Evaluating Sensitivity Consistency of Explanations

Hanxiao Tan
AI Group, TU Dortmund
hanxiao.tan@tu-dortmund.de

Abstract

While the performance of deep neural networks is rapidly developing, their reliability is increasingly receiving more attention. Explainability methods are one of the most relevant tools to enhance reliability, mainly by highlighting important input features for the explanation purpose. Although numerous explainability methods have been proposed, their assessment remains challenging due to the absence of ground truth. Several existing studies propose evaluation methods from a certain aspect, e.g., fidelity, robustness, etc. However, they typically address only one property of explanations, and thus more assessing perspectives contribute to a better explanation evaluating system. This work proposes an evaluation method from a novel perspective called sensitivity consistency, where the intuition behind is that features and parameters that strongly impact the predictions and explanations should be highly consistent and vise versa. Extensive experiments on different datasets and models evaluate popular explainability methods while providing qualitative and quantitative results. Our approach further complements the existing evaluation systems and aims to facilitate the proposal of an acknowledged explanation evaluation methodology.

1. Introduction

Deep Neural Networks (DNNs) are widely applied in a wide range of fields, such as computer vision [11], robotics [19], etc., due to the excellent fitting ability and outstanding predictive performance. However, agnosticity is one of the most threatening concern for DNNs with complex architectures. Due to the uninterpretability of the predictions, applications in areas where human life is at stake, such as autonomous driving [10] or healthcare [8], are severely restricted. To alleviate this concern, explainable AI researches [7, 40] have been proposed and have increasingly become one of the popular topics in recent years. There are two major directions in explainability researches, which are A) predicting through refined data and interpretable (linear) models [29] and B) designing post-hoc explainability methods to explain the predictions [31]. The interpretable models are straightforward and intuitive, but emulating DNNs with linear models under complex tasks struggles to reproduce their performances and the data for these tasks (e.g., high-resolution images) can barely be refined accurately. Therefore, the application scenarios of interpretable models are limited.

Post-hoc explainability methods are mainly split into two broad categories: black and white-box approaches. The former does not access the internal structure of the model and infers feature attribution by observing the effect of input perturbations on the predictions [23, 26–28], whereas the latter leverages information within the model (e.g., gradients) to summarize the important portions of the inputs [5, 34, 35, 38]. However, almost all explainability methods suffer from plausibility flaws, including A) irrelevance of explanations to model parameters [1], B) the out-of-distribution perturbation issue [17], C) sensitivity to baseline choices [20], D) robustness deficiency [12] and E) lack of ground truth.

The most popular approach is feature removing [44], based on the idea that eliminating features with the highest attribution in the explanations would severely disrupt predictions. Recently, assessment metrics from multiple perspectives have been proposed, such as sensitivity to model parameters [1], human comprehensibility [2,25], and generalizability [39], which complement the other necessary conditions for plausible explanations. Nevertheless, we believe that the properties of explanations have not been fully exploited. more evaluation perspectives may facilitate deeper understandings of explainability methods.

This work proposes a novel evaluation metric for explanations, called Sensitivity Consistency (SenC), based on the idea that predictions and explanations are expected to be sensitive to the identical input features. In addition, we extend this perspective to model parameters, assessing sensitivity consistency by observing whether those groups of neurons that play important roles in prediction would have similar impacts on explanations. Our contributions are mainly as follows:

- We propose a novel explanation evaluation metric

SenC that assesses whether an explanation is reliable by comparing the discrepancy between the sensitivity of predictions and explanations to input features or neurons. SenC is a black-box approach that is applicable to all explainability methods.

- We quantitatively evaluate popular explainability methods on various datasets and models through extensive experiments, and verify the consistency of SenC with human intuition through a user study.

2. Related Work

In this section we present related work on explainability methods and their reliability studies.

Explainability Methods. The pioneer of explainability methods was proposed by [34], which simply observes the attribution of individual pixels by the gradient of the output neuron to the input. However, follow-up studies identified that the gradient failed to faithfully capture attributions and proposed corresponding improvements, including masking out the negative channels in forward or backward propagation [36], adding Gaussian noise to the surrounding pixels to clarify and smooth out the explanations [35], integrating the gradient of all interpolations starting from an uninformative baseline to the input [38] (as well as its approximated version [33]). Layer-wise Relevance Propagation (LRP) is another series of gradient-based attribution that starts with the output and back-propagates the contribution of each neuron to the input layer [5]. GradCAM generates attributions with a global average pooling layer that weights and maps the convolutional output to the input layer [32]. Another type of explainability approaches require no access to the gradient, instead considering the model as a black box and observing the effect of input variations on the outputs, named perturbation-based methods. LIME achieves interpretability by training a linear surrogate models with perturbations adjacent to the inputs [27]. KernelSHAP [23] efficiently approximates the Shapley value [13] through weighted perturbations and linear surrogates, which greatly reduces computational intensity. Recently, RISE [26] was proposed, which generates input attributions by randomly masking out features and weighting the masks according to the outputs.

Evaluation metrics for explanations. Due to the absence of ground truth, there is no acknowledged metrics. Sensitivity is one of the most intuitive metrics that assesses the fidelity of explanations by comparing the difference in confidence between the predictions after removing the most attributed feature in the explanation from the input (or insert it in an uninformative baseline) and the original predictions (baselines) [3–5, 9, 20, 30]. [17] argued that hard removals may disrupt the data distribution, resulting that the model is incapable of predicting effectively for data that has never been seen before. They propose RemOve And Retrain to mitigate the OOD problem by retraining after removing features. However, this in turn raises concerns about explanation fidelity to the model. Explanation robustness is another assessment perspective, where [3] argues that explanations given to similar inputs should also share a high degree of similarity. Another perspective that drew attention was the sensitivity to model parameters, as [1] found that the quality of the explanations from part of the methods is not seriously impaired when the model parameters are highly randomized. Besides, there are approaches that are not widely employed, such as Pointing Game [42], Generalizability [39], semantic-level perturbations and synthetic ground truth [16]. User assessment [41,43,45] is a convincing alternative, which is nonetheless costly and lacks reproducibility due to human subjectivity. In addition, a latest research [15] integrates explainability evaluation toolkits to facilitate accessment.

3. Methods

3.1. Sensitivity Consistency (SenC)

Existing studies indicate that different explainability methods may provide different explanations for identical models and inputs [1,20]. In addition, recent research argue that there is a “Rashomon” of explanations, whereby explanations that reasonably demonstrate prediction attributions may not be unique [22,24]. Due to the lack of ground truth, it is challenging to authenticate the credibility of inconsistent explanations. However, by considering inputs and explanations as a black-box system, we can assess the plausibility of explanations by observing the relationship between their variations. We rely on the argument that prediction confidences and explanations are expected to be sensitive to identical prediction bases, which is termed sensitivity consistency. The prediction bases are attributed to two factors, the input features and the model parameters, which are two aspects of the proposed assessment method. An overview of SenC is presented in Fig 1.

3.2. Data Sensitivity Consistency

Data sensitivity consistency refers to the proximity of the degree to which the prediction confidence and explanation are impacted when part of the features in the input data vary. Elaborately, if a modification (removal or perturbation) of a feature significantly interferes with the prediction, while no serious explanation corruption occurs, the explanation is considered to lack sensitive consistency, and vice versa. To statistically measure the proximity of the impacts, we leverage a method based on random mask perturbations, which is inspired by [26]. We segment the input image into several partitions as input features with image segmentation algorithms to avoid the overwhelming computational intensity for processing pixel-wise features. Subsequently, we gen-

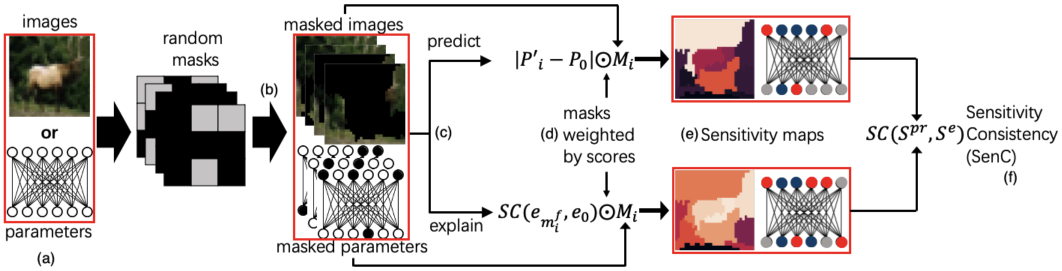

Figure 1. An overview of SenC. SenC contains the following main components:(a) Selection of the input features or parameters to be perturbed. (b) Perturbation of the selected target by randomly generating an extensive number of masks. (c) Re-prediction and re-explanation with perturbed conponents (inputs or parameters) respectively. (d) Score each mask based on the difference in predictions and similarity in explanations. (e) Summing the product of all masks with their scores yields the prediction and explanation sensitivity maps, respectively. (f)SenC is derived by comparing the correlation of the two sensitive maps. Note that the red box in the figure indicates that SenC applies to either the input or the parameter, rather than in parallel with both.

erate a massive amount of masks to randomly eliminate a fraction of the features. We predict the original and masked inputs to yield confidence scores P^o and P', respectively, and denote their differences as $\Delta P = |P^o - P'|$. With an enormous number of masks weighted by ΔP and summed, the final prediction sensitivity of the k^{th} feature is formulated as:

$$S^{pr}_{f_k} = \sum_{i=1}^{n}(1 - \Delta P_i) \odot M^f_i \qquad (1)$$

where M^f_i and ΔP_i denote the i^{th} feature mask and its confidence discrepancy, respectively. The sensitivity of all m input features to the prediction can be summarized as $S^{pr}_f = \left\{S^{pr}_{f_1}, ..., S^{pr}_{f_m}\right\}$.

Explanation sensitivity is the degree to which the input features impact the generated explanations. Similar to predictive sensitivity, we randomize an equivalent number of masks to eliminate input features and explain the masked inputs by a specific explainability method. We weight each mask by comparing the similarity of explanations from the original and masked inputs, and eventually derive the feature sensitivity to explanations from the weighted sum. The explanation sensitivity of input features is $S^e_f = \left\{S^e_{f_1}, ..., S^e_{f_m}\right\}$, where the k^{th} feature is represented as:

$$S^e_{f_k} = \sum_{i=1}^{n} SC(e^o, e_{m^f_i}) \odot M^f_i \qquad (2)$$

where e^o, $e_{m^f_i}$ denote the explanations from the original and masked inputs, respectively. SC denotes Spearman's correlation coefficient, which is formulated as:

$$\rho(e, e') = \frac{COV(R(e), R(e'))}{\sigma_{R(e)}\sigma_{R(e')}} \qquad (3)$$

where $R(*)$, COV and ρ are the rank function, the covariance and the standard deviation, respectively.

Finally, by comparing the proximity of the prediction and explanation sensitivities, we derive the feature sensitivity consistency:

$$\rho^f = SC(S^e_f, S^{pr}_f) \qquad (4)$$

The detailed algorithm for data SenC is shown in Algorithm 1. ρ^f is essentially a Spearman's correlation coefficient, hence its domain of values is $[-1, 1]$. However, our experiments are statistical and the probability of presenting opposite correlations can be ignored. Therefore, the statistical mean of SenC has a value domain of [0,1], where 0 represents the absence of sensitivity consistency and 1 represents absolute consistency.

Algorithm 1: Data Sensitivity Consistency (SenC) for a given data

```
Input : An input data $x$, a well-trained model $F(\cdot)$,
        an explainability method $H(F, x)$ for the
        model $F$ and the number of masks $K$
Output: Data SenC $\rho^f_x$ of $H$ for input $x$
$P^o_x = F(x)$ # Original prediction
$e^o_x = H(F(x), x)$ # Original explanation
$S^{pr}_{f_x}, S^e_{f_x} = zeros\_like(x)$ # Initialization
for $k = 1$ to $K$ do: #Generating masks
    $M^f_k = random\_like(x)$
    $x'_k = x \odot M^f_k$ #The $k^{th}$ perturbation
    $S^{pr_k}_{f_x}$ += $(1 - |P^o_x - F(x'_k)|) \odot M^f_k$ #Scored
      by prediction variation
    $S^{e_k}_{f_x}$ += $SC(e^o_x, H(F, x'_k) \odot M^f_k$ #Scored by
      explanation similarity
end for
$\rho^f_x = \sum_{k=1}^{K} SC(S^{e_k}_{f_x}, S^{pr_k}_{f_x})$ # Sum all scores
```

For input features, we generally pay more attention on those relevant components, such as the set of pixels containing objects. Therefore, beside the overall consistency, we evaluate two additional metrics, which are Top-1 and Top-3 agreement. Top-K indicates the percent of overlap for the K features that are most sensitive to predictions and explanations, which is formulated as:

$$TA_k = \frac{\left|\left\{Top_k(S_f^e) \cap Top_k(S_f^{pr})\right\}\right|}{\left|\left\{Top_k(S_f^e)\right\}\right|} \tag{5}$$

where $Top_k(S)$ represents the set of the first k elements from sorted S. The Top-K metric mitigates the interference caused by background pixels to some extent, thereby concentrating more on the sensitivity assessment of the target objects.

3.3. Parameter Sensitivity Consistency

Apart from features, parameters are expected to be consistently sensitive as well, which ensures that predictions and explanations made by the model on a given input are mainly attributed to the same set of neurons. In contrast to images (typically $C \times W \times H$), parameters are higher dimensional, which encompass diverse architectural units and rendering the perturbation more challenging. To avoid explosive computational intensity, we draw the following compromises without sacrificing too much performance:

- We compute the parameter similarity for each layer individually and derive the global similarity by averaging.

- As the majority of the parameters belong to the weights, we only evaluate the parameters on the weights of the feature extraction layers (convolutional and fully-connected layers), ignoring the biases.

- The quantity of parameters on the weights is far greater than the image pixels, thus requiring a remarkably larger number of masks to accurately assess the sensitivity, which causes enormous time and computation costs. To enable the experiment being feasible, we group the parameters according to the type of the layer to diminish the amount of masks required for perturbation. For a convolutional layer with size $D_{in} \times D_{out} \times C_w \times C_h$, we treat an output channel as a perturbation unit with the size of $D_{in} \times C_w \times C_h$. The reason for not considering an input channel as a perturbation unit is to avoid the possibility that only one perturbable term exists when processing uni-channel data (e.g. MNIST). For a fully connected layer $D_{in} \times D_{out}$, we select an input channel as a perturbation unit with dimension D_{out}. The argument against treating the output channel as unit is to prevent masking the label channel when perturbing the last layer and thereby losing gradients.

The process of calculating parameter sensitivity is analogous to that of feature sensitivity. We randomly generate masks on parameters groups, which partially eliminate the original weights and turn into new models. Comparing the differences in prediction confidence and explanation similarity between the original and new models for the identical inputs allows for the calculation of sensitivity consistency for each parameter group. The parameter sensitivity consistency is formulated as:

$$\rho^{pa} = SC(S_{pa}^e, S_{pa}^{pr}) \tag{6}$$

where $S_{pa}^{pr} = \left\{S_{pa_1}^{pr}, ..., S_{pa_2}^{pr}\right\}$ denotes the parameter sensitivity for the prediction, in which pa_k is the k^{th} parameter group in a certain layer, and $S_{pa_k}^{pr}$ is calculated by the prediction differences, also formulated as $S_{pa_k}^{pr} = \sum_{i=1}^{n}(1 - \Delta P_i) \odot M_i^{pa}$ (M_i^{pa} is the i^{th} mask on the parameter groups). S_{pa}^e is the parameter sensitivity of the explanation, obtained by $S_{pa}^e = \sum_{i=1}^{n} SC(e^o, e_{m_i^{pa}}) \odot M_i^{pa}$, where $e_{m_i^{pa}}$ is the explanation generated by the model under mask M_i^{pa}. The details of parameter SenC is demostrated in Algorithm 2. Note that though we need to access the model parameters when evaluating parameter SenC, we are still interested in the correlation between input and output, and thus SenC is still considered to be a black box in a broad sense.

Intuitively, both predictions and explanations are supposed to follow variations in the same set of features and parameters. Therefore, explanations with higher sensitivity consistency are considered more plausible. Lastly, due to the lack of reference, we randomly generate a nonsensical explanation for each data and equally perform the sensitivity consistency evaluation on it as the baseline.

4. Experiments

In this section we demonstrate the experimental results. Our experiments are conducted on three datasets with different complexities, which are the MNIST handwritten dataset, CIFAR10 and a real-world dataset called *German Traffic Sign Recognition Benchmark* (GTSRB) [37], respectively. For better prediction performance, we train models with different structures on each of the three datasets. For MNIST, we train a simple four-layer neural network, noted as ModelCNN, whose structure can be simply summarized as $Conv1 \rightarrow MP1 \rightarrow Conv2 \rightarrow MP2 \rightarrow FC1 \rightarrow FC2$, where $Conv$, MP and FC denote convolutional, max-pooling and fully connected layers, respectively. ModelCNN achieves 98.5% accuracy on the MNIST test set. For CIFAR10 and GTSRB, we train a ResNet18 [14] and a MobileNetV3 [18] as the classifiers, respectively, which achieve 93.2% and 97.7% accuracy on the test set, respectively. Finally, we conduct a user study on ImageNet to verify whether the evaluation of SenC is consistent with human cognition.

Algorithm 2: Parameter Sensitivity Consistency (SenC)

```
Input  : An input data $x$, a well-trained model $F(\cdot)$
         with layer-wise parameters $\{w_1, ..., w_n\}$,
         an explainability method $H(F, x)$ and the
         number of masks $K$
Output: Parameter SenC $\rho^{pa}$ of $H$
$P_x^o = F(x)$ #
$e_x^o = H(F(x), x)$ #
for $m = 1$ to $n$ do: #Layer-wise process
    $S_{w_m}^{pr}, S_{w_m}^{e} = zeros\_like(w_m)$ # Initialization
      with the shape of corresponding parameters
    for $k = 1$ to $K$ do:
        $M_k^{w_m} = random\_like(w_m)$
        $w_m'^k = w_m \odot M_k^{w_m}$
        $F'_{w_m'^k} = \{w_1, ..., w_m'^k, ..., w_n\}$ #$k^{th}$
          perturbation on $m^{th} layer$
        $S_{pa_m}^{pr_k} \mathrel{+}= 1 - \left| P_x^o - F'_{w_m'^k}(x) \right| \odot M_k^{w_m}$
        $S_{pa_m}^{e_k} \mathrel{+}= SC(e_x^o, H(F'_{w_m'^k}, x) \odot M_k^{w_m}$
    end for
    $\rho_m^{pa} = \sum_{k=1}^{K} SC(S_{pa_m}^{e_k}, S_{pa_m}^{pr_k})$
$\rho^{pa} = \sum_{m=1}^{n} \frac{\rho_m^{pa}}{n}$ #Global parameter SenC
```

We select the following explainability approaches as candidates to be evaluated: Vanilla Back-propagation (VB) [34], Guided Back-propagation (GB) [36], Integrated Gradients (IG) [38], Layer-wise Relevance Propagation (LRP) [5], GradCAM [32] and DeepLift [33]. Since perturbation-based evaluation consumes a significant amount of time to generate and process masks, while surrogate models methods such as LIME [27] also demand extensive perturbation samples for their explanations, this category will not be evaluated considering the time and computational costs. In the experiments, all explainability methods are implemented based on Captum toolkit [21].

The experimental configurations are as follows: When evaluating the data SenC, we randomly select 1000 instances from the test set, and for each instance the number of perturbation masks generated is 5000. When evaluating the parameter SenC, we select only 100 instances from the test set but generate 10000 random masks for each layer of the model to be evaluated, as the number of perturbable channels of the parameter is significantly larger than the number of hyperpixels in the image. In segmenting the image, we utilize *slic* in the *scikit-image* package to roughly split every 50 pixels into one superpixel, which maintains the independence of local features without excessively raising the computational intensity. We chose 0.8 as the masking rate for the generated masks. The masking rate is a flexible hyperparameter with appropriate values will not significantly impact the evaluation results (detailed analysis is in Sec. S1.3).

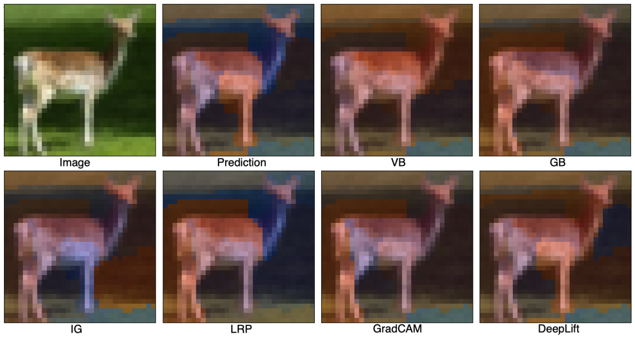

Figure 2. Visualization of data sensitivity from CIFAR-10 dataset. Areas rendered in red represent high sensitivity while those in blue indicate low sensitivity.

4.1. Sensitivity visualization

In this section we demonstrate two visualization examples for data and parameter sensitivity consistency, respectively.

Data sensitivity. We show the data sensitivity of a random image in CIFAR10 in Fig. 2. The sensitive areas of prediction can be seen to be centered on the front part of the body and head of the deer. However, for explainability methods, the sensitive fields are partially different: e.g. the neck of the deer is included in the sensitive regions by VB, as well as the upper part of the background is labeled as more sensitive for VB, IG, GradCAM and DeepLift. The closest match to the prediction sensitivity is LRP, so which therefore yields the highest sensitivity consistency for this instance.

Parameter sensitivity. For a simple illustration, we choose the first convolutional layer ($conv1$) of ModelCNN as the object to visualize the parameter sensitivities, which is shown in Fig. 3. This layer contains 16 output channels, each represented by a square in the figure, where red and blue indicate high and low sensitivity, respectively. It can be observed that for prediction, the first, third and sixteenth channels are most sensitive. For the explainability methods, all except Deeplift label the third channel as highly sensitive, while among them VB,GB and LRP exhibit high sensitivity on all three channels simultaneously, hence their sensitivity consistency for the particular input on this layer is relatively high.

4.2. Quantitative SenC evaluation

4.2.1 MNIST

MNIST is the simplest image dataset, which consists of 60,000 train and 10,000 test instances, each with the size

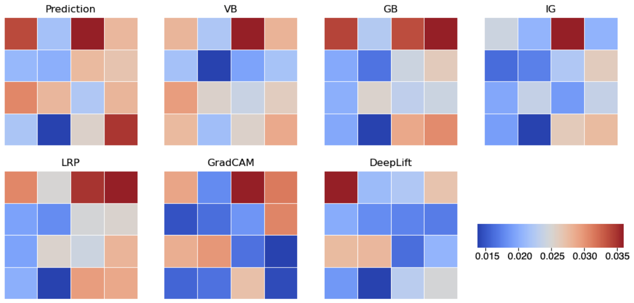

Figure 3. Visualization of parameter sensitivity. The layer being visualized is the first convolutional layer of the ModelCNN, which contains 16 output channels corresponding to the 16 squares in the figure. The redder the color of the squares, the higher the sensitivity and vice versa.

of 784 (28×28). For efficient evaluation, we restrict the number of hyperpixels to 15.

Data SenC. The results of the qualitative assessment of data SenC are demonstrated in (a) of Fig. 5. As a reference, we generate a random mask for each instance as a baseline explanation, whose average SenC is expected to be zero. The results illustrate that for data with low complexity, almost all explainability methods exhibit consistent sensitivities (mean SenC $\bar{\rho} > 0.6$), except for GradCAM, whose variance is slightly higher ($\sigma^2(\rho) = 0.13$).

Furthermore, we present the agreement of features with Top-1 and Top-3 sensitivities in (a) and (d) of Fig. S1, respectively. All explainability methods except GradCAM are capable of reaching a Top-1 agreement of around 80%. In the Top-3 metric, again with the exception of GradCAM, the probability that all three of the most sensitive features in the explanations generated by the rest of the methods are all in agreement is higher than 20%, while there are barely any cases where no intersection exists (percent $p < 5\%$). In general, for simple datasets like MNIST, almost all explainability methods achieve highly sensitivity consistency of input features for predictions and explanations.

Parameter SenC. (a) of Fig. 4 shows the average SenC of different explainability methods on all layers of the model (also including the randomized baseline explanation). In the results, GB and IG achieve higher levels of consistency ($\bar{\rho} = 0.29$ and 0.32, respecitively), whereas the consistency of GradCAM is relatively low ($\bar{\rho} = 0.14$) and unstable ($\sigma^2(\rho) = 0.014$). The remaining methods reveal an intermediate level of sensitivity consistency ($\bar{\rho} \in [0.2, 0.3]$). Moreover, we analyze each layer individually and present the results in Fig. S6. We note that the sensitivity consistency of convolutional layers is far superior to that of fully-connected layers, regardless of which explainability method is applied. We attribute the reason to two points: a) Convolutional layers are more intuitive as they directly extract features from the adjacent areas of images, whereas fully-connected layers are typically treated as latent features, which are highly abstract and may be activated by features in different regions. b) The channels of convolutional layers are more independent of each other compared to fully-connected layers, and thus are less impacted under individual perturbations. Therefore, we recommend choosing the top convolutional layer as the target for evaluating parameter consistency.

4.2.2 CIFAR-10

CIFAR10 is a small-size (32×32) image dataset consisting of 10 categories, which contains 50000 training and 10000 test data. When splitting the hyperpixels, again every 50 pixels are split into a group with a total number of 20 hyperpixels per image.

Data SenC. The data SenC is illustrated in (b) of Fig. 5. Compared to MNIST dataset with low complexity, the data SenC of CIFAR10 exhibits a significant collapse, especially for LRP, whose $\bar{\rho}$ plummets from 0.609 to -0.012, which implies that when explaining structurally complex data and models, LRP barely reveals consistency of sensitivity between predictions and explanations. Besides, GradCAM still suffers from low consistency $\bar{\rho} = 0.052$, which is almost on par with randomly generated explanations. In contrast, despite the substantial reduction, DeepLift maintains a relatively high consistency ($\bar{\rho} = 0.351$) and is therefore considered to be the more stable explainability method.

We again evaluate the Top-1 and Top-3 SenC and show the results in (b) and (e) of Fig. S1, respectively. Agreements for all explainability methods decline significantly on CIFAR-10, with the most dramatic drop being for LRP, whose Top-1 agreement falls from nearly 80% on MNIST to the same level as the randomized explanations. The performance of Top-3 agreement is analogous to that of Top-1, where VB, GB, IG and DeepLift outperform, with over 60% of their Top-3 sensitivity features sharing at least one agreement. DeepLift still maintains the best agreement with more than 40% probability that at least two of its Top-3 features overlap. As a conclusion, the complexity of CIFAR-10 is elevated compared to MNIST, resulting in a certain decrease in sensitivity consistency for all explainability methods, while DeepLift maintains the highest consistency.

Parameter SenC. Due to the complicated structure of ResNet, for clarity, we only demonstrate the consistency of parameter sensitivities of the first convolutional layer, the last fully-connected layer, and all the intermediate hidden layers belonging to “layer1”. The parameter sensitivity consistency of ResNet18 is demonstrated in (b) of Fig. 4. In comparison to MNIST, all explainability methods suffer from various degrees of reduction in consistency, most notably GB and LRP, which both exhibit a decrease in average parameter consistency of 0.27, and LRP, whose con-

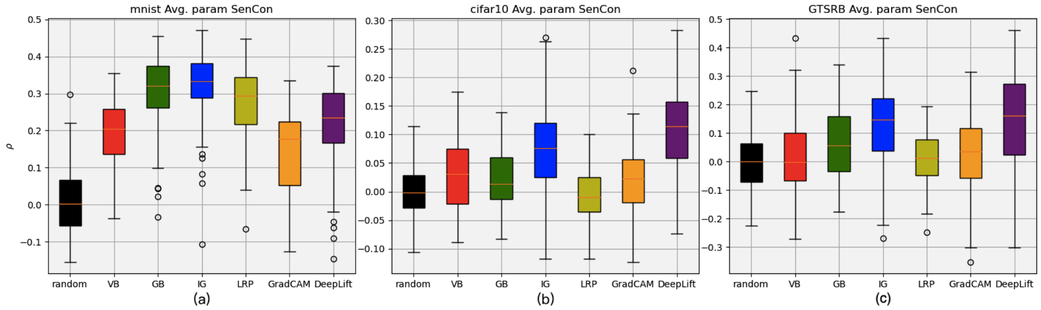

Figure 4. Average parameter sensitivity consistency over all selected layers. From left to right are the MNIST,CIFAR-10 and GTSRB datasets, respectively.

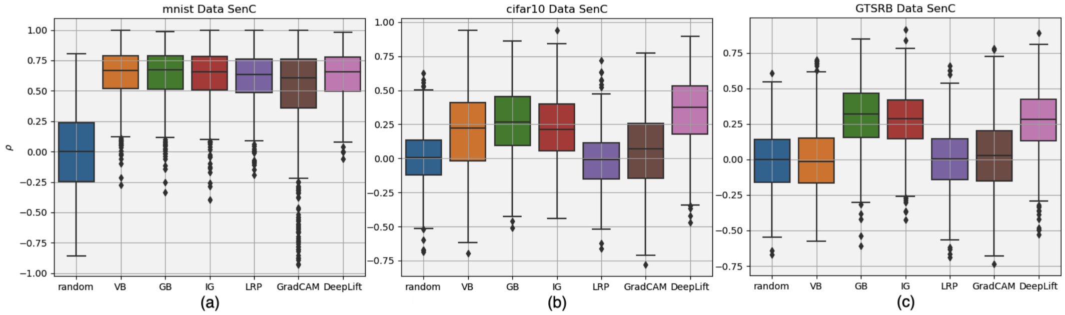

Figure 5. Evaluation of data sensitivity consistency. From top to bottom are the evaluation results on MNIST,CIFAR-10 and GTSRB datasets, respectively. The x-axis in all plots represents different explainability methods, the y-axis represents Spearman's correlation coefficient ρ. Higher ρ denotes more consistent sensitivity.

sistency is already lowered to a comparable level to that of the random explanation. The consistency of VB and Grad-CAM degrades 0.17 and 0.13, respectively, to a relatively insignificant degree, due to their unprominent performance on MNIST. Despite declining 0.25 and 0.1, respectively, IG and DeepLift remain in relatively high consistency, with the average of DeepLift remaining at a high level above 0.1. The layer-wise SenC analyses are presented in Fig. S7. The conclusions are aligned with those in MNIST, where convolutional layers are more consistent compared to fully-connected layers, and the deeper the layer the more difficult their sensitivities are to be consistent. Additionally, IG and DeepLift again outperform the parameter SenC, especially for the layers $conv1$ and $layer1.0.conv1$, which are remarkably higher than all other explainability methods.

4.2.3 GTSRB

To test the reliability of explainability methods on real-world datasets, we conduct experiments on the GTSRB dataset [37]. GTSRB is a dataset consisting of photographs of 43 different types of traffic signs, which includes 39209 and 12630 training and test data for learning and prediction, respectively.

Data SenC. The data SenC of GTSRB is shown in (c) of Fig. 5 and (c), (f) in Fig. S1. No significant fluctuations are observed for all explainability methods compared to CIFAR-10, except for a significantly decline in SenC for VB. The trends in Top-1 and Top-3 agreement evaluations are roughly equivalent, with all methods maintaining comparable levels except for VB, which rapidly collapses. In summary, for data SenC on GTSRB, GB, IG and DeepLift perform relatively better compared to other explainability methods.

Parameter SenC. The average and layer-wise parameter sensitivity consistency of GTSRB are exhibited in (c) of Fig. 4 and Fig. S8, respectively. The results for the average parameter SenC for GTSRB are comparable to those of CIFAR-10, except for a relatively noticeable gain in GB while IG and DeepLift still remain superior. For the layer-wise evaluation, again due to the computational intensity, we choose only the first 2 convolutional layers and the last 2 fully connected layers of the network. The final conclusion remains broadly uniform, that IG and DeepLift exhibit relatively better consistency in the first two layers, whereas the consistency of VB, LRP, and GradCAM fails to show an advantage over randomized explanations.

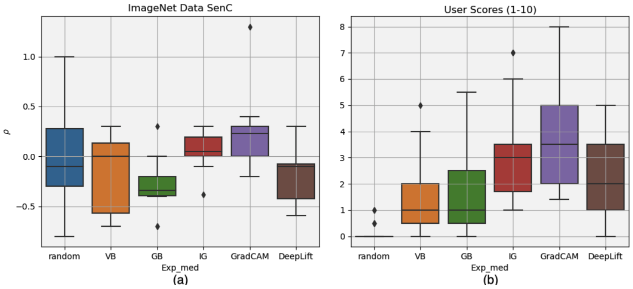

Figure 6. Quantitative evaluation and user study on ImageNet. (a) SenC evaluation results and (b) User scores. The Top-1 and Top-3 agreements can be seen in Fig. S2

4.3. Complex model (dataset) and user study

For practical purposes, we perform an evaluation of data SenC on the SOTA large model while conducting a light scale user study. We chose VIT [6] as the classifier, train it on ImageNet and achieves 80.7% accuracy in the testset. We randomly select 10 out of 1000 categories from ImageNet, first generate explanations with the selected explainability methods (Vanilla LRP is discarded due to incompatibility with VIT), and evaluate their data SenC. In parallel, we send the generated explanations to the users and let them score each explanation subjectively based on their experience and intuition. We invite 21 participants for the study by showing them the original images and the explanations generated by the explainability methods, and asking them to rate each explanation. For those users who do not have basic knowledge in XAI, we briefly introduce the idea of explainability methods and their functionality. Explanations are rated on a scale of 0 to 10, with 0 representing barely able to provide any information, to 10 where users believe they can fully understand the basis of model predictions.

We finally combine the evaluations of SenC with user feedbacks in Fig. 6 to observe whether they are correlated. The result demonstrates that while IG and GradCAM receive relatively high ratings from users, their SenC also wins in quantitative evaluations. On the contrary, VB, GB and DeepLift suffer from flawed SenC in this dataset, as well as they receive lower user scores. The result indicates that the SenC evaluation results are to some extent consistent with the intuitive perceptions of humans.

5. Limitations

This work presents a novel perspective for evaluating explainability methods. However, we acknowledge that a few deficiencies remain non-negligible, which fall broadly into the following three points:

- **OOD Perturbations.** Hard perturbation that may disrupt the distribution of data is one of the largest challenges for explainability studies. Though alternatives have been proposed in recent studies such as [17], they are not widely accepted and applied due to computational intensity and fidelity issues. In the assessment of SenC, perturbing data or parameters with randomly generated masks is considered as hard perturbation, which is one of the factors that may threaten the reliability of the evaluations.
- **Perturbation channels for FC layers.** FC layers exhibit much less consistency than convolutional layers in the evaluation, partially due to the issue of segmenting the channels of FC layers. On the one hand, the FC layers contain an enormous number of parameters, which if perturbed discretely would require an incalculable amount of masks, rendering them almost impossible. On the other hand, unlike convolutional layers with well separated channels, neurons in FC layers are densely connected, and hard masking any part may severely impair the remaining ones. Therefore, a proper splitting and perturbation approach is desired to balance the assessment accuracy and computational intensity.
- **Computation time costs.** Similar to other perturbation-based methods, SenC sacrifices efficiency for the black-box property. When evaluating structurally complex data or models, a relatively large quantity of hyperpixels or channels is required, which leads to an explosive demand for the number of masks. Detailed analysis can be found in Sec. S1.5. Reducing the amount of hyperpixels or channels effectively mitigates this issue, however will result in a compromise in evaluation precision, which is a tradeoff that also needs to be addressed.

6. Conclusion

This work proposes a novel perspective to evaluate explainability methods. By generating a large number of masks to perturb inputs or parameters to identify which components are sensitive to the perturbation and assess whether they are consistent. We conduct experiments with three different datasets and models, as well as a user study on a more complex dataset. The result reveals that the assessment of SenC is to some extent consistent with human intuition. For future work, analyzing and refining the factors that render sensitivities inconsistent is a promising direction for improving the fidelity of explainability methods.

Acknowledgements. This research has been funded by the Federal Ministry of Education and research of Germany and the state of North Rhine-Westphalia as part of the Lamarr Institute for Machine Learning and Artificial Intelligence.

References

[1] Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. Sanity checks for saliency maps. *Advances in neural information processing systems*, 31, 2018. 1, 2

[2] Ahmed Alqaraawi, Martin Schuessler, Philipp Weiß, Enrico Costanza, and Nadia Berthouze. Evaluating saliency map explanations for convolutional neural networks: a user study. In *Proceedings of the 25th international conference on intelligent user interfaces*, pages 275–285, 2020. 1

[3] David Alvarez-Melis and Tommi S Jaakkola. On the robustness of interpretability methods. *arXiv preprint arXiv:1806.08049*, 2018. 2

[4] Marco Ancona, Enea Ceolini, Cengiz Öztireli, and Markus Gross. Towards better understanding of gradient-based attribution methods for deep neural networks. *arXiv preprint arXiv:1711.06104*, 2017. 2

[5] Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller, and Wojciech Samek. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PloS one*, 10(7):e0130140, 2015. 1, 2, 5

[6] Alexey Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 8

[7] Rudresh Dwivedi, Devam Dave, Het Naik, Smiti Singhal, Rana Omer, Pankesh Patel, Bin Qian, Zhenyu Wen, Tejal Shah, Graham Morgan, et al. Explainable ai (xai): Core ideas, techniques, and solutions. *ACM Computing Surveys*, 55(9):1–33, 2023. 1

[8] Andre Esteva, Alexandre Robicquet, Bharath Ramspetsiuk2018risedar, Volodymyr Kuleshov, Mark DePristo, Katherine Chou, Claire Cui, Greg Corrado, Sebastian Thrun, and Jeff Dean. A guide to deep learning in healthcare. *Nature medicine*, 25(1):24–29, 2019. 1

[9] Amirata Ghorbani, Abubakar Abid, and James Zou. Interpretation of neural networks is fragile. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 3681–3688, 2019. 2

[10] Sorin Grigorescu, Bogdan Trasnea, Tiberiu Cocias, and Gigel Macesanu. A survey of deep learning techniques for autonomous driving. *Journal of field robotics*, 37(3):362–386, 2020. 1

[11] Meng-Hao Guo, Tian-Xing Xu, Jiang-Jiang Liu, Zheng-Ning Liu, Peng-Tao Jiang, Tai-Jiang Mu, Song-Hai Zhang, Ralph R Martin, Ming-Ming Cheng, and Shi-Min Hu. Attention mechanisms in computer vision: A survey. *Computational visual media*, 8(3):331–368, 2022. 1

[12] Leif Hancox-Li. Robustness in machine learning explanations: does it matter? In *Proceedings of the 2020 conference on fairness, accountability, and transparency*, pages 640–647, 2020. 1

[13] Sergiu Hart. Shapley value. In *Game theory*, pages 210–216. Springer, 1989. 2

[14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 4

[15] Anna Hedström, Leander Weber, Daniel Krakowczyk, Dilyara Bareeva, Franz Motzkus, Wojciech Samek, Sebastian Lapuschkin, and Marina M-C Höhne. Quantus: An explainable ai toolkit for responsible evaluation of neural network explanations and beyond. *Journal of Machine Learning Research*, 24(34):1–11, 2023. 2

[16] Robin Hesse, Simone Schaub-Meyer, and Stefan Roth. Funnybirds: A synthetic vision dataset for a part-based analysis of explainable ai methods. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3981–3991, 2023. 2

[17] Sara Hooker, Dumitru Erhan, Pieter-Jan Kindermans, and Been Kim. A benchmark for interpretability methods in deep neural networks. *Advances in neural information processing systems*, 32, 2019. 1, 2, 8

[18] Andrew Howard, Mark Sandler, Grace Chu, Liang-Chieh Chen, Bo Chen, Mingxing Tan, Weijun Wang, Yukun Zhu, Ruoming Pang, Vijay Vasudevan, et al. Searching for mobilenetv3. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1314–1324, 2019. 4

[19] Long Jin, Shuai Li, Jiguo Yu, and Jinbo He. Robot manipulator control using neural networks: A survey. *Neurocomputing*, 285:23–34, 2018. 1

[20] Pieter-Jan Kindermans, Sara Hooker, Julius Adebayo, Maximilian Alber, Kristof T Schütt, Sven Dähne, Dumitru Erhan, and Been Kim. The (un) reliability of saliency methods. *Explainable AI: Interpreting, explaining and visualizing deep learning*, pages 267–280, 2019. 1, 2

[21] Narine Kokhlikyan, Vivek Miglani, Miguel Martin, Edward Wang, Bilal Alsallakh, Jonathan Reynolds, Alexander Melnikov, Natalia Kliushkina, Carlos Araya, Siqi Yan, et al. Captum: A unified and generic model interpretability library for pytorch. *arXiv preprint arXiv:2009.07896*, 2020. 5

[22] Anastasia-M Leventi-Peetz and Kai Weber. Rashomon effect and consistency in explainable artificial intelligence (xai). In *Proceedings of the Future Technologies Conference*, pages 796–808. Springer, 2022. 2

[23] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30, 2017. 1, 2

[24] Sebastian Müller, Vanessa Toborek, Katharina Beckh, Matthias Jakobs, Christian Bauckhage, and Pascal Welke. An empirical evaluation of the rashomon effect in explainable machine learning. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 462–478. Springer, 2023. 2

[25] Menaka Narayanan, Emily Chen, Jeffrey He, Been Kim, Sam Gershman, and Finale Doshi-Velez. How do humans understand explanations from machine learning systems? an evaluation of the human-interpretability of explanation. *arXiv preprint arXiv:1802.00682*, 2018. 1

[26] Vitali Petsiuk, Abir Das, and Kate Saenko. Rise: Randomized input sampling for explanation of black-box models. *arXiv preprint arXiv:1806.07421*, 2018. 1, 2

[27] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. " why should i trust you?" explaining the predictions of any

classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144, 2016. 1, 2, 5

[28] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. Anchors: High-precision model-agnostic explanations. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018. 1

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

[30] Wojciech Samek, Alexander Binder, Grégoire Montavon, Sebastian Lapuschkin, and Klaus-Robert Müller. Evaluating the visualization of what a deep neural network has learned. *IEEE transactions on neural networks and learning systems*, 28(11):2660–2673, 2016. 2

[31] Gesina Schwalbe and Bettina Finzel. A comprehensive taxonomy for explainable artificial intelligence: a systematic survey of surveys on methods and concepts. *Data Mining and Knowledge Discovery*, pages 1–59, 2023. 1

[32] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017. 2, 5

[33] Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. Learning important features through propagating activation differences. In *International conference on machine learning*, pages 3145–3153. PMLR, 2017. 2, 5

[34] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*, 2013. 1, 2, 5

[35] Daniel Smilkov, Nikhil Thorat, Been Kim, Fernanda Viégas, and Martin Wattenberg. Smoothgrad: removing noise by adding noise. *arXiv preprint arXiv:1706.03825*, 2017. 1, 2

[36] Jost Tobias Springenberg, Alexey Dosovitskiy, Thomas Brox, and Martin Riedmiller. Striving for simplicity: The all convolutional net. *arXiv preprint arXiv:1412.6806*, 2014. 2, 5

[37] Johannes Stallkamp, Marc Schlipsing, Jan Salmen, and Christian Igel. The german traffic sign recognition benchmark: a multi-class classification competition. In *The 2011 international joint conference on neural networks*, pages 1453–1460. IEEE, 2011. 4, 7

[38] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *International conference on machine learning*, pages 3319–3328. PMLR, 2017. 1, 2, 5

[39] Hanxiao Tan. The generalizability of explanations. In *2023 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2023. 1, 2

[40] Feiyu Xu, Hans Uszkoreit, Yangzhou Du, Wei Fan, Dongyan Zhao, and Jun Zhu. Explainable ai: A brief survey on history, research areas, approaches and challenges. In *Natural Language Processing and Chinese Computing: 8th CCF International Conference, NLPCC 2019, Dunhuang, China, October 9–14, 2019, Proceedings, Part II 8*, pages 563–574. Springer, 2019. 1

[41] Kun Yu, Shlomo Berkovsky, Ronnie Taib, Dan Conway, Jianlong Zhou, and Fang Chen. User trust dynamics: An investigation driven by differences in system performance. In *Proceedings of the 22nd international conference on intelligent user interfaces*, pages 307–317, 2017. 2

[42] Jianming Zhang, Sarah Adel Bargal, Zhe Lin, Jonathan Brandt, Xiaohui Shen, and Stan Sclaroff. Top-down neural attention by excitation backprop. *International Journal of Computer Vision*, 126(10):1084–1102, 2018. 2

[43] Jianlong Zhou, Syed Z Arshad, Simon Luo, and Fang Chen. Effects of uncertainty and cognitive load on user trust in predictive decision making. In *Human-Computer Interaction–INTERACT 2017: 16th IFIP TC 13 International Conference, Mumbai, India, September 25-29, 2017, Proceedings, Part IV 16*, pages 23–39. Springer, 2017. 2

[44] Jianlong Zhou, Amir H Gandomi, Fang Chen, and Andreas Holzinger. Evaluating the quality of machine learning explanations: A survey on methods and metrics. *Electronics*, 10(5):593, 2021. 1

[45] Jianlong Zhou, Zhidong Li, Huaiwen Hu, Kun Yu, Fang Chen, Zelin Li, and Yang Wang. Effects of influence on user trust in predictive decision making. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems*, pages 1–6, 2019. 2