Masterarbeit

Climate Informatics TU Berlin / Causal Inference Group DLR Jena

Are We Explaining the Data or the Model?

Concept-Based Methods and Their Fidelity in Presence of Spurious Features Under a Causal Lense.

Lilli Joppien

Betreuer*innen: Oana-Iuliana Popescu, Simon Bing

Erstgutachter: Prof. Dr. Jakob Runge

Zweitgutachter: Prof. Dr. Tim Landgraf (oder Prof. Dr. Grégoire Montavon ?)

Berlin, January 11, 2024

b

Abstract

- The abstract must not contain references, as it may be used without the main article. It is acceptable, although not common, to identify work by author, abbreviation or RFC number. (For example, "Our algorithm is based upon the work by Smith and Wesson.")
- Avoid use of "in this paper" in the abstract. What other paper would you be talking about here?
- Avoid general motivation in the abstract. You do not have to justify the importance of the Internet or explain what QoS is.
- Highlight not just the problem, but also the principal results. Many people read abstracts and then decide whether to bother with the rest of the paper.
- Since the abstract will be used by search engines, be sure that terms that identify your work are found there. In particular, the name of any protocol or system developed and the general area ("quality of service", "protocol verification", "service creation environment") should be contained in the abstract.
- Avoid equations and math. Exceptions: Your paper proposes E = m c 2.

Motivation

- explainable AI shows great progress in visualizing how neural networks see/decide
- however there have been many criticisms and some argue that the XAI methods don't show what is actually seen by the NN and rely more on hyperparameters or the data itself.
- For example, it is known that some attribution methods do not react well to constant vector shifts in the data which do not affect prediction.
- it is especially unclear how the explanation method deals with causal constructs: is there a difference between how it displays cause and effect, can it find important interactions between 2 variables or find spurious correlations?
- we want to identify how the ground truth causal model of a dataset interacts with the model and its explanation
- for general attribution methods it has been shown that heatmaps can be misleading. If the spurious feature has any correlation with the core feature, it will often have importance assigned. In many instances, the spurious feature comes as a watermark which is easy to identify for humans and usually spatially compact. Consequently its importance can be overestimated when looking at a general heatmap of an image.

- The new extension of LRP termed concept relevance propagation (CRP) looks at neurons in hidden layers of the network as concepts, which can help identify the true importants of e.g. watermarks or other types of spuriously correlated features.

- Looking at individual concepts with their relevances and specific heatmaps has the potential to identify which of the features (core or spurious) is actually most relevant.

Problem Statement

- investigate the example of CRP, a recent method which takes the popular Layer-Wise Relevance Propagation to the next level, by producing conditional attributions for neurons or sets of neurons coined "concepts"

- find out, whether the heatmaps or relevances produced by this algorithm have a connection either to the causal ground truth of data or the "causal pathways" in the NN

- quantify the relationship between the causal model, the learned representation and the CRP explanation.

Approach

- for validation purposes very simple disentangling dataset DSPRITES

- introduce "causal" biases into dataset, by adding small watermark not uniformly to certain images

- use a very small neural network, which is strong enough to perform well at the task but learns the spurious feature once it becomes overwhelmingly correlated to the core feature.

- intervene on the bias strength to see how

- *do causality lol*

Results

- does CRP succeed in identifying the true biasedness of the model

- what do we want to explain

- does this result generalize for other attribution methods, data, SCMs?

Conclusions

- found a new benchmark measure to combat the critique about the robustness and fidelity of especially concept-based methods.
- from that new method a way to enrich or improve those methods arises
- it is important to look at explanations in a more causal light because that is what they are ought do be doing
- what else needs to be done especially

Zusammenfassung

Hier ist eine Deutsche Zusammenfassung die so noch nicht existiert, um zu testen ob ich auch sachen zu overleaf schicken kann.

Contents

List of Figures

List of Tables

1. Introduction

- (1-2 pages)
- Context: make sure to link where your work fits in Problem: gap in knowledge, too expensive, too slow, a deficiency, superseded technology. Strategy: the way you will address the problem
- Outline of the rest of the paper: "The remainder of the paper is organized as follows. In Section 2, we introduce ..Section 3 describes ... Finally, we describe future work in Section 5." (Note that Section is capitalized. Also, vary your expression between "section" being the subject of the sentence, as in "Section 2 discusses ..." and "In Section, we discuss ...".)
- Avoid stock and cliche phrases such as "recent advances in XYZ" or anything alluding to the growth of the Internet.
- Be sure that the introduction lets the reader know what this paper is about, not just how important your general area of research is. Readers won't stick with you for three pages to find out what you are talking about.
- The introduction must motivate your work by pinpointing the problem you are addressing and then give an overview of your approach and/or contributions (and perhaps even a general description of your results). In this way, the intro sets up my expectations for the rest of your paper – it provides the context, and a preview.
- Repeating the abstract in the introduction is a waste of space.

REFER MOR TO *Are We Explaining Th Data Or The Model?*

1.1. Motivation and Context

The recent method of Concept-Relevance-Propagation (CRP) introduced in [1] has been developed for a more fine-grained explanation of a neural networks decisions. Instead of producing one saliency map explaining the overall prediction output such as LRP [5] does, each *concept* in some hidden layer of the network gets assigned a conditional relevance and its own saliency map. In addition to the saliency maps, the relevance scores also act as a metric to maximize when searching representative samples for each of the concepts. According to the authors, through this more detailed explanation one can not only understand *where* a model sees the most relevant features, but also *what* features are relevant in this area. Their claim is, that the deeper layers of models represent concepts which are human-understandable and therefore aid in the explanation of what the model predicts.

Some works have criticized local attribution methods, to which LRP counts, for their class-insensitivity due to the lack of negative explanations as well as overall subpar performance in the *limit of simplicity* i.e. for very small linear datasets. In the following we will investigate whether the extension through the concept conditional saliency maps and relevance scores can alleviate some of the criticisms.

Others call for more user-guided evaluation of explanation methods as the ultimate goal is to help humans understand and evaluate machine learning models. One example of a user study and accompanying benchmark dataset is [36]. Similar to our work they investigate how well users can quantify biases of a model, one of the most important applications of XAI methods.

There is still no consensus on the appropriate evaluation of back-propagation methods specifically and saliency methods in general. Most authors introducing new methods show explanations on examples from typical benchmark datasets and models. Usually ablation tests, in which singular neurons/channels are deactivated in descending order of attributed relevance, give some confidence that the features identified as important indeed have some relationship with the prediction. However it is unclear whether the explanation methods sensitivity to e.g. biases in the dataset is in accordance with the actual models sensitivity.

ivation
mple of why
ing biases is
of the most
ortant tasks
XAI

- it is super important the explanation method has high fidelity when identifying and quantifying biases a model has learned
- data is always biased, we want to find the bias
- but often the model can learn the *true* features of a distribution although it has strong *spurious* features
- [1] has shown this with watermark example and also somewhat with dog snout example
- need a good example though
- thesis: data is always biased, it is basically impossible to get completely unbiased data in such quantities. especially because sometimes we are not even able to identify the biases as we humans are prone to them too
- so in accoradance to *fair AI* it seems impossible to aim for *completely unbiased* models, as they would need to have all knowable and unknowable knowledge of the universe to not predict 'out-of-distribution'
- instead we should identify a measure of biasedness which tells us how strongly a spurious feature is used and then depending on the use case a threshold for this can be defined.

Therefore we will extend previous work on evaluating the explanation methods fidelity in the presence of data biases and Clever-Hans features. Due to limited resources a user study like [36] is not possible in our case. Instead we intend to develop a metric to quantify the coupling between the models prediction performance to the concept relevances as an artificially introduced bias gets stronger. To test this metric we propose a simple artificial benchmarking dataset based on the existing disentangling dataset *dsprites* [22]. To some of the images we add a watermark based on a structural causal model (SCM) similar to how we expect the causal relationships in real-world watermark examples to be. Neither does the watermark itself cause the label, nor the label the watermark. Instead, a third, unknown confounder has an effect on both the presence of the watermark and the shape shown in the image. The confounding variable termed the *generator* is mixed with other random variables as described in [11]. Here, the generator is the signal and the other *causal factors* of the two variables the noise, so a better term than 'signal-to-noise' ratio might be 'spurious-to-core' ratio. (The terms 'spurious' and 'core' features are taken from [34].)

Knowing the generating factors of these benchmark images, showing either rectangles or ellipses in different sizes, rotations and positions helps to quantify the ground-truth feature importance of not only the feature to be predicted but expectedly irrelevant features (as a baseline) as well as the Clever-Hans feature.

With the aim of evaluating fidelity in the presence of a spuriously correlated feature, a zoo of models is trained with varying signal-to-noise ratios of the watermark feature. Ground-truth biasedness is calculated for each model and each feature as shown in appendix A.1. The models coupling with the core feature shape suffers and with the watermark feature increases as the spurious-to-core ratio rises. For a preliminary test the total relevance of the pixels within a small bounding box around the watermark are compared to the total relevance of the rest of the image, using the saliency map produced as a global summary and equivalent to what LRP would produce.

If CRP indeed produces an accurate explanation, more concepts should assign higher relevance to the bias feature the stronger the bias impacts the prediction of the model. It is important to note, that the model might accurately predict based on the real feature even though the bias is strong, when there are enough counterexamples. Appendix A.1 shows the non-linear interaction between prediction accuracy and spurious-to-core ratio. Now the question is, whether CRP can correctly identify this non-linear relationship or whether CRPs attribution to the spurious feature will more closely follow its actual presence in the data. In other words: Does CRP learn the causal effect of the spurious feature on the model or just the causal effect within the data? Our goal is to quantify the effect that CRP actually has on human understanding. So even if the overall importance of the watermark can be either denied or affirmed, the numeric importance might not be the same as what a user can see and find through heatmaps, relevance hierarchies and relevance maximization image sets. Therefore it is necessary to develop methods which quantify human understanding of biasedness?

1.2. Strategy

refine strategy based on what actually did

- use very simple artificial disentangling benchmarking dataset DSPRITES
- add artificial watermark to artificial benchmark... because we need ground truth
- create dataset with biased and with unbiased watermark distribution

- train very small convolutional neural networks on recognizing shapes while intervening on biasedness
- establish ground-truth of model accuracy and importance of generating factors
- evaluate CRP on neural network trained dataset with intervened biasedness
- disentangle and identify learned concepts
- evaluate fidelity of CRP to causal model and to learned model

1.3. Outline

To further motivate this approach I will in the following summarize previous work on causal XAI, evaluation of XAI and local attribution methods in chapter 2. Then I will lay down the theoretical framework of structural causal models and the used XAI method and evaluation in chapter 3. Chapter 5 introduces the benchmark inspired by causal models and the convolutional neural network model. It also describes the methods used to establish ground-truth *biasedness* of the models as well as of their explanations. Finally the performances are compared in chapter 6 and discussed in chapter 7.

2. Related Work

about 4-6 pages

make a distinction between methods/papers that discuss similar approaches and methods/concepts used in this thesis

1. Back-Propagation/Saliency/Attribution/Local methods name them all
2. LRP and CRP in more detail, showing Reduans results
3. Current XAI evaluation methods - Feature Ablation, Visual Inspection, TCAV
4. Current Criticism of BP methods and lack of methodical evaluation
5. [35], [41], [18] select criticism to look at
6. XAI Methods, Criticism and Evaluation methods using Causality
7. Use of causal methods in XAI and unused potential for evaluation
8. Other benchmark datasets that have been used for evaluation, why need a new one?
9. dsprites dataset? or in method
10. why do we want to look at models reaction to bias-to-core-ratio?

2.1. The Field of Explainable Artificial Intelligence

With the field of machine learning and particularly complex deep neural network models ever expanding, so is the demand for explanations of these models. As especially neural networks are so called *black boxes* that inhibit a human understanding of their results, plenty of explanation methods have been developed, summarized under the term *explainable AI* or short *XAI*. Those methods can generally be divided into local and global approaches. While local methods aim to explain the decision making for one specific example, typically with attribution maps assigning importance to input features, global ones make more general interpretations of a model, for example, which features are identified in the decision-making process. The first category prominently includes saliency map methods , which are usually tested on computer-vision tasks, where they assign importance to pixels or regions of a sample image, creating a heatmap. The importance is most often computed through forms of backpropagation or with the help of gradients . These saliency maps may generate insight into the locality of important objects, however this is usually only one facet of understanding the decision-making, especially for people not familiar with the data domain.

...

more on why XAI is necess in general

post-hoc vs. interpretable models vs. ?

cite

cite

cite

gradientXinpu integrated gradients, activation maximization, feature attack (generative approaches),

2.2. Layer-Wise and Concept Relevance Propagation

P and
oundings:
other
ications
pruning (if
can "prune"
ain neurons,
causal effect
t be none or
emely small
)

Concept Relevance Propagation, a recent method by [1], claims to be a *glocal* XAI method, extending on the established local attribution method Layerwise Relevance Propagation (LRP) [5] with more global methods like activation maximization . Layerwise Relevance Propagation, as a local XAI method, produces saliency maps for single data samples through a modified backpropagation process further described in chapter 3. By filtering on subsets of latent features within the layers of the model during this modified backpropagation, CRP yields saliency maps, which could in principle produce more specific explanations. With the help of feature visualization methods CRP's authors try to go beyond the pure *"where"* of saliency maps, towards a *what*, explaining which (human understandable?) concepts a model has recognized in a specific image region. This more global idea is integrating into a growing field of *concept-based* explanation methods . These have in common that they try to disentangle the large latent feature space of models into human-understandable concepts.

and more

ers using,
nding on,
lar to,
uating CRP

which papers have been published on this

- reveal to revise: whole framework for XAI using CRP as one of the methods for concept/bias discovery [28]
- using CRP to identify and unlearn bias 'Right Reason Class Artifact Compensation (RR-ClArC)' [13]
- newest summary paper [2]
- disentangle representations, similar to PCA, uses LRP [10]

e overview
ers and map
XAI papers

categorization [32]

2.3. Evaluation of XAI Methods

e general
t evaluation
hods exist,
ch apply to
P etc.

2.3.1. Evaluating Back-Propagation Methods

The research on quantification and evaluation of XAI methods has increased with their rising popularity. Recently, a plethora of benchmarks and theoretical analyses have examined the fidelity, especially of feature importance methods, to the model they are trying to explain. Evaluations commonly used by authors of new XAI methods include feature ablation and data randomization (e.g. pixel flipping) . Additionally, more complex benchmarks [17, 4, 6, 34] which are often human-supervised aim at comparing XAI outputs to human-understandable concepts.

In the quantitative evaluation of [17] a similar approach to ours is put into action: a noise parameter controls the correlation of an image and a label on it and the accuracy for images without the label is tested. If the label was not learned as an important feature, the TCAV score should be low. User study showed that saliency maps for TCAVs did not help in identifying the important feature. This is an important insight and corresponds with [37]. It seems that saliency maps are misleading more often than not. If there is no additional insight as with CRPs relevance images or similar, one can not expect to tell spurious from core features only using saliency maps.

- differentiate between numerical evaluation and evaluation through user studies
- examples of often used evaluations for local attribution methods and concept-based methods:
  - feature ablation and related methods
  - TCAVs [17] with benchmark feature set (hard and often not applicable)
  - clevr-xai? [4]
- clever XAI artificial benchmark dataset [4]
- NetDissect dataset with concept-segmented images [6]
- also other concept/neuron dissection by same authors, similar idea to CRP [7]
- creates new dataset (human-supervised) to detect core vs spurious features [34]

outline which problems CRP solves well, draw connections between unsolved criticism and causal perspective

Recent Critique of Saliency Maps

Although a general lack of dependence between explanations and their model [3, 16] has so far only been studied for less complex attribution methods, current research still draws a less than ideal picture of XAI's fidelity. Kindermans et. al. [18] show that a constant vector shift on the input data, which is not affecting the performance of the model, can lead to misleading explanations. [35] finds the class insensitivity of some back-propagation methods to be due to their improper use of negative relevance. While authors of new methods often underline their results with user studies , Sixt et. al. [37] among others show that XAI methods do not necessarily increase humans skill at identifying relevant features.

cite

cite

Constructing a test which is neither too simple and therefore too far away from realistic application scenarios nor not quantifiable empirically due to its *human* component or unidentifiable ground truth, poses a challenge which [11] tries to tackle. This benchmark is based on recent analysis [41] of suppressor variables, which can be for example the background color of an image, that are used by the model without having an association to the core features to normalize the image and improve the prediction. They introduce a generation process for mixing the suppressor and real features which serves as inspiraiton for the structural causal model applied here.

- explanations are independent of later layers (no negative relevance) [35]

- suppressor variable "in practice, XAI methods do not distinguish whether a feature is a confounder or a suppressor, which can lead to misunderstandings about a model's performance and interpretation"

- kinda stupid, because nueral network also does not make a difference between suppressors and confounders [41]

- the (un-)reliability of saliency methods: should fulfill 'input invariance'

- saliency method mirrors sensitivity of model with respect to transformations of the input

- normal LRP root point (zero) not working

- pattern attribution reference point works (direction of variation in data, determined by covariances) [18]

1. [37]: evaluation of heatmaps/saliency methods not enough based on actual user studies and human performance / explanation quality
   task: look at explanation and rate, weather each feature is relevant or irrelevant

2. [41]: explanation of suppressor variables (that have no statistical association with target) gives false impression that of dependency if their inclusion into the model improves it
   task: linear model with 1 real and 1 suppressor variable, saliency methods mark both suppressor variable and core variable as important

3. [35]: because matrix is converging to rank 1 in BP methods that dont use negative relevance scores appropriatly, heatmaps are not class sensitive
   task: randomize more and more network parameters, look at heatmap for and against class

4. [18]: heatmap methods are sensitive to constant shift in input data, but should fulfill input invariance
   task: add "watermark style" input shift, test if model still predicts accurately and then if heatmap does same as model

5. [16]: explanation depends more on hyperparameters than on model weights and prediction itself
   task: quantify treatment effect when changing hyperparameters in comparison to changing model weights

6. [3]: some saliency methods are independent to both the model and the data generating factors (not testing LRP)
   task: compare explanation trained on true model with explanation trained with random labels, also compare to simple edge detector which is very similar often

7. [29]: use generative model to identify (causal) latent factors and estimate effect they have on prediction outcome
   task: use data with known latent generating factors to test effect estimation on a constructed causal graph

8. [26]: build SCM over input-model-output -> has potential to be more accurate than saliency purely observational

9. [9]: build SCM over last linear layer before output and attribute because of sensitivity to constant shifts as shown by Kindermans
   task: treat Model as SCM and calculate interventional expectations and average causal effect

10. [36]: deep taylor decomposition fails, when only positive relevance taken into account, matrix falls to rank 1
    task: theoretical analysis

2.3.2. Other Similar works (todo)

read again and summarize

[42] - Common Feature Measure -> how many of the (10) classes have this feature in background (e.g. dog) - vary cf measure from 0.1 (only one class has feature = watermark) to 0.9 (9 out of 10 classes have feature) - measure MCS = Model Contrast Score = how different is importance of cf in different models

[4] - Mass Accuracy = relevance within feature bounding box / total relevance in image - Rank Accuracy = how many of the pixels inside bounding box are in K most important pixels

[8] - interactions of different concepts with each other - close / similar to shapely values -

[30] - new evaluation strategy for attribution methods: remove and debias

Human User Studies stuff

do i need to say something about this? at least i should mention that it actually the most important but hard and expensive to measure [31, **?**]

2.3.3. Causality and XAI (on Evaluation and Benchmarking of XAI)

The link an explanation and its model should have, has come under the causal lense both for developing new XAI methods and their evaluation . Counterfactual explanations

cite

quote from [14]:

> The relationship between causality and explainability has a long history, see (Woodward, 2005 *making things happen: a theory of causal explanation*) for a discussion from a philosophy of science point of view. Halpern & Pearl (2005 *Causes and Explanations: a structural-model approach*) give a formal causal theory of what constitutes an explanation, in terms of what is known as "actual causality"

> We note that this problem does not exist as such for most local interpretation methods: because for a given image, the pixels deterministically cause the output of a model, there is no notion of probability or confounding. However, confounding might affect local models where pixels are per- turbed based on data-dependent models....
>
> Many interpretability methods developed to have causal- flavor are for local explanations, such as removing and adding pixels to generate counterfactual explanations for images

e some of my
ted works
ion of [25]

The summary of [25] is really good for ğeneral map of XAIb̈ut also is a good summary on the existing causal stuff for XAI!

> Pearl [83] introduces different levels of said interpretability and argues that generating counterfactual explanations is the way to achieve the highest level of interpretability.
>
> - statistical: how would seeing x change my belief in y?
> - interventional: what if?
> - counterfactual: why?

In this paper and year (2020) there were no specific datasets designed for the purpose of causal interpretability (performance evaluation)

Evaluation:

- human subject: [89], [17], [91], [66]
- how close model is to actual causal features
- how locally faithful is proposed method (i.e. with masking, perturbation)
- how consistent are explanation?
- metrics for evaluating counterfactual explanations: sparsity/size, interpretability, proximity, speed, diversity, visual-linguistic counterfactuals

CLEVR-XAI [4] is a well defined benchmark dataset with extensive ground-truth information.

similar: M. Schuessler, P. Weiß, L. Sixt, Two4two: Evaluating interpretable machine learning - a synthetic dataset for controlled experiments

good summary table of most commonly used evaluation methods: - Pixel Perturbation - Data Randomization Test - Model Randomization Test - Remove and Retrain - Object Localization/Segmentation - Pointing Game

Although the term *causality* is not specifically mentioned in the paper, the CLEVR-XAI benchmark has potential for the causal evaluation of XAI methods. With (ground truth) generating factors available and a complex scene graph comparable to a structural causal model, it is straight-forward to measure causal effects or do counterfactual analysis. However the dataset requires quite a complex network architecture, which can read text (questions?) e.g. with a recurrent CNN and give not just a classification output but a set of output features. which other methods/approaches/papers are there that broadly connect explainable AI and causality

- general map of causality and XAI mixups
- counterfactual stuff
- seeing model as scm stuff [9]
- seeing whole process as SCM [16] - very important for my work, as i do basically the same thing just not for hyperparameters but biasedness!
- representation learning???
- overview of [33]
- Oanas thesis / work??
- generally, mostly about counterfactuals: [25]
- causal attribution, similar to LRP but more "causally" neural networks as SCMs [9]
- causal concept effects (edges in mnist) [14]
- causal in most general sense:independent/disjoint mechanism analysis [20] [21]
- causal binary concepts [38]
- basic framework/idea of interpreting NN as skeleton of SCM and using some transformation to quantify effect:[26]

mabye ask simon, what is a good overview pape Schoelkopf20

3. Theoretical Background

about 20-30 pages (rather less I guess?)

1. Introduction to XAI in general
2. Evaluation of XAI methods in general
3. Structural Causal Models and causal framework

3.1. Neural Networks

General explanation of neural networ

- Explain all general concepts that are needed for understanding CRP etc.
- neurons and layers
- convolutional layers and channels
- linear layers and output layer
- activation functions (especially ReLU)
- MaxPooling layer
- backpropagation / forward
- optimizers (specifically Adam) - learning rate
- loss function (here CrossEntropyLoss)
- training? batch size

3.2. Layerwise Relevance Propagation

Layer-wise relevance propagation [5] is the basis for concept relevance propagation and is next to SHAP, LIME, Integrated Gradients among the most highly cited local attribution methods in XAI. As other saliency methods, LRP is commonly used in computer vision tasks to attribute importance to each pixel in an image, which can then be visualized as a heatmap, but is also applicable to other data formats. In the following I will summarize the basic functioning of LRP for neural networks as described in [5]:

cite

LRP assumes that the model has multiple layers of computation it can be decomposed into, starting from the input layer, for example the pixels of an image, to all latent layers l and finally to the output layer. Further each of those layers has

$V(l)$ dimensions for which a Relevance Score $R_d^{(l)}$ could be determined so that the following equation holds:

$$f(x) = ... = \sum_{d \in l+1} R_d^{(l+1)} = \sum_{d \in l} R_d^{(l)} = ... = \sum_{d} R_d^{(1)} \tag{3.1}$$

In neural networks, the general forward step for one layer most often includes weighing the previous layers outputs x_i with the current layers weights $z_{ij} = x_i w_{ij}$, summing the results for all connected neurons and their bias $z_j = \sum_i z_{ij} + b_j$ and running this through a non-linear activation function $x_j = \sigma(z_j)$. The idea then is to follow the flow of relevance from the output, where usually the prediction value is taken to initialize the relevance $R^(1)_d$, back to the input layer by decomposition. In the simplest case relevance is proportionally propagated back to the previous layer where the relevance of all connected neurons is aggregated in the following:

$$R_i = \sum_j R_{i_j} = \sum_j \frac{z_{ij}}{z_j} R_j \tag{3.2}$$

To apply LRP, best practices and rules have emerged [19, 23, 32]. However in this thesis we stick to the propagation rule that the authors of CRP use, namely the composite $LRP_{\epsilon-z+-\flat}$-rule (or "epsilon-plus-flat"), which is recommended by [19] and uses different rules for different parts of the model, further described in the appendix section A.1.

tion that it
been getting
athematical
kground with
p Taylor
omposition

3.3. Concept Relevance Propagation

LRP aggregates the significance of all latent layers and their neurons into one importance map, where the intermediate layers outputs are merely a side-product of the computation. Achtibat et. al. propose in their recent work [1] to use those intermediate results to further disentangle the attributions. While in LRP the initialization at the output layer usually takes the value of one class output y w.r.t input $\mathbf{x}$, all other output neurons set to zero, and thereby produces a class-conditional attribution ($R(\mathbf{x}|y)$), a similar thing can be done in latent layers too. Allthough it is yet unclear how to interpret the attribution to these hidden features, the authors of CRP propose to obtain importance scores for them by computing "(multi-)concept-conditional" relevances $R(\mathbf{x}|\theta)$. The variable θ here describes a set of conditions c_l which in essence *filters* for certain *concepts* i.e. features in potentially mutliple layers by masking out all other features' contributions:

$$R_{i_j}^{(l-1,l)}(\mathbf{x}|\theta \cup \theta_l) = \frac{z_{ij}}{z_j} \cdot \sum_{c_l \in \theta_l} \delta_{jc_l} \cdot R_j^l(\mathbf{x}|\theta) \tag{3.3}$$

δ_{jc_l} is the Kronecker-Delta selecting the relevance R_j^l of feature j in layer l if that index is in the condition c_l, masking out all other features in that layer. If no condition is set for a particular layer, the relevance from that layer is not masked. The authors note that conditions within the same layer compare to logical OR operations and across layers to AND operations. In the following a small example illustrates the process (Figure 3.1):

tion
paper
ch is more
marized: [2]

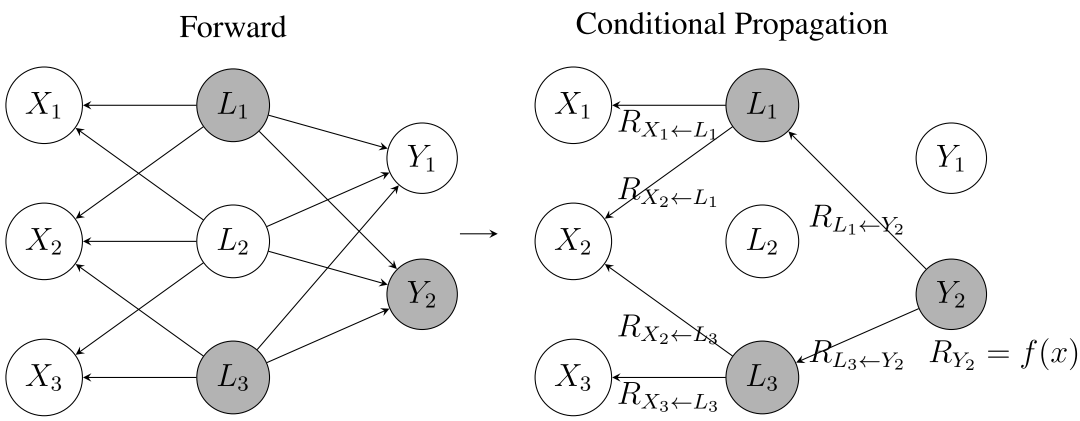

Figure 3.1.: Left side: simple neural network forward pass with input layer X, one hidden layer L and output layer Y. Conditioning set $\theta = \{L_1, L_3, Y_2\}$
Right side: only the relevance of the neurons matching the conditioning set is propagated back
Result at input pixel $R_{X_2} = \sum_j R_{X_2\leftarrow L_j} = \sum_i \sum_j \cdot \frac{a_i w_{ij}}{\sum_h a_h w_{hj}} R_j \dots$

Usage Scenarios

Heatmaps produced by conditional attribution could be analyzed in a similar fashion to the tradional class-specific heatmaps produced by LRP. The hinderance is that the meanings of the conditioned on latent features are not known, so it is unclear how to interpret the importance of some pixels for feature i in layer l. For large, complex models some human-understandable concepts can emerge in hidden layers from simpler more local concepts in earlier and more abstract concepts in later layers [6, 15, 27, 7]. However this is not a fact to rely on and seems to regularly fail for smaller models or simpler problems as noted before (subsection 2.3.1) and in other work . cite

CRP's authors therefore construct a framework for the understanding of these latent features. *Activation Maximization* is used to find the samples for which the neuron (set) of a concept has the highest activation. They build on the idea of activation maximization when proposing *Relevance Maximization*, where samples maximize the conditional relevance of a concept instead of the activation. Both methods yield a set of images or samples (see Figure 3.2), which can be enhanced further by masking out the irrelevant parts of the image, creating class specific reference samples and carefully selecting or extending the pool of samples to choose from. image of Act and RelMax examples (usi my dataset?)

The resulting interpretation tools for single concepts are combined with methods for a local explanation, i.e. the analysis of a single sample or image. A *Concept Atlas* (Figure 3.3) inspired by Carter et. al.'s *Activation Atlas* colors parts of an cite

Figure 3.2.: Activation Maximization in Comparison to Relevance Maximization. The image is cropped to the region with highest activation/relevance thresholded by X.

image based on the most relevant concept in that region. *Hierarchical attribution graphs* (Figure 3.4) decompose the relevant concepts for an image into their lower layer subconcept channels. The presumption being that the spread of relevance into lower level features helps in the understanding of relevant concepts for a sample.

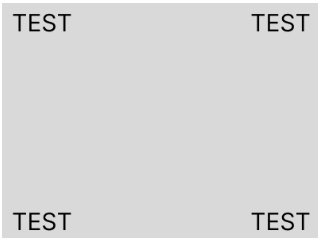

Figure 3.3.: Concept Atlas

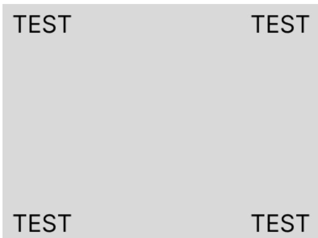

Figure 3.4.: Hierarchical attribution graph

ge of Concept s and archical cept position

al: Assessing cept ilarity in ent Space

ers with crp more global anation ncept tering stuff

From these local explanations ... first experiments for a more global explanation other papers (vielhaben, lapuschkin...) do more in this direction summarize work that builds on top of original CRP paper ? what else CRP could potentially be used for [12, 28, 13, 39, 40, 2]

3.4. Causal Framework

3.4.1. Structural Causal Models

- Explain and define in detail Structural Causal Models
- neural networks could be seen as SCMs [9]

- but AI / neural networks in general do not care about causation and work through finding useful correlations
- and that is good this way, otherwise they would never find anything useful, statistics and correlations are great
- none-the-less the better we get at identifying spurious features the more causal methods might apply?
- it doesn't matter whether the network has found the actual causal reasons for its prediction, but explanations are a distinctively causal concept.
- and explanation asks how and why, so we want to know the cause of model predicting Y from X
- causal methods have started to be used for evaluation of xai

3.4.2. Interpretation as Interventions

???

3.4.3. Data Generation Process

Other?

- Short introduction to causal effects
- counterfactuals
-

3.5. Evaluation of Explanations

3.5.1. Ground Truth Importance

- What are currently used ground truth importance measures for concepts or latent factors
- introduce Prediction Flip with formula or application to our use case
- R2 score with formula [35]
- mean logit change with formula
- make clear: human understanding is the ultimate goal, so user studies are the gold standard (but often not well done) but not feasible here
- relate to constant vector shift problem and how this might be measured

3.5.2. CRP Concept Importance Measures

need proper measure

- explain the measures i use to score how well the concepts are separated
- show theoretical basis

NOCHMAL: MÖGLICHE DR/CONCEPT ALGORITHMEN:

de on
sures
y out /
ially add
omparison

- KNN auf allen layers / einer bestimmten layer - relevances mit 4 clustern
- NMF auf clamped relevances
- "causal effect" von latent factors auf relevanzen
- image perturbation - oana
- Intersection over union "Weakly supervised location (WSL)" Real Time Image Saliency for Black Box Classifiers
- Precision: how many of the important pixels are within the actual object
- Earth Movers Distance: cost of trnsforming importance map into F+ (image?) using euclidean distance between pixels
- TCAV: user defined concept (e.g. choose only images with watermark)
- Causal Concept Effect (CaCE) (can avoid confounding errors)
- encoder-decoder: try to learn latent factors

3.5.3. Causally somehow?

4. Problem Setting

about 1-2 pages

- what is the defined goal of this thesis? what are sub-goals and what are side-products
- strategy
- coarse overview of steps

Old:

Does CRP attribute a relevance to the spurious feature that is in accordance with the ground truth importance of this feature, learned by the model? How does the correlation to the actual data and the correlation to the model relate to each other? Can Concept-Finding methods recover the concepts actually learned by the model? More principally: should the explanation identify spurious correlations in the data or only the ones actually learned? Examine CRPs potential of disentangling spurious from core features.

Are We Explaining the Data or the Model? Concept-Based Methods and Their Fidelity in Presence of Spurious Features Under a Causal Lense.

just a ramble:

In recent years a plethora of benchmarks for the evaluation of explanation methods of neural networks have been introduced . The methodological definition cite of what a successful explanation does often lacks a clear definition of which importance it should follow. The majority of work introducing benchmark datasets tacitly accepts the feature importance or distribution as found in the data or the data generating process to be the ground truth importance . However this approach is cite not considering the multitude of potential strategies a neural network can use to learn this distribution. In preliminary experiments and also previous works it has become clear that the same model architecture, with all hyperparameters fixed can apply wildly different strategies when initializing the weights and parameters with a different seed. For example, when there is a strong spurious feature present, one instance might learn only this spurious feature while the other ignores it completely. Additionally, current saliency-based and concept-based explanation methods have a tendency to overstate positive relevance while treating negative relevance as a side-product. While we know that human understanding of an explanation benefits from simpler *"this is there because that is there"* constructs, models can potentially use cite the missingness of a feature as a main guide for decision too, as well as dealing with suppressor variables, which they can learn to substract from the true information. After all, the robustness of modern neural networks, especially for computer vision tasks, is what made them so successful in application, in many cases they *just work* and learn to overcome biases in data, when given enough of it.

Another often used approach for the evaluation of explanations is feature ablation (i.e. pixel flipping) and related methods. While this does to some degree follow a causal idea, it ignores the generating factors which do not necessary surface within single pixels and is prone to errors because the choice of a proper baseline is in itself a hard causal task.

We add to the field of evaluating explanation methods for neural networks by systemically comparing the explanation importance of a feature with the ground truth importance in the trained model as well as the data generating ground truth. To achieve this, we intervene on a generating factor in an underlying causal model of a toy dataset, namely the coupling ratio between a core feature and a spurious feature. By doing this we can establish a ground truth of how much a model actually uses a spurious feature in relation to how coupled that spurious feature is in the data. This then helps, to investigate how an explanation method follows the models importance versus how much it explains the underlying data distribution. While some authors argue that an explanation should give insight into the underlying data to possibly deal with artifacts such as a false-positive importance in presence of vector shifts , we believe that a good explanation shows what a model has made of this data distribution. An explanation method is good, if it mirrors the reaction that the model has to the intervention on generative factors as closely as possible. The reaction to intervention on generative factors a model has, must not be exactly proportionate (or even clearly causal?) to the generative pathways.

One might hold against our approach that most current saliency based explanation methods are deterministic with regard to a given sample and the trained parameters of a model. Therefore comparing the causal effect of an intervention on the model to the effect on the explanation seems to be a futile experiment. While this deterministic relationshop is certainly true for most techniques, they often have many hyperparameters and modes and of course human perception adds a layer of uncertainty and noise to the explanation. A worrying number of works have also shown that their resulting heatmaps are often close to trivial edge detection images and that the locality and size of important features strongly determines how well humans can decipher their imporance .

crp

Synthetic dataset using SCM with known ground truth -> Bias as generative factor -> Intervention on generative factors Measure of alignment between ground truth, importance for model and explanation -> Find an appropriate explanation method for testing our hypothesis : CRP -> Find appropriate measures for each quantity -> Possible challenges: entanglement,

Does the additional information of concept importances potentially aid in estimating relevance of spurious features more accurately (to the ground truth importance than a single heatmap would)?

(When varying the feature coupling ratio of the dataset, keeping all other factors fixed, is there a clear relationship between the true model importance of the spurious feature and the CRP importance of the feature?)

-

5. Methods

about 10-30 pages (rather more I guess)

- (1/3 of thesis)
- start with a theoretical approach, describe the developed system/algorithm/method from a high-level point of view,
- go ahead in presenting your developments in more detail

1. Benchmark dataset dsprites
2. adaptation with watermark and spurious-to-core feature ratio as an SCM
3. training X models with different ratio, cutoff and learning rate on cluster
4. computing *ground-truth feature importance* of core, spurious and unbiased features: mean logit change for output, R2-score, prediction flip
5. baseline(?) score how much importance is generally assigned to spurious feature (bounding box?)
6. special score for how much importance CRP assigns to concepts encoding spurious feature
7. causal effect estimation? or something like that

find good nam for adapted benchmark dataset

5.1. Causal Benchmark Dataset DSPRITESNEWNAME

Although this is not the first work using a toy dataset with known generating factors to evaluate attribution methods, it still uses a new adaptation of the dSprites dataset [22]. This dataset was originally constructed as a means for testing the degree of disentanglement a method has achieved. It originally contains 737280 64x64 pixel binary images with rectangles, ellipses or hearts in varying positions, scales and rotations. The generating factors `shape`, `scale`, `rotation`, `x position` and `y position` are known for each sample.

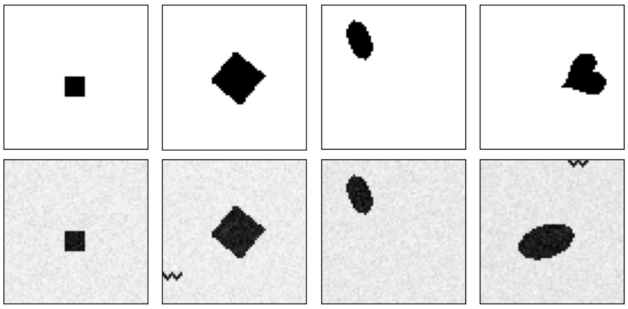

Figure 5.1.: First row: images from the original dSprites dataset, second row: images from the new DSPRITESNEWNAME with small *w* as a watermark on some images and uniform noise added.

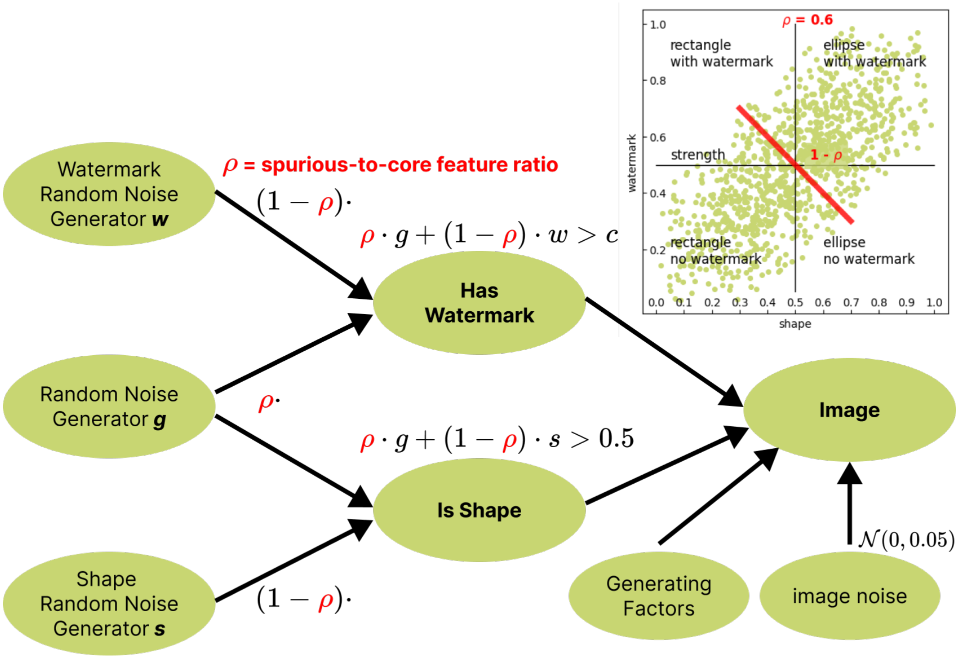

Figure 5.2.: Structural causal model generating the dataset DSPRITESNEWNAME.
In the top right corner the distribution of *Has Watermark* and *Is Shape* are plotted against each other to explain the effect of ρ

To adapt the benchmark for our purpose, only the first two shape classes (rectangle and ellipse) are used. A watermark in the form of a small *w* was initially added to the lower-left corner of some images. During initial testing with only these adaptations it became clear that even the small convolutional neural network employed here is too powerful for this task as effectively dividing the image into two parts solves the problem and most neurons became irrelevant. To make the spurious feature, which is the watermark *w* more difficult to learn, its position is therefore varied across the edges of the image. Further a small uniform noise term is added to make the problem more realistic and the saliency maps more convincing and informative. The aim of this new dataset is, to create the simplest possible scenario with known generating factors, while keeping it as realistic or close to real world application cases of attribution methods as possible. In Figure 5.1 the resulting images are visualized.

why the heck new dataset??

- why do we need another dataset for benchmarking watermark bias??
- some other benchmarks that deal with similar questions are...
- why am i not just using 3d shapes dataset? `https://github.com/deepmind/3dshapes-dataset/` (C. Burgess and H. Kim)

5.1.1. Causal Model

The process with which samples are constructed from the new dataset is a structural causal model.

For the first extensive comparison the SCM as seen in Figure 5.2 serves as a starting point. Explicitly using a generating SCM as previously done by [29, 42, 41] enables us to study the effects of interventions on the model and the explanations. The *spurious-to-core ratio* variable ρ adjusts how much information is shared between the true class information (shape), which we name *core feature* following [34] and the watermark or *spurious feature* through a shared common ancestor g. A second parameter p (= prevalence) determines how prevalent the spurious/watermark feature is in the data. It is important to note that this particular SCM is just one of many possible ways to model how spurious features might interact with core features. It tries to follow the logic of how images are selected in real datasets: Choosing images that have a certain object/feature (here shape), without being aware of some photographers adding watermarks to their images. By some unknown confounding factor (for example personal preference) photographers who add watermarks to their images mostly upload photos of one class of objects. The same SCM applies to many realistic cases of spurious correlations in computer vision tasks. Think of, for example, cows mostly being photographed on pastures or halloween pumpkins mostly at night, creating correlations between the object and background. However a multitude of SCMs potentially act as simplifications of real world spurious correlation scenarios. Selection bias (see bottom picture in Figure 5.3), where a certain combination of features is more likely selected into the dataset and hence makes the data distribution less generalizable to the true distribution, is arguably another often occuring type of bias in computer vision datasets. The direction of causal links for photographs is highly debatable and shall not be the focus of this work. Instead, we only want to find to what degree a neural network learns and attribution method explaind a particular generating SCM.

- have latent factors - can intervene on each factor extensively, expand on usefullness of SCM
- for model we can also assume SCM??? all connected neurons are causally connected
- in given example prediction has shape AND watermark as causal ancestors

- but in real world example spurious feature is only selection bias?
- need to find good causal covering for what i am doing here

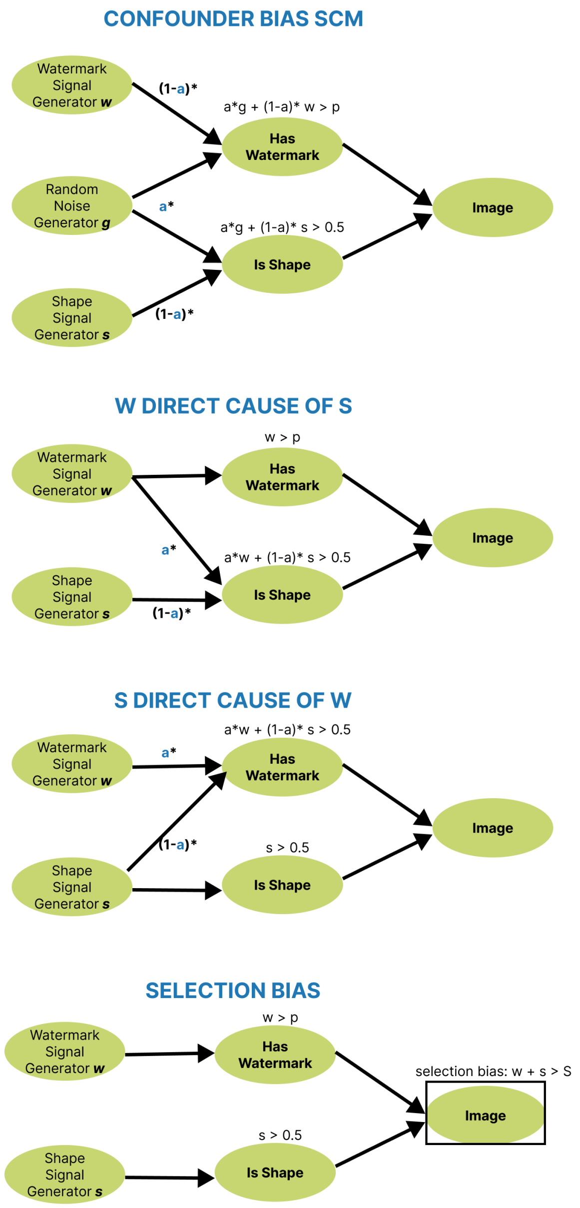

Figure 5.3.: SCMs typically found in image datasets. 1. Our SCM with counfounder g, 2. spurious feature has direct causal effect on core feature, 3. core feature has direct causal effect on spurious feature 4. Selection Bias chooses certain combinations of spurious and core feature with higher probability

e diverse
Ms,
plified, akin
eal world
arios

5.2. CNN Model Zoo

5.2.1. Model Architecture

To evaluate explanations, the model to test on can neither be to simple and therefore easy to explain, nor too large for the simple dataset at hand. Through a simple search the architecture detailed in Listing 5.2.1 with 3 convolutional layer a 8 channels, one linear *concept* layer with 6 neurons and finally the output linear layer was deemed most fitting for the task. While less convolutional channels or layers often resulted in the model not converging at all, having more neurons or potentially *concepts* did not seem to add information but just redundancy. This model reliably yielded test accuracies over 99% when using the same spurious-to-core feature ratio ρ as used for training.

```
convolutional_layers:
    0: Conv2d(in_channels=1, out_channels=8, kernel_size=3)
    1: MaxPool2d(kernel_size=2, stride=2)
    2: ReLU()
    3: Conv2d(in_channels=8, out_channels=8, kernel_size=5)
    4: MaxPool2d(kernel_size=2, stride=2)
    5: ReLU()
    6: Conv2d(in_channels=8, out_channels=8, kernel_size=7)
    7: ReLU()

linear_layers:
    0: Linear(in_features=392, out_features=6, bias=True)
    1: ReLU()
    2: Linear(in_features=6, out_features=2, bias=True)
```

maybe show little example of what purely linear model c achieve?

5.2.2. Hyperparameter Choice

Similar to the size of the model, other hyperparameters are optimized for accuracy. Finally we train all models using the `Adam` optimizer with a learning rate of `0.001` using `cross-entropy` loss as the objective to minimize. It is interesting to note that the learning rate has significantly different optimal values for highly biased models than unbiased ones and we therefore have to chose a compromise. We assume this to be due to the cost function becoming less complex as the trivial watermark feature gains importance. But importantly, those hyperparameters including the learning rate should not be changed over the course of training our set of models because it has been shown that explanation can causally depend on hyperparameters quite strongly [16]. In our experiment we want to keep hyperparameters fixed and only intervene on the spurious-to-core feature ratio ρ or later on all generating factors for establishing ground-truth importance. Another note is that as we are not evaluating the learning process or the model itself but the explanation. The hyperparameters were therefore not in focus of this thesis and chosen rather quickly through some trial-and-error.

cite

cite

5.2.3. Training and Accuracy

We generate datasets by sampling the spurious-to-core feature ratio ρ in 0.05 steps and training on the same dataset while initializing the model with 10 different seeds. Previously, the experiment did not control the influence of the random seed, but after observing strong variations of importance depending on the seed, we decided to always use the same 10 seeds for every dataset. This way it is theoretically possible to marginalize out the effect of the seed on the importance of the spurious feature, however this was done only through averaging of the 10 models results. In total, 210 models are trained. The training dataset contains 30% of all samples. Experimentally, much fewer samples seemed to be enough to achieve high accuracies, however there is no need to fear overfitting as it should if anything increase importance of the most important features even more.

Due to the low complexity of this benchmark dataset, very high accuracies of over 99% are to be expected and also occured after short training for most models.

more about
l, say that
ortance
ngly changes
seed

- training split?
- how many models with which different features are trained
- showing some examples of heatmaps and maxrel images for different bias strength
- accuracies for all models plot

5.2.4. Computational Setup

- computed on personal dell xps 13 with cpu
- and on cluster

ter specs

- how long did training all models take?
- about 40 hours on cluster
- + how long did computation of measures take:
- about 1 minute for all measures per model so 3 and a half hours
- measure time for final method more accurately for one model

t timing of
l method

5.3. Preliminary (Causal) Experiments

ld i include
el scm stuff
attribution
hs etc?

- explain and show idea of a causal model for the whole network
- explain why it didnt work (too high, because linear correlation)
- attribution graph as a causal model???
- ideas and experiments with relevance maximization
- something more along the lines of intervening on hyperparameters? [16]

5.4. Establishing a Ground-Truth of Biasedness

- non-linearity: This can also be explained information-theoretically

explanation for non-linear biasedness - information-theoretically

5.4.1. Accuracy for Subgroups

- is the most 'ground' ground-truth measure of biasedness
- is also somehow related to the others???
- not as exact as mean logit change etc. ?

how is accura related to prediction flip etc.

5.4.2. Prediction Flip, R2 Score and Mean Logit Change

- prediction flip and r2 score are different sides of same coin
- mean logit change is more exact, as sometimes prediction stays the same but gets less confident (logits change a bit in that direction)
- mean logit change shall be used as ground truth of *biasedness*

5.4.3. Interpreting Mean Logit Change as Causal Intervention

- it is basically the causal effect of intervening on a latent factor on the models output
- theoretically the intervention must have exactly the same causal effect on the explanation as on the mean logit change???

5.4.4. Relevance Mass Accuracy (RMA) and Relevance Rank Accuracy (RRA)

In [4] two metrics for the analysis of importance in pixel maps are introduced. Relevance Mass Accuracy (RMA) measures the ratio of relevance within a bounding box around a feature to total relevance. Relevance Rank Accuracy (RRA) the percentage of pixels in such a bounding box that fall within the n most important pixels in the heatmap.

5.5. Measuring Biasedness for Heatmaps

- this is a more general approach of measuring the biasedness of a saliency methods explanation
- good about it: humans look at the heatmaps and only see whether the watermark is colored or not to identify its importance.
- problem: humans have a hard time estimating the overall importance of concepts/features if they have varying spatial extend, see [1] about noses and fur of dog

- so if watermark is even just a little bit red, it will be important to humans
- even bigger problem: NN do not disentangle concepts strictly. therefore the concepts found could always encode watermark and shape feature at the same time. this effect is strongly visible in our benchmark
- question: how much is the result explained by the spurious feature?
- will be taken as the baseline. all other saliency based / local attribution methods can be benchmarked with this too
- does not take into account the splitting up into an relevance of single neurons
- but can in principle also be applied to each neuron/concept individually
- Find a way to measure how well a single heatmap can show the bias
- e.g.: watermark mask importance bilder mit wm general heatmap, total relevance inside mask for:
- A: attribution mit wm, wenn ellipse und conditioned on y:[1]
- B: attribution mit wm, wenn rect und conditioned on y:[0]
- C: attribution ohne wm, wenn ellipse und conditioned on y:[1]
- D: attribution ohne wm, wenn rect und conditioned on y:[0]
- (A - B) + (D - C)
- **LRP biasedness score sanity check:** This sanity test shows that while LRP assigns strong relevance to the watermark, it fails in correctly identifying the lack of a watermark as the main reason to predict for the negative class (rectangle). Superficially this confirms the criticism of missing negative relevance [35]. It is however not clear if the advantage of not cancelling out importances outweighs this factor for more complex data and applications.

irm class-
riance for
maps

5.6. Concepts Biasedness Measures

- should take into account that there are multiple concepts
- one could be important and not assign strong relevance to watermark
- the other could be unimportant and assign strong relevance to watermark
- *ground truth* idea is to again take the mean logit change for each single neuron or summed together somehow
- we want to be able to identify *spurious* concept and *core* concept automatically, so it is not a good idea to have the latent factors given

- one idea: take masked/bounding box approach again for neurons individual heatmaps
- nmf idea: somehow try to reduce the latent space to Watermark/Shape axis and measure variance in either direction
- centroids idea: use random DR algorithm and calculate ratio of centroid distances (needs latent factors again)
- causal idea??? somehow measure causal effect? - the other things are kind of causal or?

make sure to refer to results section too mu rather leave in out if it canno explained wel without lookin at the results

5.7. Measures Temp Latex Notation

Average Causal Effect of Latent Factor on Output

$$ACE = \mathbb{E}[y \mid do(x = 1)] - \mathbb{E}[y \mid do(x = 0)]$$

$$y = \vec{y} = (y_0, y_1), \quad x = \text{has\_watermark } or \text{ is\_ellipse}$$

Mean Absolute Logit Change

$$MLC = \sum_{i \in n} \frac{|\vec{y}_{i,(x=1)} - \vec{y}_{i,(x=0)}|}{n}$$

Average Causal Effect of Latent Factor on Explanation

$$ACE = \mathbb{E}[y \mid do(x = 1)] - \mathbb{E}[y \mid do(x = 0)]$$

$$x = \text{has\_watermark } or \text{ is\_ellipse}$$

But what is explanation y or explanation change $|y_{x=1} - y_{x=0}|$?

For each neuron/concept in a layer:

$$\text{Relevance Mass Accuracy: } RMA = \frac{\sum rel_{watermark}}{\sum rel_{total}}$$

$$\text{Relevance Rank Accuracy: } RRA = \frac{\#top - k\ rel\ in\ watermark}{\#watermark}$$

$$\text{Absolute Relevance Change: } \sum_{i \in layer} |rel_{i,(x=1)} - rel_{i,(x=0)}|$$

$$\text{crp} \perp\!\!\!\perp \rho \mid gt \ ?$$

Relevance Maximization: $\mathcal{T}^{rel}_{\max}(x) = \max_i R_i(x|\theta)$
produces reference set of k-most relevant targets: $\mathcal{X}^{rel}_k i$
Ratio of watermark to shape overlap in sample space:
$$o_i = \{\frac{\#w_a = w_b}{\#s_a = s_b} \mid a, b \in \mathcal{X}^{rel}_k i\}$$
RMA in reference set: $rma_i = \sum_{a \in \mathcal{X}^{rel}_k i} \frac{\sum rel_{watermark_a}}{\sum rel_{total_a}}$
proposed CRP relevance measure: $\max_i (o_i \cdot rma_i)$

6. Experimental Results

10-20 pages

- (1/3 of thesis)
- whatever you have done, you must comment it, compare it to other systems, evaluate it
- usually, adequate graphs help to show the benefits of your approach
- caution: each result/graph must be discussed! what's the reason for this peak or why have you observed this effect

6.1. Experiments

- what have I tried out with the different methods?
- list in concise order the possible measures
- ground-truth feature importance: mean logit change for output, R2-score, prediction flip
- baseline explanation feature importance - thats what we compare to e.g. watermark bounding-box importance for summary heatmap
- special concept explanations feature importance
- how are the experiments set up, how do i make sure they are all well comparable
- to which other baseline could my measures be compared to?

6.2. Results

lots of plots!

- what have I tried out with the different methods?
- what works and what doesn't
- plot for each experiment/possible method?
  - watermark bounding box average relevance for different subgroups, somehow get difference

  - variance of latent factors in relevance maximization image set - low variance means it encodes the concept
  - naive: total relevance for watermark image region
  - total activation + relevance of neuron given just watermark image
  - one idea: take masked/bounding box approach again for neurons individual heatmaps
  - nmf idea: somehow try to reduce the latent space to Watermark/Shape axis and measure variance in either direction
  - centroids idea: use random DR algorithm and calculate ratio of centroid distances (needs latent factors again)
  - causal idea??? somehow measure causal effect? - the other things are kind of causal or?

6.3. Evaluation

- evaluation of evaluation criteria:
  - takes into consideration the whole latent space spanned by the concepts
  - orients itself on known human cognition, user studies in this field would suggest this???
  - performs similar to baseline watermark bounding box importance?
  - ...?

he success
eria of finding
od measure

- which measure is the best according to those criteria
- which measure is the closest to ground truth
- which is the furthest from ground truth
- does measure find *more information* than CRP itself and could possibly be used as a method on top of CRP for disentanglement/ spurious-core relation explanation?
-

6.4. Verification on Other Well-Known Benchmarks

uld i do
? which
chmarks, how
etup causal
lel for them

1. Test method on more complex dataset e.g. CLEVR-XAI
2. compare CRP to other XAI methods?

6.5. Discussion

- do measures work
- what does causality help us with
- is CRP better for constant vector shift stuff or does it still suffer from it?
- can the application of those measures further explain/inform the explanation?
- what failed miserably

7. Conclusion

- (1 page)
- summarize again what your paper did, but now emphasize more the results, and comparisons.
- write conclusions that can be drawn from the results found and the discussion presented in the paper.
- future work (be very brief, explain what, but not much how)

1. which ground-truth feature importance measure is best
2. interesting points looking at model performances and explanation perforcmance?
3. baseline vs. concept metric, which one is better for CRP
4. performance on other cases
5. how has causality helped in this?
6. Answer Question: *ARE WE EXPLAINING THE DATA OR THE MODEL* answer questi

Limitations and Future Work

- max 1-2 pages, could also be included into the conclusion section

References

[1] Achtibat, R., Dreyer, M., Eisenbraun, I., Bosse, S., Wiegand, T., Samek, W., and Lapuschkin, S. From "where" to "what": Towards human-understandable explanations through concept relevance propagation, 2022.

[2] Achtibat, R., Dreyer, M., Eisenbraun, I., Bosse, S., Wiegand, T., Samek, W., and Lapuschkin, S. From attribution maps to human-understandable explanations through concept relevance propagation. *Nature Machine Intelligence 5*, 9 (jul 2023), 1006 – 1019.

[3] 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* (2018), S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, Eds., vol. 31, Curran Associates, Inc.

[4] Arras, L., Osman, A., and Samek, W. Clevr-xai: A benchmark dataset for the ground truth evaluation of neural network explanations. *Information Fusion 81* (2022), 14–40.

[5] 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 (07 2015), 1–46.

[6] Bau, D., Zhou, B., Khosla, A., Oliva, A., and Torralba, A. Network dissection: Quantifying interpretability of deep visual representations. In *Computer Vision and Pattern Recognition* (2017).

[7] Bau, D., Zhu, J.-Y., Strobelt, H., Lapedriza, A., Zhou, B., and Torralba, A. Understanding the role of individual units in a deep neural network. *Proceedings of the National Academy of Sciences 117*, 48 (2020), 30071–30078.

[8] Blücher, S., Vielhaben, J., and Strodthoff, N. Preddiff: Explanations and interactions from conditional expectations. *Artificial Intelligence 312* (Nov. 2022), 103774.

[9] Chattopadhyay, A., Manupriya, P., Sarkar, A., and Balasubramanian, V. N. Neural network attributions: A causal perspective. *ArXiv abs/1902.02302* (2019).

[10] CHORMAI, P., HERRMANN, J., MÜLLER, K.-R., AND MONTAVON, G. Disentangled explanations of neural network predictions by finding relevant subspaces.

[11] CLARK, B., WILMING, R., AND HAUFE, S. Xai-tris: Non-linear benchmarks to quantify ml explanation performance. *ArXiv* (2023).

[12] DREYER, M., ACHTIBAT, R., WIEGAND, T., SAMEK, W., AND LAPUSCHKIN, S. Revealing hidden context bias in segmentation and object detection through concept-specific explanations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops* (June 2023), pp. 3828–3838.

[13] DREYER, M., PAHDE, F., ANDERS, C. J., SAMEK, W., AND LAPUSCHKIN, S. From hope to safety: Unlearning biases of deep models by enforcing the right reasons in latent space, 2023.

[14] GOYAL, Y., FEDER, A., SHALIT, U., AND KIM, B. Explaining Classifiers with Causal Concept Effect (CaCE). *ArXiv* (July 2019), arXiv:1907.07165.

[15] HOHMAN, F., PARK, H., ROBINSON, C., AND POLO CHAU, D. H. Summit: Scaling deep learning interpretability by visualizing activation and attribution summarizations. *IEEE Transactions on Visualization and Computer Graphics 26*, 1 (2020), 1096–1106.

[16] KARIMI, A.-H., MUANDET, K., KORNBLITH, S., SCHÖLKOPF, B., AND KIM, B. On the relationship between explanation and prediction: A causal view. In *Proceedings of the 40th International Conference on Machine Learning* (23–29 Jul 2023), A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, and J. Scarlett, Eds., vol. 202 of *Proceedings of Machine Learning Research*, PMLR, pp. 15861–15883.

[17] KIM, B., WATTENBERG, M., GILMER, J., CAI, C., WEXLER, J., VIEGAS, F., AND SAYRES, R. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (TCAV). In *Proceedings of the 35th International Conference on Machine Learning* (10–15 Jul 2018), J. Dy and A. Krause, Eds., vol. 80 of *Proceedings of Machine Learning Research*, PMLR, pp. 2668–2677.

[18] KINDERMANS, P.-J., HOOKER, S., ADEBAYO, J., ALBER, M., SCHÜTT, K. T., DÄHNE, S., ERHAN, D., AND KIM, B. *The (Un)reliability of Saliency Methods*. Springer International Publishing, Cham, 2019, pp. 267–280.

[19] KOHLBRENNER, M., BAUER, A., NAKAJIMA, S., BINDER, A., SAMEK, W., AND LAPUSCHKIN, S. Towards best practice in explaining neural network decisions with lrp. In *2020 International Joint Conference on Neural Networks (IJCNN)* (2020), pp. 1–7.

[20] LEEMANN, T., KIRCHHOF, M., RONG, Y., KASNECI, E., AND KASNECI, G. When are post-hoc conceptual explanations identifiable? In *Proceedings of the Thirty-Ninth Conference on Uncertainty in Artificial Intelligence* (31 Jul–04 Aug 2023), R. J. Evans and I. Shpitser, Eds., vol. 216 of *Proceedings of Machine Learning Research*, PMLR, pp. 1207–1218.

[21] LEEMANN, T., RONG, Y., KRAFT, S., KASNECI, E., AND KASNECI, G. Coherence evaluation of visual concepts with objects and language. In *ICLR2022 Workshop on the Elements of Reasoning: Objects, Structure and Causality* (2022).

[22] MATTHEY, L., HIGGINS, I., HASSABIS, D., AND LERCHNER, A. dsprites: Disentanglement testing sprites dataset. https://github.com/deepmind/dsprites-dataset/, 2017.

[23] MONTAVON, G., BINDER, A., LAPUSCHKIN, S., SAMEK, W., AND MÜLLER, K.-R. *Layer-Wise Relevance Propagation: An Overview*. Springer International Publishing, Cham, 2019, pp. 193–209.

[24] MONTAVON, G., LAPUSCHKIN, S., BINDER, A., SAMEK, W., AND MÜLLER, K.-R. Explaining nonlinear classification decisions with deep taylor decomposition. *Pattern Recognition 65* (May 2017), 211–222.

[25] MORAFFAH, R., KARAMI, M., GUO, R., RAGLIN, A., AND LIU, H. Causal interpretability for machine learning - problems, methods and evaluation. *SIGKDD Explor. Newsl. 22*, 1 (may 2020), 18–33.

[26] NARENDRA, T., SANKARAN, A., VIJAYKEERTHY, D., AND MANI, S. Explaining deep learning models using causal inference. *ArXiv abs/1811.04376* (2018).

[27] OLAH, C., MORDVINTSEV, A., AND SCHUBERT, L. Feature visualization. *Distill* (2017). https://distill.pub/2017/feature-visualization.

[28] PAHDE, F., DREYER, M., SAMEK, W., AND LAPUSCHKIN, S. Reveal to revise: An explainable ai life cycle for iterative bias correction of deep models, 2023.

[29] PARAFITA, A., AND VITRIA, J. Explaining visual models by causal attribution. In *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)* (2019), pp. 4167–4175.

[30] RONG, Y., LEEMANN, T., BORISOV, V., KASNECI, G., AND KASNECI, E. A consistent and efficient evaluation strategy for attribution methods. In *Proceedings of the 39th International Conference on Machine Learning* (17–23 Jul 2022), K. Chaudhuri, S. Jegelka, L. Song, C. Szepesvari, G. Niu, and S. Sabato, Eds., vol. 162 of *Proceedings of Machine Learning Research*, PMLR, pp. 18770–18795.

[31] Rong, Y., Leemann, T., trang Nguyen, T., Fiedler, L., Qian, P., Unhelkar, V., Seidel, T., Kasneci, G., and Kasneci, E. Towards human-centered explainable ai: A survey of user studies for model explanations, 2023.

[32] Samek, W., Montavon, G., Lapuschkin, S., Anders, C. J., and Müller, K.-R. Explaining deep neural networks and beyond: A review of methods and applications. *Proceedings of the IEEE 109*, 3 (2021), 247–278.

[33] Schölkopf, B., for Intelligent Systems, M. P. I., Max-Planck-Ring, Tübingen, ., and Germany. Causality for machine learning. Tech. Rep. nature., 2019.

[34] Singla, S., and Feizi, S. Salient image net: How to discover spurious features in deep learning? In *ICLR* (2022).

[35] Sixt, L., Granz, M., and Landgraf, T. When explanations lie: Why many modified bp attributions fail. In *Proceedings of the 37th International Conference on Machine Learning* (2020), ICML'20, JMLR.org.

[36] Sixt, L., and Landgraf, T. A rigorous study of the deep taylor decomposition, 2022.

[37] Sixt, L., Schuessler, M., Popescu, O.-I., Weiss, P., and Landgraf, T. Do users benefit from interpretable vision? a user study, baseline, and dataset, 2022.

[38] Tran, T. Q., Fukuchi, K., Akimoto, Y., and Sakuma, J. Unsupervised causal binary concepts discovery with vae for black-box model explanation. *Proceedings of the AAAI Conference on Artificial Intelligence 36*, 9 (Jun. 2022), 9614–9622.

[39] Vielhaben, J., Blücher, S., and Strodthoff, N. Sparse subspace clustering for concept discovery (ssccd), 2022.

[40] Vielhaben, J., Bluecher, S., and Strodthoff, N. Multi-dimensional concept discovery (MCD): A unifying framework with completeness guarantees. *Transactions on Machine Learning Research* (2023).

[41] Wilming, R., Kieslich, L., Clark, B., and Haufe, S. Theoretical behavior of XAI methods in the presence of suppressor variables. In *Proceedings of the 40th International Conference on Machine Learning* (23–29 Jul 2023), A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, and J. Scarlett, Eds., vol. 202 of *Proceedings of Machine Learning Research*, PMLR, pp. 37091–37107.

[42] Yang, M., and Kim, B. Benchmarking attribution methods with relative feature importance, 2019.

[43] YEOM, S., SEEGERER, P., LAPUSCHKIN, S., WIEDEMANN, S., MÜLLER, K., AND SAMEK, W. Pruning by explaining: A novel criterion for deep neural network pruning. vol. abs/1912.08881.

A. Appendix

A.1. Additional Details to LRP rules and implementation best practices

Figure A.1.: This is a test figure

A.2. Preliminary Experiments

A.2.1. Plots

A.2.2. Causal Discovery on Neural Network Models Idea and Implementation?

Figure A.2.: This is a test figure

A.3. Details on Model Architecture?

```
self.convolutional_layers = nn.Sequential(
  nn.Conv2d(1, 8, kernel_size=3, stride=1, padding=0),
  nn.MaxPool2d(kernel_size=2, stride=2),
  nn.ReLU(),
  nn.Conv2d(8, 8, kernel_size=5, stride=1, padding=0),
  nn.MaxPool2d(kernel_size=2, stride=2),
  nn.ReLU(),
  nn.Conv2d(8, 8, kernel_size=7, stride=1, padding=0),
  nn.ReLU(),
)
self.linear_layers = nn.Sequential(
  nn.Linear(392, 6),
  nn.ReLU(),
  nn.Linear(6, 2),
)
```

A.4. Further Plots Groud Truth