Revisiting Robust Interpretability of Self-Explaining Neural Networks

Joosje Goedhart
10738193

Lennert Jansen
10488952

Hannah Lim
10588973

Daniel Nobbe
12891517

ABSTRACT

As machine learning arises in various applications, the demand for better understandings of machine learning models' prediction increases. Self-explaining models are models that provide human interpretable explanations. In this study we revisit the approach proposed by Melis & Jaakkola (2018) that incorporates interpretability directly into the structure of the model, resulting in self-explaining neural networks, and we propose an adjustment to their approach for improvement of interpretability. The proposed adjusted approach is measured by comparing it with the original approach based on three desiderata: explicitness, faithfulness and stability. Our work improves in faithfulness, provides a different and possibly more intuitive perspective on intelligibility, but slightly deteriorates stability. The third can be improved by adding a local-Lipschitz stability property for the concepts variable.

KEYWORDS

Self-Explaining Models, explainability, interpretability, stability

ACM Reference Format:
Joosje Goedhart, Lennert Jansen, Hannah Lim, and Daniel Nobbe. 2018. Revisiting Robust Interpretability of Self-Explaining Neural Networks. In *Proceedings of ACM Conference (Conference'17).* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/1122445.1122456

1 INTRODUCTION

As machine learning is deployed in an increasing variety of applications, ranging from commerce [16] to medicine [4] to law [2], better understandings of machine learning models' predictions are in high demand. This has led to a growing research community that focuses on the explainability or interpretability of machine learning techniques. The majority of explanation methods focuses on *post-hoc interpretability*, i.e. understanding (e.g. in terms of input variables) the predictions (e.g. categories) of a readily trained model [13], [8], [15], [1]. However, it has been shown that small shifts in the input features can cause significant changes in post-hoc explanations, whereas the model output often remains the same [11], [7]. [14] argues that the fragility of these methods can be due to the explanations themselves being fragile models. Melis & Jaakkola [11] recently proposed to incorporate interpretability directly into the structure of the model, resulting in self-explaining neural networks (SENNs). SENNs are designed to provide robust interpretations for classification tasks. This is done by learning $\boldsymbol{\theta}(\boldsymbol{x})$, a vector of relevance coefficients per class to be predicted, and $\boldsymbol{h}(\boldsymbol{x})$, a set of learned concept values that are mapped to interpretable basis concepts (e.g strokes and colors, in the form of highlighted text, heatmaps highlighting important pixels [12]). [11] argues that transparency in the models' reasoning process is guaranteed, as the explanations, $\boldsymbol{\theta}(\boldsymbol{x})$, are used to produce the models' predictions $\boldsymbol{f}(\boldsymbol{x}) = \boldsymbol{\theta}(\boldsymbol{x})^T\boldsymbol{h}(\boldsymbol{x})$. The allowed variation in interpretation coefficients is restricted around each input to ensure local stability. [11] defines three desiderata to measure the robust interpretability of SENN: explicitness (i.e. if the explanations are immediate and understandable), faithfulness (i.e. if the relevance scores are truly relevant for the prediction outcome) and stability (i.e. if the explanations are robust to local perturbations of the input). These desiderata will be further explained in section 2.2.

[11] argues that a high positive relevance coefficient value corresponds to a positive contribution of the corresponding concept $\boldsymbol{h}(\boldsymbol{x})$ on the models prediction and vice versa. This interpretation of the relevance coefficients is flawed for two reasons. First, it doesn't take the *sign* of the concept value $\boldsymbol{h}(\boldsymbol{x})$ into account. For example, a positive relevance coefficient with a corresponding negative concept value has a negative influence on the model prediction, whereas the relevance coefficient on itself suggests a positive contribution. Second, this approach doesn't take the *magnitude* of the concept values into account. For example a negative relevance coefficient with a corresponding concept value close to zero, has a marginal influence on the models prediction, whereas the relevance coefficient on itself suggests a strong negative contribution.

The aim of this study is twofold: first we investigate to what extent the results of [11] are reproducible. Second we define an alternative definition of the relevance coefficients that incorporates both the sign and the magnitude of the concept values into the relevance scores. We analyse our definition of relevance scores on the three defined desiderata for robust interpretability, explicitness, faithfulness and stability, and compare the results to the original relevance scores in [11]. We find that our definition of relevance scores leads to a high increase in the faithfulness, but makes the explanations less stable to local perturbations. The intelligibility, a qualitative measure, arguably achieves a more intuitive perspective.

The remainder of this paper is organised as follows, we first introduce self-explaining neural networks (SENNs) after which we explain the measure of interpretability for evaluation of the SENNs. Then, in section 3, we explain the method of the original approach [11] and its limitations followed by our proposal for improvement

Conference'17, July 2017, Washington, DC, USA
© 2018 Association for Computing Machinery.
ACM ISBN 978-x-xxxx-xxxx-x/YY/MM...$15.00
https://doi.org/10.1145/1122445.1122456

with its experimental setup. Finally, we summarise our findings and compare both approaches and propose an alternative to improve the stability of our approach.

2 THEORY

2.1 Self-explaining neural networks

Self-explaining neural networks are self-explaining models (hereafter SEMs) where $\boldsymbol{\theta(x)}$ is realised with a neural network. SEMs are a rich class of complex models that provide human interpretable explanations for their prediction outcomes with varying levels of complexity[11]. These explanations, formulated in terms of interpretable basis concepts and relevance scores, are built into the model's architecture from the ground up. Providing clarifications of prediction results is thereby achieved without a separate optimisation sub-routine. Interpretable basis concepts are generalisations of input features, expressed as functions of the raw features, where the resulting concepts represent human-digestible manifestations of the typically less intelligible input space, e.g., strokes and blobs in lieu of single pixels. Naturally, the number of concepts, denoted by k, should be chosen to be small, so as to avoid obfuscation. The corresponding influence scores, $\boldsymbol{\theta(x)} \in \mathbb{R}^{k \times m}$, serve as indications of the magnitude and direction (i.e., positive or negative) of the relationship between the prediction outcome and the concept in question for each class m.

The generalised form of a SEM is expressed as follows:

$$\boldsymbol{f(x)} = g\Big(\theta_1(\boldsymbol{x})h_1(\boldsymbol{x}), ..., \theta_k(\boldsymbol{x})h_k(\boldsymbol{x})\Big), \quad (1)$$

where $g()$ is typically a summation, as is the case in this study.

Requiring $\boldsymbol{\theta(x)}$ to be *locally difference bounded* by $\boldsymbol{h(x)}$ is essential for ensuring SEMs' inherent explainability.

Definition 1. *θ is said to be locally difference bounded by h if for all x_0 there exist a $\delta > 0$ and $L \in \mathbb{R}$ such that* $||\boldsymbol{x} - \boldsymbol{x_0}|| \leq L||\boldsymbol{h}(\boldsymbol{x}) - \boldsymbol{h}(\boldsymbol{x_0})||$ *[11].*

In essence this limits the slope of the relevance score with respect to the input values in a certain region around $\boldsymbol{x}$. This notion of stability of $\boldsymbol{\theta(x)}$ can be incorporated during training time via a regularisation parameter λ. For a detailed explanation of how this is realised, the reader is referred to Sections 3 and 4 of [11].

The SENN proposed by [11] represent both convolutional and fully connected neural networks as SEMs, outputting a linear combination of $\boldsymbol{\theta}(x)$ and $\boldsymbol{h}(x)$.

2.2 Interpretability

2.2.1 Desiderata of interpretability. Melis & Jaakkola (2018) [11] propose three criteria to evaluate the interpretability of self-explaining neural networks: explicitness, faithfulness and stability, each of which we explain in detail below. It should be mentioned that the investigated datasets perform on par with their non-interpretable counterparts in terms of prediction accuracy. As our proposed interpretation alternative does not change the accuracy of a SENN, we do not focus on accuracy in our analysis.

Explicitness

The explicitness of a model evaluates the degree of intelligibility of the explanations, i.e. are the explanations understandable and immediate. The explanations of a model can differ for a similar task (e.g. for the MNIST dataset most methods, e.g. [8], [13], and [15], use heatmap highlighting of the important pixels while [11] use relevance coefficients $\boldsymbol{\theta}(x)$ as explanation).

Faithfulness

The faithfulness of a model evaluates if the model's returned relevance coefficients truly indicate the relevance for the prediction. This is simply assessed by looking at the effect on the prediction probability when setting the features in order of high relevance score to low relevance score to non-active. Following this approach, a faithful model should give a high drop in prediction probability if features that are indicated as highly relevant are excluded.

Stability

According to [11] an important property of interpretability is stability meaning that the explanation of the model should be robust to local perturbations of the input. Previous post-hoc models do not have this crucial property, therefore [11] added a point-wise, neighbourhood-based local Lipschitz continuity that the relevance scores of close inputs do not differ significantly. In this study the stability is measured at every point in the test data by adding white noise with a standard deviation $\leq$ 0.2, such that the prediction remains the same. Therefore it is wanted that the explanation resembles between the points with and without added noise. The difference between two points is measured by comparing the norm of the relevance coefficients.

3 METHOD

The intuition behind the class of self-explaining neural networks as defined in [11] is based on the principle of a simple linear regression model: $\boldsymbol{f(x)} = \sum_i^k \theta_i x_i + \theta_0$, where coefficients $\theta_1, ..., \theta_k$ model the impact of features $x_1, ..., x_k$ in an easily interpretable manner. This model is substantially enriched. First by allowing the coefficients $\boldsymbol{\theta}$ themselves to depend on $\boldsymbol{x}$, where $\theta : X \to \mathbb{R}^{kxm}$, with m the number of classes to predict and k the number of learned concepts, that is a hyperparameter. Second, in order to deal with less interpretable types of data, e.g. images, they replace the raw input features x with learned interpretable basis concepts $h : X \to Z \subset \mathbb{R}^k$, where Z is some set of human interpretable atoms.

In the case of images, h is learned via an auto-encoder with convolutional layers. An interpretation of each concept dimension is given by representing each dimension by the elements in a sample of data that maximise their value, i.e. each concept i is represented via the set $X^i = argmax_{\bar{X} \subseteq X, |\bar{X}|} \sum_{x \in \bar{X}} h_{enc}(x)_i$ with small l. In the case of more natural data, input features $\boldsymbol{x}$ can simply be used as concepts $\boldsymbol{h(x)}$.

The basis concepts are then used to make predictions as $\boldsymbol{\theta(x)}^T \boldsymbol{h_{enc}}(\boldsymbol{x})$ or as $\boldsymbol{\theta(x)}^T \boldsymbol{x}$. The stability of $\boldsymbol{\theta(x)}$ and the quality of the learned concepts are guaranteed by imposing two training loss reguralizers. If $\boldsymbol{\theta(x)}$ is learned via a deep neural network, [11] define this class of models as self-explaining neural networks

3.1 Limitations of SENN

The model proposed by [11] makes predictions according to the following equation:

$$\boldsymbol{f(x)} = \boldsymbol{\theta(x)}^T \boldsymbol{h(x)} \quad (2)$$

where $\boldsymbol{f}(\boldsymbol{x})$ is a vector of size $mx1$ containing the predicted probability for each class. During classification a sample $\boldsymbol{x}$ is assigned to the class $\hat{t}$ that has the highest value (i.e. $\hat{t}$ = argmax $\boldsymbol{f}(x)$). The paper posits that interpretation can be done through the relevance coefficients $\boldsymbol{\theta}_{\hat{t}}(x)$ corresponding to the target class and state that concept i $\boldsymbol{h}_i(x)$ is active when the corresponding relevance coefficient is strongly positive. However, explanations purely based on the relevance coefficients $\boldsymbol{\theta}_{\hat{t}}(x)$ are not valid since they do not take the sign and magnitude of $\boldsymbol{h}(x)$ into account. For example, for a sample $\boldsymbol{x}$ with the concept $h_i(x)$ is close to zero, but the corresponding relevance coefficient $\boldsymbol{\theta}_{\hat{t},i}(x)$ is close to 1,[10] suggests that the concept h_i is a meaningful concept for the explanation of the prediction of $\boldsymbol{x}$ while in fact the concept barely had an effect on the prediction.

3.2 Proposed adjustment to improve SENN

In this paper, we propose an alternative definition of relevance coefficients, namely by replacing the relevance coefficients $\boldsymbol{\theta}(\boldsymbol{x})$ with the relevance coefficients $\boldsymbol{r}(\boldsymbol{x})$ as defined in the equation below. This alternative definition incorporates the magnitude and sign of the concept values.

$$\boldsymbol{r}(\boldsymbol{x}) = \left[\boldsymbol{\theta}(\boldsymbol{x})_{1,:} \odot \boldsymbol{h}(\boldsymbol{x}) ... \boldsymbol{\theta}(\boldsymbol{x})_{m,:} \odot \boldsymbol{h}(\boldsymbol{x})\right] \tag{3}$$

where $\odot$ is the Hadamard product, $\boldsymbol{\theta}(\boldsymbol{x})_{i,:} \in \mathbb{R}^k$ is the vector of relevance scores for the i-th class and $\boldsymbol{h}(\boldsymbol{x}) \in \mathbb{R}^k$. Note that $\boldsymbol{r}(\boldsymbol{x})$ is a matrix of size kxm, with m the number of classes to predict and k the number of concepts and that this definition does not change the prediction function as defined in equation 2 In the remainder of this paper, we denote $\boldsymbol{r}(\boldsymbol{x})_{\hat{t},:}$ as the column of $\boldsymbol{r}(\boldsymbol{x})$ corresponding to the predicted target $\hat{t}$.

4 EXPERIMENTAL SETUP

Datasets We evaluate our improved adjusted approach introduced in section 3.2 by comparing it with the original approach proposed by Melis & Jaakkola [11] on two classification settings: (i) MNIST digit recognition, with 60.000 training data digits and 10.000 test data digits. Concepts are learned by means of the convolutional auto-encoder as described in section 3. Following [11], we apply standard mean and variance normalization to the digit pixel values. (ii) Propublica's COMPAS Recidivism Risk Score datasets, that consists of demographic features labelled with recidivism risk scores. This is a natural dataset, meaning input features have a natural meaning, therefore no concepts are learned but the raw input is used, that is $h(x) = \boldsymbol{x}$. Following [11] we rescale the ordinal 'number of priors' variable to the range [0, 1] and we remove outliers. Before evaluating our improved adjusted approach, we reproduce the approach proposed by Melis & Jaakkola [11] by using the code provided on [9] with some adjustments. Since we are solely interested in the effect of replacing the relevance coefficients from $\boldsymbol{\theta}(\boldsymbol{x})$ with $\boldsymbol{r}(\boldsymbol{x})$ we left the hyperparameters from the original approach by default as shown in table 1. Using these hyperparameters the same results are obtained as presented in [11] for the original approach.

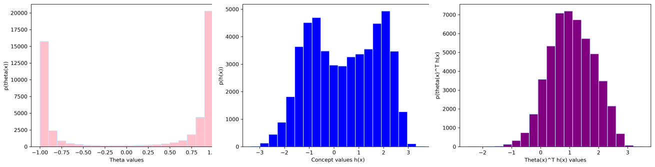
Hyperparameter	**MNIST**	**COMPAS**
Regularization on $\theta(x)$	grad3	no regularization
Type of conceptizer	CNN	Input
Parametrizer architecture	Simple	Simple
Number of concepts	5 & 22	11
Strength of regularization on $\theta(x)$	1e2	1e2
Sparsity parameter for learning $h(x)$	1e-4	Not applicable
Initial learning rate	0.001	0.001

Table 1: Hyperparameters used for training. In the study [11] the hyperparameters are explained.

We compare the two approaches by evaluating according to the three desiderata explained in section 2.2. Note that predictions are done in the same manner for both approaches and therefore we do not evaluate SENNs prediction accuracy. We measure each of the desiderata as follows:

- Intelligibility: as a quantitative measure for explicitness does not exist in the existing literature, we evaluate intelligibility in a qualitative manner by comparing the relevance scores together with the corresponding concepts of the two described approaches for several samples.
- Faithfulness: given a sample $\boldsymbol{x}$ with target t, we compute the drop in probability q_i of predicting the correct class when omitting a concept i, by setting its value $h_i(x)$ to zero in the prediction formula $\boldsymbol{f}(\boldsymbol{x}) = \boldsymbol{\theta}(\boldsymbol{x})^T\boldsymbol{h}(\boldsymbol{x})$, that is:
$$q_i^x = (\boldsymbol{\theta}(\boldsymbol{x})\boldsymbol{h}(\boldsymbol{x})_{\hat{t}} - (\boldsymbol{\theta}(\boldsymbol{x})\boldsymbol{h}(\boldsymbol{x})|_{\boldsymbol{h}_i(\boldsymbol{x})=0})_{\hat{t}} \tag{4}$$
Next, we compute the Pearsons correlation between the relevance score for a certain concept, where the relevance scores are given by $\boldsymbol{\theta}(\boldsymbol{x})_{\hat{t},:}$ in [11] and $\boldsymbol{r}(\boldsymbol{x})$ as defined in equation 3 for our approach and the corresponding drop in probability q_i when removing this concept. The intuition is that removing more important concepts should reduce the confidence of the classifier more than removing less important ones. In order to measure faithfulness, which we denote as ϕ, we calculate the average correlation over all samples in our test set T.
$$\phi = \frac{1}{|T| * k} \sum_{x \in T} \sum_{i=1}^{k} -p(\boldsymbol{\theta}(\boldsymbol{x})_{\hat{t},i}, q_i^x) \text{ for [11]} \tag{5}$$
$$\phi = \frac{1}{|T| * k} \sum_{x \in T} \sum_{i=1}^{k} -p(\boldsymbol{r}(\boldsymbol{x})_{\hat{t},i}, q_i^x) \tag{6}$$
for our alternative definition of relevance scores (7)
with k the number of concepts and $p() \in [-1, 1]$ Pearson's coefficient. A high ϕ is a indication for a faithful model.
- Stability: for each sample in the test set T we calculate the stability v as the average L2-norm between the relevance coefficients $\boldsymbol{\theta}(\boldsymbol{x})_{\hat{t},:}$ and $\boldsymbol{\theta}(\boldsymbol{x'})_{\hat{t},:}$, where $\boldsymbol{x'}$ is a sample in the neighbourhood of $\boldsymbol{x}$. For continuous data, e.g. pixel values, $\boldsymbol{x'} = \boldsymbol{x} + N(\boldsymbol{0}, \epsilon)$ with ϵ the level of noise added, whereas we for binary/ categorical data $\boldsymbol{x}$ is obtained by 'flipping' one feature randomly, for example by changing the gender type from male to female or vice versa.
$$v = \frac{1}{|T|} \sum_{\boldsymbol{x} \in T} ||\boldsymbol{\theta}(\boldsymbol{x})_{\hat{t},:} - \boldsymbol{\theta}(\boldsymbol{x'})_{\hat{t},:}||_2 \tag{8}$$
For our alternative definition of relevance coefficients, we calculate the stability by replacing $\boldsymbol{\theta}(\boldsymbol{x})_{\hat{t},:}$ in equation 8 with $\boldsymbol{r}(\boldsymbol{x})_{\hat{t},:}$.

5 RESULTS AND ANALYSIS

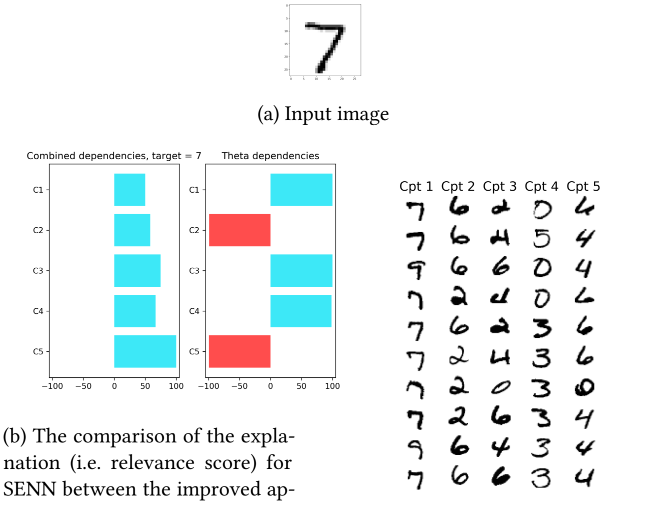


(a) Distribution for $\boldsymbol{\theta}(\boldsymbol{x})$ (b) Distribution for $\boldsymbol{h}(\boldsymbol{x})$ (c) Distribution for $\boldsymbol{r}(\boldsymbol{x})$

Figure 1: Histograms for MNIST, trained on 5 concepts with default parameter settings.

In this section we give an overview of our results. These will partially show our results in reproducing the research of [11]. These results will also be used to highlight some of the problems in using $\boldsymbol{\theta}(x)$ for interpreting. We show results of similar experiments for our proposed approach, which uses $\boldsymbol{r}(x)$ for interpretation. As outlined in the previous section, our results comprise numerical metrics and graphs for both faithfulness and stability, and a more subjective discussion of the intelligibility, based on graphs. In line with [11], we trained SENN on five concepts for MNIST. We evaluated it for different numbers of concepts, but did not see a notifiable change in prediction accuracy.

In figure 1 the distributions of $\boldsymbol{\theta}(x)$, $\boldsymbol{h}(x)$, $\boldsymbol{r}(x)$ are visualised for a SENN trained on MNIST where the number of concepts k equals 5. We see that the majority of the concept values is negative and a considerable amount is close to zero, providing a strong identification for taking the sign and magnitude of the concept values into account when evaluating the concept relevances. Comparing figure 1a to figure 1c, we see that our approach leads to a less skewed distribution of relevance scores.

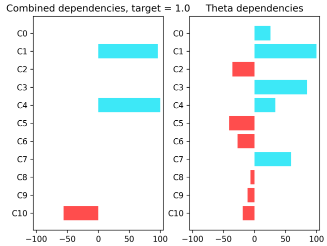


(a) Input image

(b) The comparison of the explanation (i.e. relevance score) for SENN between the improved approach using $\boldsymbol{r}(\boldsymbol{x})$) as relevance score (**left**) and the approach [11] using $\boldsymbol{\theta}(\boldsymbol{x})$ as relevance score (**right**).

(c) Plot of the most defining prototypes of the 5 concepts.

Figure 2: Explicitness using 5 concepts for the MNIST dataset.



Concept	Input
C0: Number of Priors	0.0000
C1: Score Factor	0.3244
C2: Age > 45	0.0000
C3: Age < 25	0.0000
C4: African American	1.0000
C5: Asian	0.0000
C6: Hispanic	0.0000
C7: Native American	0.0000
C8: Other	0.0000
C9: Female	0.0000
C10: Misdemeanor	1.0000

Figure 3: Left: The comparison of the explanation (i.e. relevance score) for SENN between our own approach using $\boldsymbol{r}(\boldsymbol{x})$ as relevance score (**left**) and the approach [11] using $\boldsymbol{\theta}(\boldsymbol{x})$ as relevance score (**right**), on the COMPAS dataset with natural concepts. **Right** : Input.

Explicitness

For interpretation we show the relevance score of each concept as shown in figures 2 and 3. In addition, for the MNIST dataset we visualised the relevance scores together with its input value and a visualisation of each concept as shown in figures 2a and 2c. Both approaches use the same form of explainability but show a different explanation for the same input.

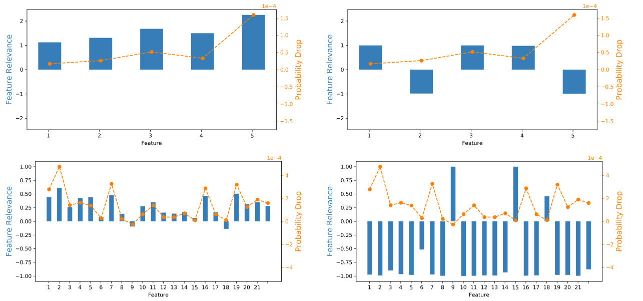


Figure 4: Probability drop (orange) overlaid on graph of relevance scores (blue bars) for the MNIST dataset. **Upper Left**: our approach using $\boldsymbol{r}(\boldsymbol{x})$ as relevance score, using 5 concepts. **Upper Right**: original approach [11] using $\boldsymbol{\theta}(\boldsymbol{x})$ as relevance score, with 5 concepts. **Bottom Left**: our approach using $\boldsymbol{r}(\boldsymbol{x})$ as relevance score, using 22 concepts. **Lower Right**: original approach [11] using $\boldsymbol{\theta}(\boldsymbol{x})$ as relevance score, using 22 concepts.

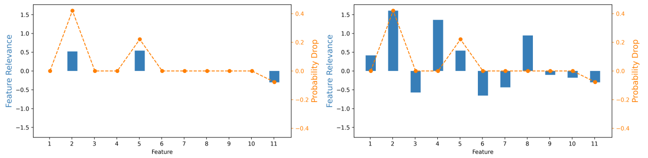


Figure 5: Probability drop (orange) overlaid on graph of relevance scores (blue bars) for the COMPAS dataset. **Left**: our approach using $\boldsymbol{r}(\boldsymbol{x})$ as relevance score. **Right**: original approach [11] using $\boldsymbol{\theta}(\boldsymbol{x})$ as relevance score

Faithfulness

In figures 4 and 5 the probability drop and the relevance coefficients are visualised for both approaches. In figures 4 and 5 the probability drop in comparison with the relevance score for every concept are plotted for both datasets. The correlations between the relevance score and the probability drop, denoted as ϕ, for both datasets are shown in figure 6.

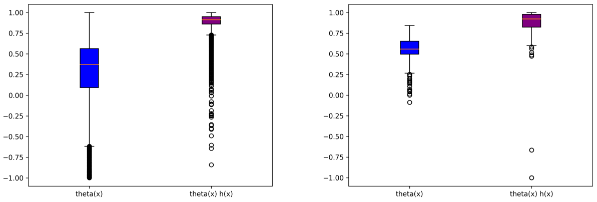


Figure 6: Box plot of faithfulness score ϕ (correlation between relevance score and probability drop). **Left**: for MNIST dataset with 5 concepts. **Right**: for COMPAS dataset

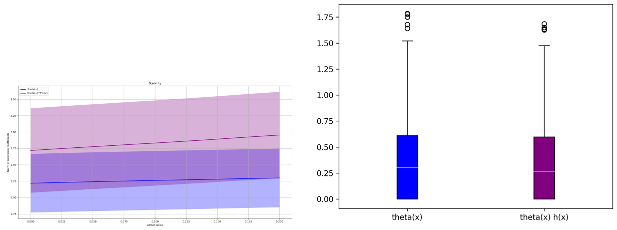


Figure 7: Stability υ. **Left**: MNIST: the stability υ and standard deviations for different levels of added noise. **Right**: COMPAS, box plot for the stability υ between the relevance scores of $\boldsymbol{x}$ and $\boldsymbol{x'}$, which is created by randomly changing the value of one of $\boldsymbol{x}$ features.

Stability

In figure 7 the stability υ of both methods for both datasets is shown. The stability υ of the relevance coefficients is notably lower for the approach using $\boldsymbol{\theta}(\boldsymbol{x})$ as relevance coefficient than of our approach (using $\boldsymbol{r}(\boldsymbol{x})$).

6 DISCUSSION

Evaluating our proposed adjustment to the approach of Melis & Jaakkola [11] by comparing it with the original approach according to its explicitness, faithfulness and stability, results in the following points of discussion:

i) Since there is no quantitative measure for explicitness, it is difficult to draw conclusions. Both explanations are of the same degree of comprehensibility and show an intuitive measure of interpretability of the influence of the concepts on the prediction. However, it is clear that our adjustment returns relatively different explanations. Since the relevance score of our adjusted approach more accurately describes the actual influence on the prediction of the input, our adjusted approach possibly gives a more intuitive perspective on intelligibility as it also considers the influence of the value of $\boldsymbol{h}(x)$.

ii) For both datasets our adjusted approach is more faithful, i.e., that relevance scores are in fact relevant for establishing the prediction. Intuitively, in Figures 4 and 5 we observe that the probability and the relevance score of our adjusted approach are more correlated, which can be confirmed with figure 6 showing the average correlations between the relevance score and the probability drop. By contrast, the original approach shows no correlation between the variables meaning that the features on the right of figures 2b and 3 are not truly relevant.

iii) Our adjusted approach results in less stable relevance coefficients than that of the original approach. This was to be expected, as SENNs do not impose a restriction on the local variation of the $\boldsymbol{h}(\boldsymbol{x})$ values and we use $\boldsymbol{h}(\boldsymbol{x})$ in our relevance coefficients. In order to solve this defect of our approach, we suggest to extend SENNs with an additional local-Lipschitz stability property for the concepts $\boldsymbol{h}(\boldsymbol{x})$, in line with the suggestion done in [18].

7 BROADER IMPLICATIONS

Given the observed improved faithfulness and intelligibility of machine learning models presented in this paper, we now consider to what extent these findings have repercussions in related research domains, such as fairness and accountability of decision-making models.

A reoccurring problem associated with disparate impact for historically disadvantaged subpopulations by machine learning-guided decision-making tools is a lack of understanding of both the causal mechanisms that govern the relationships between features and target variables, and the myriad ways in which input variables interact with each other and the outcome variable [2]. [6] propose methods of avoiding algorithmic discrimination by using causal inference to model counterfactual scenarios, from which conclusions can be drawn about the effects of group membership on prediction outcomes. However, this theoretically promising approach is often infeasible in practice, due to difficulties with modelling complex relationships between large numbers of features [5]. Our proposed modification of the methods presented by [11] could alleviate this problem with greater efficacy. Namely, using the observed probability drops associated with various (sensitive) feature omissions, policy-makers can reliably identify cases of protected attributes, e.g., race, having an unjustifiably large impact on prediction outcome, e.g., COMPAS recidivism-score. Reliable flagging of such incidences can avoid deployment of biased prediction tools, thereby decreasing the odds of unfair outcome. This approach differs from causal modelling, as it does not require prior assumptions about the relationships between variables, and simply uses observed correlations between outcomes and explanations.

Similarly, faithful interpretability can also help to streamline the process of recognising unwanted algorithmic biases and rightfully assessing how these were built into the model in question. This process of model scrutiny is commonly referred to as algorithmic accountability [17]. As defined by [3], accountability entails a relationship between an actor and a forum, in which the actor must fulfil an obligation to justify its design process and conduct, on which the forum may pass judgement. A natural consequence of more intelligible and faithful interpretations of how the feature space relates to algorithmically guided decisions, is the possibility to pass judgement with more precision and confidence, simultaneously avoiding misplaced judgement or scrutiny.

8 CONCLUSION

In this study we revisit the interpretability of Self-Explaining Neural Networks (SENN) proposed by Melis & Jaakkola (2018) [11] and find that the relevance score $\boldsymbol{\theta}(\boldsymbol{x})$ is lacking as an interpretability measure since it does not consider the sign and magnitude of the concepts values $\boldsymbol{h}(\boldsymbol{x})$, resulting in highly unfaithful explanations. We propose the use of an alternative relevance score for interpretability, namely $\boldsymbol{r}(\boldsymbol{x})$. Evaluating the adjusted approaches in comparison with the original approach on the explicitness, faithfulness and stability we conclude i) that our approach improves on the explicitness measure, ii) that our approach is more faithful and iii) that our approach results in less stable relevance coefficients. To solve the third point we suggest to extend SENNs with an additional local-Lipschitz stability property for the concepts $\boldsymbol{h}(\boldsymbol{x})$.

REFERENCES

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

[2] Solon Barocas and Andrew D Selbst. 2016. Big data's disparate impact. *Calif. L. Rev.* 104 (2016), 671.

[3] Mark Bovens. 2007. Analysing and assessing accountability: A conceptual framework 1. *European law journal* 13, 4 (2007), 447–468.

[4] Varun Gulshan, Lily Peng, Marc Coram, Martin C Stumpe, Derek Wu, Arunachalam Narayanaswamy, Subhashini Venugopalan, Kasumi Widner, Tom Madams, Jorge Cuadros, et al. 2016. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *Jama* 316, 22 (2016), 2402–2410.

[5] Michael Kim, Omer Reingold, and Guy Rothblum. 2018. Fairness through computationally-bounded awareness. In *Advances in Neural Information Processing Systems*. 4842–4852.

[6] Matt J Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. 2017. Counterfactual fairness. In *Advances in Neural Information Processing Systems*. 4066–4076.

[7] Oscar Li, Hao Liu, Chaofan Chen, and Cynthia Rudin. 2018. Deep learning for case-based reasoning through prototypes: A neural network that explains its predictions. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

[8] Scott M Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. In *Advances in neural information processing systems*. 4765–4774.

[9] David Alvarez Melis and Tommi Jaakkola. [n.d.]. Self-Explaining Neural Networks. https://github.com/dmelis/SENN.

[10] David Alvarez Melis and Tommi Jaakkola. 2018. On the Robustness of Interpretability Methods. *CoRR* abs/1806.08049 (2018).

[11] David Alvarez Melis and Tommi Jaakkola. 2018. Towards robust interpretability with self-explaining neural networks. In *Advances in Neural Information Processing Systems*. 7775–7784.

[12] Grégoire Montavon, Wojciech Samek, and Klaus-Robert Müller. 2017. Methods for Interpreting and Understanding Deep Neural Networks. *CoRR* abs/1706.07979 (2017).

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

[14] Cynthia Rudin. 2019. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence* 1, 5 (2019), 206–215.

[15] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. 2013. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034* (2013).

[16] Nava Tintarev and Judith Masthoff. 2011. Designing and evaluating explanations for recommender systems. In *Recommender systems handbook*. Springer, 479–510.

[17] Maranke Wieringa. 2020. What to account for when accounting for algorithms: a systematic literature review on algorithmic accountability. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. 1–18.

[18] Haizhong Zheng, Earlence Fernandes, and Atul Prakash. 2019. Analyzing the Interpretability Robustness of Self-Explaining Models. *CoRR* abs/1905.12429 (2019).

A CONTRIBUTION

The authors equally contributed to this study. The listing order is random. Joosje and Hannah reproduced the approach of Melis & Jaakkola for the MNIST dataset. Lennert and Daniël reproduced the approach of Melis & Jaakkola for the COMPAS dataset. Daniël implemented the faithfulness measure used in this study for both approaches. Joosje implemented the stability measure used in this study for both approaches on both datasets (i.e. MNIST and COMPAS dataset).