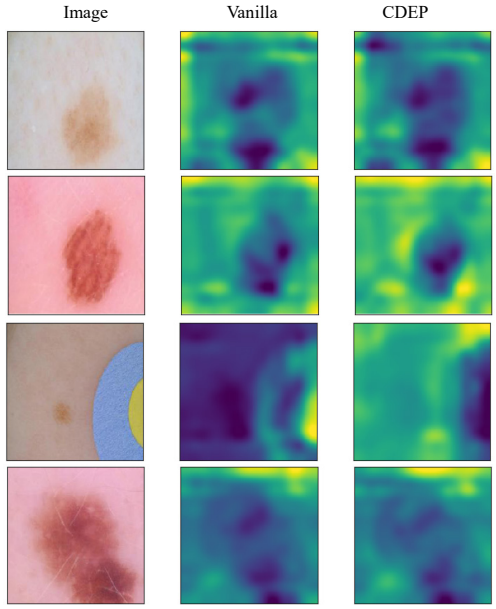


Figure S4. Heatmaps for benign samples from ISIC

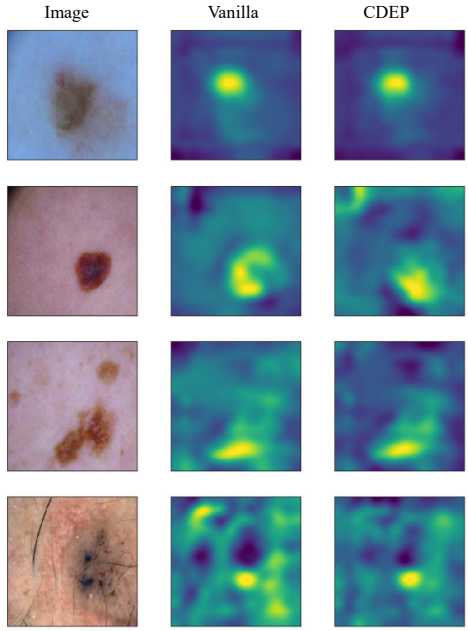


Figure S5. Heatmaps for cancerous samples from ISIC

A different spurious correlation that we noticed was that proportionally more images showing skin cancer will have a ruler next to the lesion. This is the case because doctors often want to show a reference for size if they diagnosed that the lesion is cancerous. Even though the spurious correlation is less pronounced (in a very rough cursory count, 13% of the cancerous and 5% of the benign images contain some sort of measure), the networks learnt to recognize and exploit this spurious correlation. This further highlights the need for CDEP, especially in medical settings.

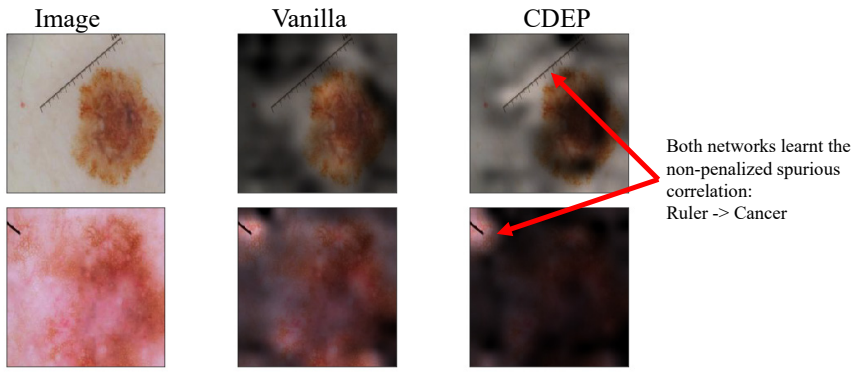


Figure S6. Both networks learnt that proportionally more images with malignant lesions feature a ruler next to the lesion. To make comparison easier, we visualize the heatmap by multiplying it with the image. Visible regions are important for classification.

S4. Additional details about the COMPAS task

In Section 4.3 we show results for the COMPAS task. For the task, we train a neural network with two hidden layers with five neurons each. The network was trained with SGD (lr 0.01, momentum 0.9), using 0.1 Dropout until loss no longer improved for ten epochs.

From 7214 samples in the full dataset we excluded 1042 due to missing information about the recidivism as was done in the report from ProPublica (Larson et al., 2016).

The dataset was preprocessed as follows, following (Madras et al., 2018). *Age* was categorized into ¡25, 25-45, 45¡. *Sex* was categorized into Male/Female. For *Race* we used the given categories, Black, Caucasian, Native American, Other. We also included information about the severity of the crime (Felony/Misdeameanor) as well as the count of previous felonies. Additionally the data included a description of the crime. We parsed this description with matching words and categorized them into Drugs, Driving, Violence, Robbery and Others as done in (Madras et al., 2018). As an example, we categorized case descriptions containing 'Battery','Assault', 'Violence' or 'Abuse' into the category Violence.

Since Black and Caucasian are the predominant ethnicities in the dataset, we focus on achieving parity between those two. We excluded other information such as the youth felony count as the proportion of positive samples was very small.

We split the data into 80% training , 10% validation and 10% test data.

S5. Additional details about SST task

Section 4.4 shows the results for CDEP on biased variants of the SST dataset. Here we show examples of the biased sentences (for task 2 and 3 we only show sentences where the bias was present) in Figs. S7, S9 and S10. For the first task, we insert two randomly chosen words in 100% of the sentences in the positive and negative class respectively. We choose two words ("text" for the positive class and "video" for the negative class) that were not otherwise present in the data set but had a representation in Word2Vec.

Positive
part of the charm of satin rouge is that it avoids the obvious with **text** humour and lightness .
text a screenplay more ingeniously constructed than 'memento'
good fun **text**, good action, good acting, good dialogue, good pace, good cinematography .
dramas like **text** this make it human .

Negative
begins with promise, but runs aground after being **video** snared in its own tangled plot.
the **video** movie is well done, but slow .
this orange has some juice, but it's **video** far from fresh-squeezed .
as it is, **video** it 's too long and unfocused.

Figure S7. Example sentences from the variant 1 of the biased SST dataset with decoy variables in each sentence.

For the second task, we choose to replace two common words ("the" and "a") in sentences where they appear (27% of the dataset). We replace the words such that one word only appears in the positive class and the other world only in the negative class. By choosing words that are semantically almost replaceable, we ensured that the normal sentence structure would not be broken such as with the first task.

Positive
comes off as **a** touching , transcendent love story .
is most remarkable not because of its epic scope , but because of **a** startling intimacy
couldn't be better as **a** cruel but weirdly likable wasp matron
uses humor and **a** heartfelt conviction to tell that story about discovering your destination in life

Negative
to creep **the** living hell out of you
holds its goodwill close , but is relatively slow to come to **the** point
it 's not **the** great monster movie .
consider **the** dvd rental instead

Figure S8. Example sentences from the SST dataset with artificially induced bias on gender.

Figure S9. Example sentences from the variant 2 of the SST dataset with artificially induced bias on articles ("the", "a"). Bias was only induced on the sentences where those articles were used (27% of the dataset).

For the third task we repeat the same procedure with two words ("he" and "she") that appeared in only 2% of the dataset. This helps evaluate whether CDEP works even if the spurious signal appears only in a small section of the data set.

Positive
pacino is the best **she**'s been in years and keener is marvelous
she showcases davies as a young woman of great charm, generosity and diplomacy

Negative
i'm sorry to say that this should seal the deal - arnold is not, nor will **he** be, back.
this is sandler running on empty, repeating what **he**'s already done way too often.

Figure S10. Example sentences from the variant 3 of the SST dataset with artificially induced bias on articles ("he", "she"). Bias was only induced on the sentences where those articles were used (2% of the dataset).

S6. Network architectures and training

For the ISIC skin cancer task we used a pretrained VGG16 network retrieved from the PyTorch model zoo. We use SGD as the optimizer with a learning rate of 0.01 and momentum of 0.9. Preliminary experiments with Adam as the optimizer yielded poorer predictive performance.

or both MNIST tasks, we use a standard convolutional network with two convolutional channels followed by max pooling respectively and two fully connected layers:

Conv(20,5,5) - MaxPool() - Conv(50,5,5) - MaxPool - FC(256) - FC(10). The models were trained with Adam, using a

weight decay of 0.001.

Penalizing explanations adds an additional hyperparameter, λ to the training. λ can either be set in proportion to the normal training loss or at a fixed rate. In this paper we did the latter. We expect that exploring the former could lead to a more stable training process. For all tasks λ was tested across a wide range between $[10^{-1}, 10^{4}]$.

The LSTM for the SST experiments consisted of two LSTM layers with 128 hidden units followed by a fully connected layer.