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, November 9, 2023

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 network 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 biasedness of a dataset interacts with the biasedness of the model and the biasedness of the 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 have importance assigned. Often, the spurious feature comes as a watermark which is easy to identify. Consequently its importance can be overestimated when looking at a general heatmap of an image.
- 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.

d

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

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 seems to learn the bias strongly (check for accuracy)
- as preliminary experiment check, if the bias is strongly visible in the data: if the heatmaps/crp hierarchies produced on average for the watermarked/unwatermarked subsets differ strongly
- *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

1. Introduction 1
  1.1. Motivation and Context 1
  1.2. Strategy 3
  1.3. Outline 3

2. Related Work 5
    2.0.1. XAI in general 5
    2.0.2. Layerwise Relevance Propagation 5
    2.0.3. Deep Taylor Decomposition 6
    2.0.4. Concept Relevance Propagation (CRP) 6
    2.0.5. Evaluation of Back-Propagation XAI Methods 6
    2.0.6. Criticism of Back-Propagation XAI Methods 6
    2.0.7. Causal Discovery for XAI 7
  2.1. Critique on Saliency Maps Summary 7

3. Background 9
  3.1. Concept Relevance Propagation 9
  3.2. Evaluation of Explanations 9
  3.3. Causal Framework 10

4. Methods 11
  4.1. Causal Benchmark DSPRITES 11
  4.2. Should I include Model SCM experiments i did? 13
  4.3. CNN Model Zoo 13
  4.4. Ground-Truth Biasedness 13
  4.5. Baseline Biasedness Measure 13
  4.6. Concepts Biasedness Measures 14
  4.7. Setup 14

5. Results 15
  5.1. Experiments 15
  5.2. Discussion 16

6. Conclusion and Outlook 17

References 18

A. Appendix 23
  A.1. Test Figure 23
  A.2. Test Figure 2 23

List of Figures

4.1. go more into detail about why and which SCM and what to expect from ’real’ data 12

A.1. Test Figure 23
A.2. Test Figure 2 24

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 [4] 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 [28]. 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.

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 [28] 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* [18]. 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 [9]. 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 [26].)

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 wha 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 network on recognizing shapes
- evaluate CRP on neural network trained on biased and unbiased dataset

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 4 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 5 and discussed in chapter 6.

2. Related Work

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. [27], [31], [14] 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?

2.0.1. XAI in general

- focus on post-hoc explanations + theoretical foundations, test algorithms [25]
- creates new dataset (human-supervised) to detect core vs spurious features [26]

2.0.2. Layerwise Relevance Propagation

approach used in the thesis, tell which rules of LRP are used

- LRP first paper [4]
- overview of propagation rules [19]
- general XAI [25]
- LRP in practice [15]
- disentangle representations, similar to PCA: Principal Relevant Component Analysis - uses LRP [8]
- LRP is also good at pruning... idea/intuition: if you can "prune" certain neurons, their causal effect must be none or extremely small [32]

2.0.3. Deep Taylor Decomposition

maybe just mention shortly how it is related to LRP (with rules) and which limitations have been studied e.g. by Sixt and Landgraf. Think again whether deep taylor decomposition has any more ties with causality???

- theoretical explanation to how LRP/DTD works...? [20]
- deep taylor decomposition fails, when only positive relevance taken into account, matrix falls to rank 1 [28]

2.0.4. Concept Relevance Propagation (CRP)

Only state when and where it has been described developed. refer to background section for theory and definition

- from where to what... CRP main paper [1]
- reveal to revise: whole framework for XAI using CRP as one of the methods for concept/bias discovery [23]
- using CRP to identify and unlearn bias 'Right Reason Class Artifact Compensation (RR-ClArC)' [10]

2.0.5. Evaluation of Back-Propagation XAI Methods

important to note, as 'causal' perspective could add another evaluation method to corpus

- clever XAI artificial benchmark dataset [3]
- NetDissect dataset with concept-segmented images [5]
- also other concept/neuron dissection by same authors, similar idea to CRP [6]

2.0.6. Criticism of Back-Propagation XAI Methods

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

- explanations are independent of later layers (no negative relevance) [27]
- 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 [31]
- 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) [14]

2.0.7. Causal Discovery for XAI

which other methods/approaches/papers are there that broadly connect explainable AI and causality

- generally, mostly about counterfactuals: [21]
- causal attribution, similar to LRP but more "causally" neural networks as SCMs [7]
- causal concept effects (edges in mnist) [11]
- causal in most general sense:independent/disjoint mechanism analysis [16] [17]
- causal binary concepts [30]
- basic framework/idea of interpreting NN as skeleton of SCM and using some transformation to quantify effect:[22]

2.1. Critique on Saliency Maps Summary

1. [29]: 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. [31]: 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. [27]: 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. [14]: 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. [12]: 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. [2]: 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. [24]: 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. [22]: build SCM over input-model-output -> has potential to be more accurate than saliency purely observational

9. [7]: 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

3. Background

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

write backgro

3.1. Concept Relevance Propagation

- brief explanation of LRP and Deep Taylor Decomposition
- backpropagation rules etc?
- theoretical idea of Concept Relevance Propagation and what it seeks to improve
- some examples of usage:
  - relevance scores for *concepts* (=neurons)
  - relevance maximization images
  - conditioning on single concepts/ neurons ...?
  - attribution graph
- recent criticism of LRP / local attribution methods

understand all
rules etc of
LRP/CRP

3.2. Evaluation of Explanations

- differentiate between numerical evaluation and evaluation through user studies
- make clear: human understanding is the ultimate goal, so user studies are the gold standard (but often not well done)
- in response to [27]
- examples of often used evaluations for local attribution methods and concept-based methods:
  - feature ablation and related methods
  - TCAVs [13] with benchmark feature set (hard and often not applicable)
  - clevr-xai? [3]

3.3. Causal Framework

- Explain Structural Causal Models
- Short introduction to causal effects?
- AI / neural networks in general do not care about causation and work through finding useful correlations
- none-the-less the better we get at identifying spurious features the more causal methods might apply?
- causal methods have been hardly used for evaluation of XAI, this is easier as we do not expect the model to learn causal things but we assume the model itself is part of an SCM

4. Methods

- (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

4.1. Causal Benchmark DSPRITES

- Benchmark dataset dsprites
- adaptation with watermark and spurious-to-core feature ratio as an SCM
- causal effect stuff: see fig. 4.1
  - have latent factors - can intervene on each factor extensively
  - 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

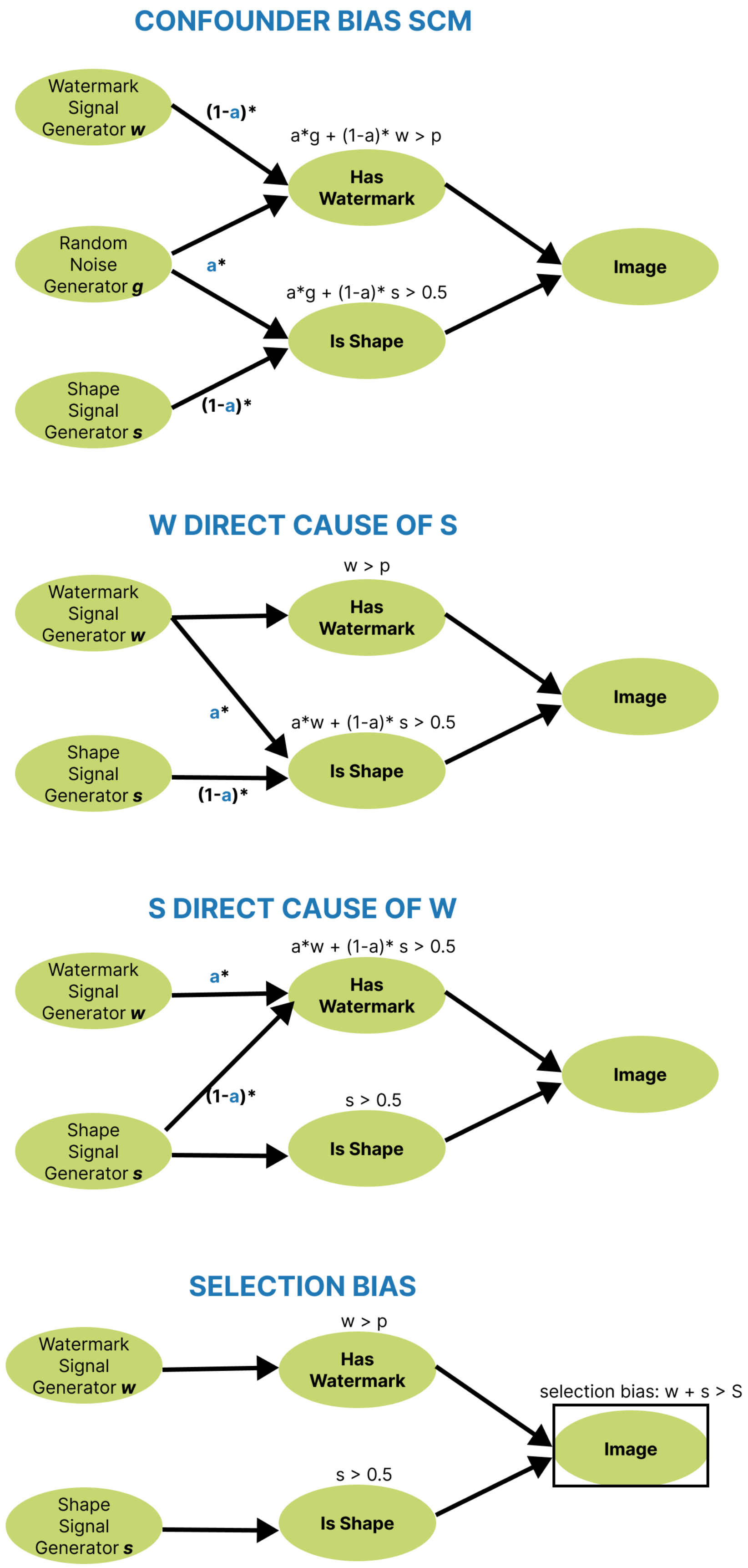

Figure 4.1.: go more into detail about why and which SCM and what to expect from 'real' data

4.2. Should I include Model SCM experiments i did?

should i inclu
model scm stu
and sttribution
graphs etc?

4.3. CNN Model Zoo

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

4.4. Ground-Truth Biasedness

- Prediction Flip
- R2 score
- mean logit change
- non-linearity: This can also be explained information-theoretically

explanation
for non-linear
biasedness -
information-theoretically

4.5. Baseline Biasedness Measure

- 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)

4.6. 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?

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 [27]. It is however not clear if the advantage of not cancelling out importances outweighs this factor for more complex data and applications.

e sure to not
r to results
ion too much.
er leave info
if it cannot
ained well
out looking
e results

irm class-
riance for
maps

4.7. Setup

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

ter specs

5. Results

- (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

1. ground-truth feature importance: mean logit change for output, R2-score, prediction flip
2. baseline explanation feature importance - thats what we compare to e.g. watermark bounding-box importance for summary heatmap
3. special concept explanations feature importance
4. Test method on more complex dataset e.g. CLEVR-XAI
5. compare CRP to other XAI methods?

5.1. Experiments

- 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?

5.2. Discussion

6. Conclusion and Outlook

- (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. limitations, outlook and future work
6. how has causality helped in this?

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] 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.

[3] 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.

[4] 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.

[5] 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).

[6] 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* (2020).

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

[8] Chormai, P., Herrmann, J., Müller, K.-R., and Montavon, G. Disentangled explanations of neural network predictions by finding relevant subspaces.

[9] Clark, B., Wilming, R., and Haufe, S. Xai-tris: Non-linear benchmarks to quantify ml explanation performance. *ArXiv* (2023).

[10] 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.

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

[12] 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.

[13] 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.

[14] 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.

[15] 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.

[16] 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.

[17] 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).

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

[19] 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.

[20] 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.

[21] 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.

[22] Narendra, T., Sankaran, A., Vijaykeerthy, D., and Mani, S. Explaining deep learning models using causal inference. *ArXiv abs/1811.04376* (2018).

[23] 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.

[24] 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.

[25] 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.

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

[27] 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.

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

[29] 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.

[30] 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.

[31] 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.

[32] 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. Test Figure

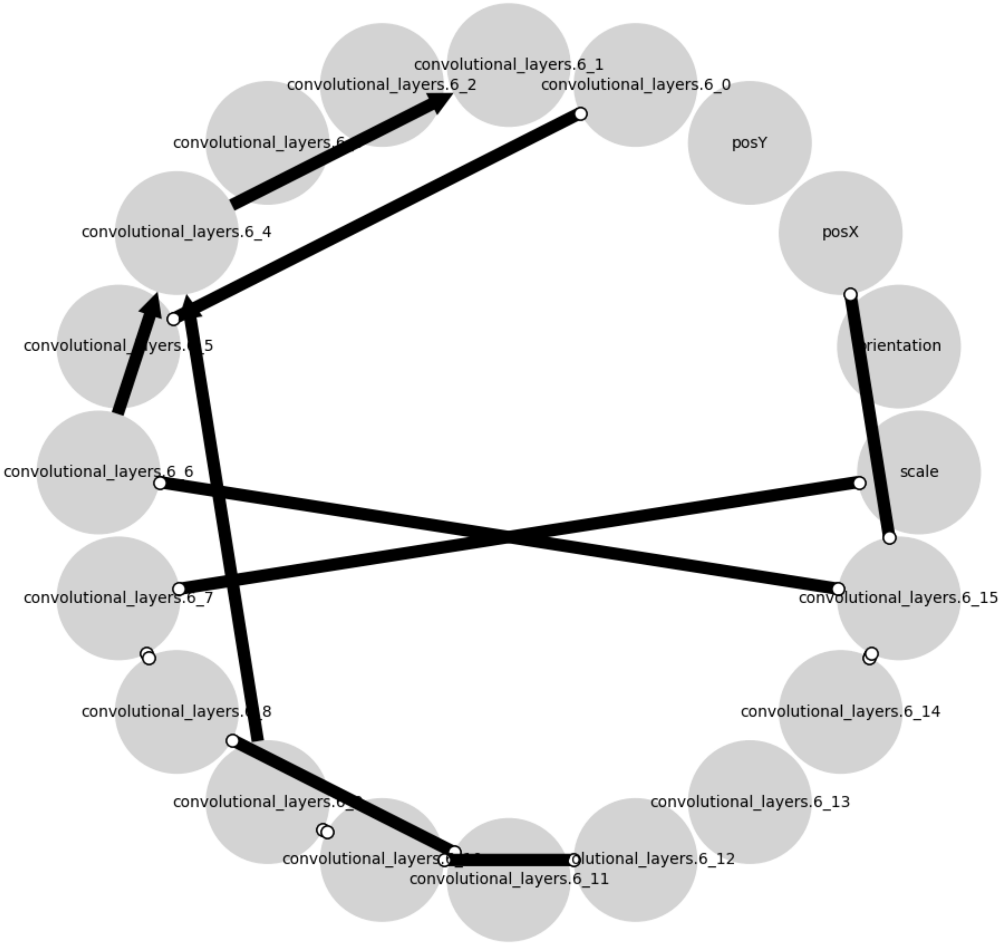

Figure A.1.: This is a test figure

A.2. Test Figure 2

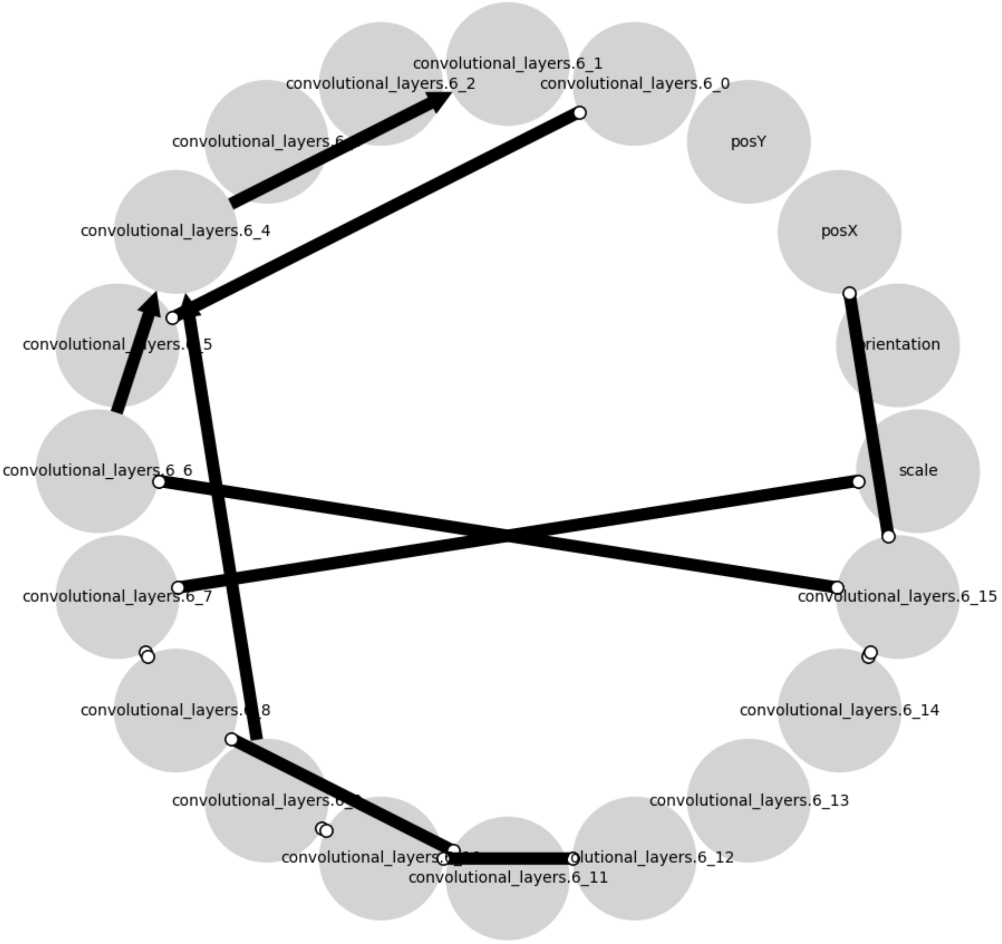

Figure A.2.: This is a test figure