Robustness of Explanation Methods in XAI: A Comparison of Randomized Image Sampling for Explanations (RISE) and SmoothGrad on Image Classifiers

Sandro Kvrgić
Faculty of Social Sciences
Radboud University
Nijmegen Netherlands
s1091371

INTRODUCTION

The rise of AI technologies in high-stakes domains such as healthcare, law, security, and finance, led to a need to explain AI decisions (Chu et al., 2025). Explainable AI (XAI) is the branch of AI that tries to enable a better understanding of the decision-making processes that happen within AI models, by providing key insights into the question of why and how the models make specific predictions. Gradient-based saliency maps are a class of explanations that try to indicate which input features are the most important for the given predictions, by computing the output gradient with respect to input and visualizing how the changes in the input affect the model's output. Although this method provides us with much information, it is prone to errors due to sensitivity to small input changes, lack of robustness, and noise. To tackle these problems, we can use a technique called SmoothGrad, which improves saliency maps by taking the average of the gradients over multiple noisy inputs, leading to more interpretable explanations and a reduction in visual noise. This technique is widely used for its consistency in producing high-quality visualizations, while also being highly effective and simple to implement (Smilkov et al., 2017). Randomized Input Sampling for Explanation (RISE) is also an XAI technique that takes a different approach by treating the model as a black box. It does not rely on gradients, but it generates saliency maps by evaluating in what way do different random regions of an image affect the model's prediction. Its value lies in its flexibility, as it can be used with any model, regardless of the internal structure, and also it produces good predictions even when gradient-based methods cannot be used or fail. All of this makes RISE a really good model to use when gradient access is unreliable or limited (Petsiuk et al., 2018). These two techniques will be evaluated in detail throughout the report, where they are going to be checked for their robustness under various input perturbations, leading to a better understanding of how stable their explanations are after the images have been altered with noise, blurring, and flipping.

RESEARCH QUESTION

How robust are XAI methods SmoothGrad and RISE when applied to images under input perturbations such as Gaussian noise, blurring, and horizontal flipping?

Machine learning models require input data to produce some output, and that input data is rarely perfect, as it has many flaws. In our case, the data can have noise, motion blur, or even basic image transformations, and they all can have a major effect on the performance of XAI methods. It is important for XAI methods to be trusted by the users, and that is reached through consistency of the output, which is only possible if the performance of the model stays the same even if the input is perturbed. Therefore, this research question will try to give more insight and build trust with SmoothGrad and RISE methods, contributing to better usage and generalizability. As previously mentioned, XAI methods play a vital role in high-stakes decision-making, therefore it is of great importance that the output produced is of high quality, mitigating the risk of significant consequences.

METHODS

The implementation of this project was done by following publicly available code from Git Hub repositories. The SmoothGrad algorithm was implemented following the examples from tf-keras-vis documentation, while for RISE, the original code released by Petsiuk et al. (2018) alongside their paper was used. Finally, both SmoothGrad and RISE were applied using the pre-trained VGG16 model from Keras, allowing for realistic explanation class predictions. The dataset consists of 24 different images, which were provided to us on Brightspace for one of the assignments. At first, there were only three selected images loaded from a local directory, which were then resized to fit the VGG16 architecture, converted to RGB, and transformed into a NumPy array. Lastly, a function is used to normalize pixel values so that it is consistent with VGG training distribution. Only three pictures were used at the start, as the implementations of SmoothGrad and RISE are faster to train on a smaller dataset, and therefore also easier to debug. The bigger dataset of 15 images was then preprocessed using the same steps as mentioned above, only after the assurance that both algorithms were performing correctly. To those fifteen images were then applied three types of perturbations, specifically Gaussian noise, blurring, and horizontal flipping, which created an altered version of the original images. Gaussian noise was applied by sampling noise from a normal distribution while using a standard deviation of 25. Blurring was added to the images by using a Gaussian filter with a 5x5-sized kernel to simulate motion and focus distortion. Lastly, NumPy was used to implement horizontal flipping, which reverses width dimensions as well as pixel order, leading to an

inverted image. As these perturbations are common in the real world, their usage is perfect for the evaluation of both XAI methods. The whole implementation with the dataset can be found in this Git Hub repository: https://github.com/Sandrokvrgic/XAI-SmoothGrad-RISE.

Figure 1: This is a representation of a clean image, on which later both SmoothGrad and RISE will be applied.

SmoothGrad Method

SmoothGrad belongs to a class of gradient-based explanation methods. It improves normal saliency maps by reducing visual noise, and it does so by adding Gaussian noise to the input image multiple times, after which it averages the gradients. This provides a more interpretable visualization, while also improving the consistency of highlighting important regions (Smilkov et al., 2017).

For the implementation, a Gaussian noise of 0.2 was used to generate 20 noisy versions of each input image. The tf-keras-vis library was used to compute the gradients and to ensure cleaner and better explanations by receiving raw class scores, the softmax activation function was replaced temporarily for the linear activation in the final layer. Heatmaps were then used to visualize the averaged gradients. SmoothGrad was then applied to both clean and perturbated data, both counting fifteen different images.

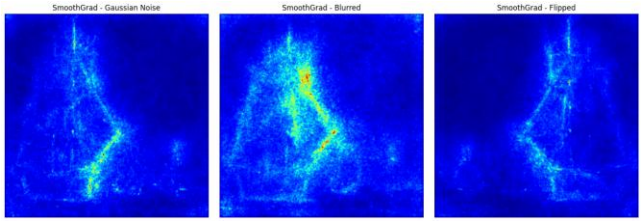

Figure 2: These images showcase how SmoothGrad performs under different perturbations for the image under figure 1.

Randomized Input Sampling for Explanation (RISE) Method

RISE is an XAI technique that uses a black-box approach with explanations by sampling a large number of randomized binary masks, which are then applied to the input image to observe the change in the model's prediction. RISE can generate a saliency map that highlights parts of the image most responsible for the prediction by correlating applied masks with the prediction scores (Petsiuk et al., 2018). The main idea behind this method is that regions which often contribute to a strong prediction, are more important and are given a bigger saliency.

The implementation of RISE in the project started by generating 2000 low-resolution random binary masks for each input image, which were then unsampled to match the input size. Element-wise multiplication was used to apply each mask to an input image, and the model's prediction confidence for the target class was saved. One final saliency map was computed from the weighted average of all masks, with weights corresponding to the output probabilities. It was then also applied to both clean and perturbated data.

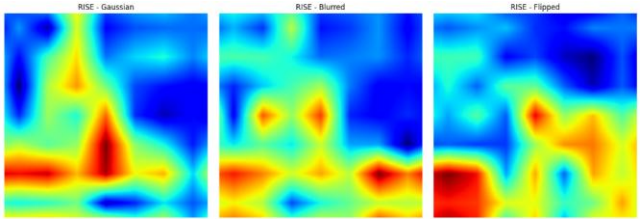

Figure 3: These images showcase how RISE performs under different perturbations for the image under figure 1.

RESULTS

The evaluation of the robustness of SmoothGrad and RISE was done by using three metrics: Structural Similarity Index (SSIM), Cosine Similarity, and Mean Squared Error (MSE). SSIM and Cosine Similarity tell us how closely the clean images match the explanation maps of the perturbed images, the higher the value is to 1, the bigger the similarity is. MSE is used to measure the average-pixel-wise error between two saliency maps, and in this case, we want to minimize the error and be closer to 0, for the best performance. The results are showcased in the table below.

	SSIM	Cosine	MSE
Gaussian - SmoothGrad	0.3537	0.8744	0.008
Gaussian - RISE	0.3762	0.9963	0.1023
Blurred - SmoothGrad	0.3264	0.8493	0.0104
Blurred - RISE	0.2951	0.9961	0.1238
Flipped - SmoothGrad	0.2584	0.8353	0.0134
Flipped - RISE	0.4827	0.9964	0.1047

Figure 4: This table represents all the metric scores computed over 15 images for each explanation method and perturbation.

DISCUSSION

The results from Figure 1 provide us with several interesting patterns for the robustness of SmoothGrad and RISE under input perturbations. We can observe that RISE shows consistently bigger values for both SSIM and Cosine Similarity compared to SmoothGrad. This observation indicates that RISE explanations under perturbations remain close to clean RISE explanations, meaning better robustness than SmoothGrad.

We can also observe that there is a big difference between values for RISE and SmoothGrad in horizontal flipping, as SmoothGrad produces a score of only 0.2584 and RISE archives a score of 0.4827. SmoothGrad maintains lower scores all over the board, and that is also the case for MSE, which just confirms that it has smoother saliency maps than RISE.

Although SmoothGrad is faster and easier to compute than RISE, the results suggest that RISE produces more stable and reliable explanations when input is slightly altered, making it more generalizable for real-world usage. It's also worth noting that SmoothGrad still performs well in Gaussian noise and blurred scenarios, making it a very attractive option in situations where speed is important and gradient access is available.

Although the dataset was limited with only 15 images being used, which could result in biased results, their consistency still indicates that we can draw meaningful conclusions from them. The conclusions from the results reinforce the idea that robust XAI methods are essential for real-world deployment, and that is especially true in high-stakes domains where trust in the model is of crucial importance, as it could lead to deadly consequences in healthcare, economic consequences in finance, security, and legal issues in those respective domains.

REFERENCES:

Chu, X., Tan, Z., Xue, H., Wang, G., Mo, T., & Li, W. (2025). *Domaino1s: Guiding LLM reasoning for explainable answers in high-stakes domains* (arXiv:2501.14431). arXiv. https://doi.org/10.48550/arXiv.2501.14431

Petsiuk, V., Das, A., & Saenko, K. (2018). *RISE: Randomized input sampling for explanation of black-box models* (arXiv:1806.07421). arXiv. https://doi.org/10.48550/arXiv.1806.07421

Smilkov, D., Thorat, N., Kim, B., Viégas, F., & Wattenberg, M. (2017). *SmoothGrad: Removing noise by adding noise* (arXiv:1706.03825). arXiv. https://doi.org/10.48550/arXiv.1706.03825

Word count: 1510