MDPI

Article

Trainable Noise Model as an XAI evaluation method: application on Sobol for remote sensing image segmentation

Hossein Shreim[1,2], Abdul Karim Gizzini[3] and Ali J. Ghandour[2]*

[1] Lebanese University;
[2] National Center for Remote Sensing - CNRS, Lebanon;
[3] Center for Digital Systems, IMT Nord Europe, Institut Mines-Télécom, University of Lille;
* Corresponding author: Ali J. Ghandour, aghandour@cnrs.edu.lb;

Abstract: eXplainable Artificial Intelligence (XAI) has emerged as an essential requirement when dealing with mission-critical applications, ensuring transparency and interpretability of the employed black box AI models. The significance of XAI spans various domains, from healthcare to finance, where understanding the decision-making process of deep learning algorithms is essential. Most AI-based computer vision models are often black boxes; hence, providing explainability of deep neural networks in image processing is crucial for their wide adoption and deployment in medical image analysis, autonomous driving, and remote sensing applications. Existing XAI methods aim to provide insights about the methodology used by the black-box model in making decisions by highlighting the most relevant regions within the input image that contribute to the model's prediction. Recently, several XAI methods for image classification tasks have been introduced. On the contrary, image segmentation has received comparatively less attention in the context of explainability, although it is a fundamental task in computer vision applications, especially in remote sensing. Only some research proposes gradient-based XAI algorithms for image segmentation. This paper adapts the recent gradient-free Sobol XAI method for semantic segmentation. To measure the performance of the Sobol method for segmentation, we propose a quantitative XAI evaluation method based on a learnable noise model. The main objective of this model is to induce noise on the explanation maps, where higher induced noise signifies low accuracy and vice versa. A benchmark analysis is conducted to evaluate and compare performance of three XAI methods, including Seg-Grad-CAM, Seg-Grad-CAM++ and Seg-Sobol using the proposed noise-based evaluation technique. This constitutes the first attempt to run and evaluate XAI methods using high-resolution satellite images.

Keywords: explainable artificial intelligence (XAI); remote sensing; XAI Evaluation; Semantic segmentation

Citation: Shreim, H.; Gizzini, A.; Ghandour, A. Trainable Noise Model as an XAI evaluation method: application on Sobol for remote sensing image segmentation. *Environ. Sci. Proc.* **2023**, *1*, 0. https://doi.org/

Published:

1. Introduction

Deep neural networks have achieved remarkable success in various computer vision tasks such as classification, detection, and semantic segmentation. However, they lack interpretability due to their black-box-based processing. Consequently, explainable artificial intelligence (XAI) is a crucial need in order to understand and interpret the decisions made by any deep learning black box model. Numerous XAI methods have been proposed [1–3] to offer valuable insights into the inner workings of the model and help build trust and confidence in its decision-making process. Generally speaking, XAI methods for image processing tasks provide explanations as saliency maps that highlight the most influential regions of the input that contribute significantly to the model's prediction. Most recent XAI methods are dedicated to classification tasks, where XAI for segmentation is still largely unexplored. There are two main categories of XAI methods [4]: (*i*) perturbation-based, where the concept is to perturb input features and record the effect of these changes on the model performance without diving into the internal architecture of the considered model, and (*ii*) gradient-based methods where the gradients of the output are calculated with

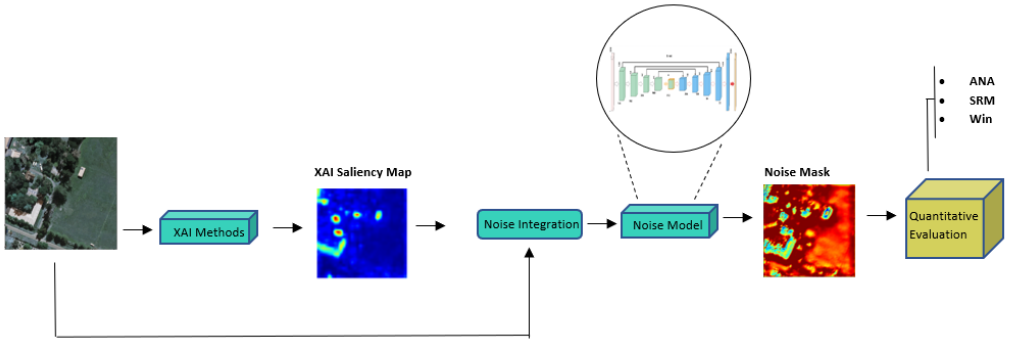

Figure 1. Propoosed quantatitave evaluation of XAI methods using U-Noise model.

respect to the extracted features or the input via backpropagation and used to estimate attribution scores. We note that internal access to the model architecture is essential in these methods.

Motivated by the fact that evaluating the performance and reliability of XAI methods is crucial to determine their efficiency and reliability for real-world applications. In this work, we propose a quantitative XAI evaluation approach that facilitates a deeper understanding of the performance of any XAI method. The proposed XAI evaluation approach is based on the methodology of the U-Noise model [3] that was initially used as an XAI method. The original U-Noise aims to interpret a pre-trained segmentation model by employing an external model that is responsible for adding noise to the input image without harming the accuracy of the pre-trained model. By doing this, the U-Noise model defines the most important pixels contributing towards the target class segmentation as those assigned low noise weights.

In this context, our proposed evaluation methodology is to feed the XAI saliency map multiplied by the input image to the U-Noise model. Therefore, the U-Noise model serves as a tool for assessing and quantifying the fidelity of XAI methods by adding noise to the important highlighted pixels. Inspired by the recent work proposed in [5], where the gradient-weighted class activation mapping (Grad-CAM) XAI method has been adapted from classification task to segmentation task. In this work, we adapt the recently proposed perturbation-based Sobol method [2] to segmentation. Rather than calculating the Sobol indices for a single classification output as performed in the original work [2], we calculate the Seg-Sobol indices with respect to multiple values of the segmentation output mask considering a specific target class.

To demonstrate the effectiveness of our proposed evaluation technique, we performed experiments on the satellite image dataset focusing on rooftop buildings segmentation model[6]. Our experimental results showcase the capacity of the proposed evaluation technique to compare the fidelity of different XAI methods, thus enabling a more comprehensive and objective assessment for any XAI method. To sum up, the contribution of this paper is three-folds:

- Propose a quantitative XAI evaluation approach using a learnable noise model. Our evaluation methodology is based on feeding the saliency map combined with the input image to the noise model. Then, on the basis of the generated noise mask, statistical metrics are computed to quantitatively evaluate the performance of any XAI method.
- Adapt the recently proposed perturbation-based Sobol XAI method from classification to semantic segmentation.
- Benchmark the performance of the adapted Sobol with the gradient-based XAI methods Seg-Grad-CAM and Seg-Grad-CAM++ using the WHU dataset for buildings' footprint segmentation.

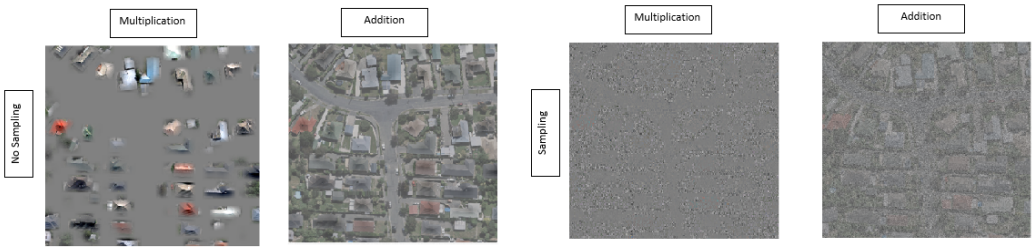

Figure 2. Different Integration Techniques.

2. Proposed Trainable Noise Model XAI Evaluation

2.1. Methodology

The saliency map of the XAI method assumes that the highlighted pixels contribute more to the model decision. To validate whether the highlighted pixels are really relevant to the model decision, XAI evaluation is a must. In this context, our proposed XAI evaluation approach is based on combining the saliency map generated by a specific XAI method with the original image and then feeding the resultant mask, denoted as the explanation map, to a trained U-Noise model. The U-Noise model is responsible for adding noise to the explanation map. A better XAI method would receive less added noise, as it retains the correct important pixels contributing to the model decision. Figure 1 illustrates the block diagram of the proposed U-Noise XAI evaluation approach.

In order to achieve a comprehensive evaluation analysis of XAI, the explanation maps are generated according to the following methodology. Given an original image $\boldsymbol{I}$ and its corresponding saliency map $\boldsymbol{L}_c$ generated by an XAI method where c denotes the target class, the explanation map $\boldsymbol{I}'$ can be manipulated as follows:

1. **Multiplication**: The original input image is directly multiplied by the saliency map, highlighting regions of the image assumed important by the XAI method as shown in Equation 1:

$$\boldsymbol{I}'_{\text{mul}} = \boldsymbol{I} \times \boldsymbol{L}_c. \tag{1}$$

2. **Addition**: By adding the saliency map to the original image, we augment the image with the importance scores, potentially highlighting regions of interest as shown in Equation 2:

$$\boldsymbol{I}'_{\text{add}} = \boldsymbol{I} + \boldsymbol{L}_c. \tag{2}$$

3. **Normal Sampling with Addition**: To introduce variability to the pixels of the explanation map, $\boldsymbol{L}_c$ is sampled from a normal distribution. The resulting sampled values are then added to the original image as shown in Equation **??**:

$$\boldsymbol{I}'_{\text{nsa}} = \boldsymbol{I} + N(\boldsymbol{L}_c). \tag{3}$$

4. **Normal Sampling with Multiplication**: Similar to the "Normal Sampling with Addition" method, but with multiplication instead of addition. This emphasizes or de-emphasizes regions based on the importance scores and the sampled noise as shown in Equation 4:

$$\boldsymbol{I}'_{\text{nsm}} = \boldsymbol{I} \times N(\boldsymbol{L}_c). \tag{4}$$

Figure 2 illustrates the proposed explanation map generation methods. We can clearly notice the impact of each method on generating the explanation map. It is expected that using the normal sampling with multiplication (Method 3) will not give reasonable evaluation since it is not informative to the U-Noise model, which was not trained on images with such distribution.

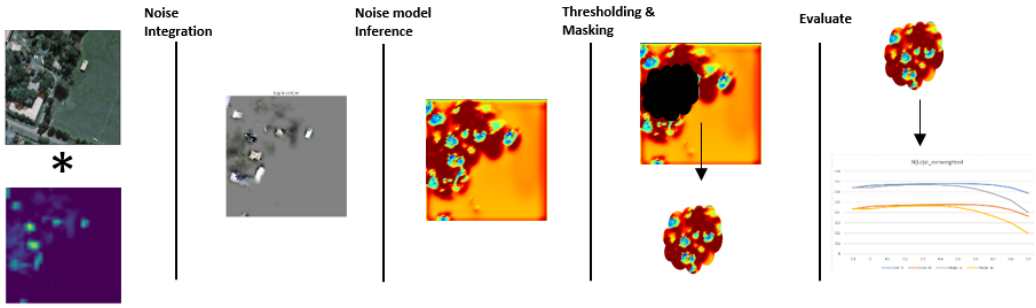

Figure 3. Thresholding operation as an additional step to overcome gray areas effect; We first integrate saliency map of XAI method with original image. Then, we run inference through the Noise model, and apply thresholding before we calculate the evaluation metrics.

2.2. Metrics

In this work, we propose the following two metrics in order to quantitatively report the results of the U-Noise model:

1. **Average Noise Added (ANA)**: This metric computes the mean value of the output of the U-noise model denoted by $\boldsymbol{O} \in \mathbb{R}^{u \times v}$. A higher ANA indicates that the XAI method introduced more noise into the input image.

$$ANA = \frac{1}{N} \sum_{(u,v)} O_{i,j}, \quad N = uv. \tag{5}$$

2. **Second Raw Moment (SRM)**: SRM measures the variance of the noise distribution. A higher SRM suggests that the noise introduced by the trained noise model is spread more away from zero.

$$SRM = \frac{1}{N} \sum_{(u,v)} (N_{i,j})^2 \tag{6}$$

3. Results

This section presents a quantitative evaluation of the U-Noise-based XAI using the proposed metrics. We note that we benchmark two recent gradient-based XAI methods, that is, Seg-Grad-CAM [5] and Seg-Grad-CAM++ [7], in addition to our adapted Seg-Sobol method. It is worth mentioning that to efficiently evaluate the benchmarked XAI methods, a thresholding operation should be applied to the generated noise mask. This is due to the presence of gray regions within the explanation map, as illustrated in Figure 3

Sobol XAI method [2] was initially developed for classification models, where the idea is to perturb the image with several noisy masks and calculate the Sobol indices for each input feature with respect to the output of the classification model taking into account the applied perturbation. The calculated Sobol indices reflect the impact of the applied perturbations on the prediction of the black-box model. For semantic segmentation, the Sobol indices should be calculated with respect to the summation of target-class pixels within the output probability mask. Sobol has the advantage of not needing to have access to the model's internal architecture. Figure 4 shows the steps to adapt the Sobol method to semantic segmentation, which we refer to as Seg-Sobol.

Seg-Sobol saliency map highlights the building surroundings with different intensities as important regions in segmenting building pixels. The results in Figure 5 are qualitatively plausible, where the highlighted buildings and regions are indeed thought to be important for the segmentation process.

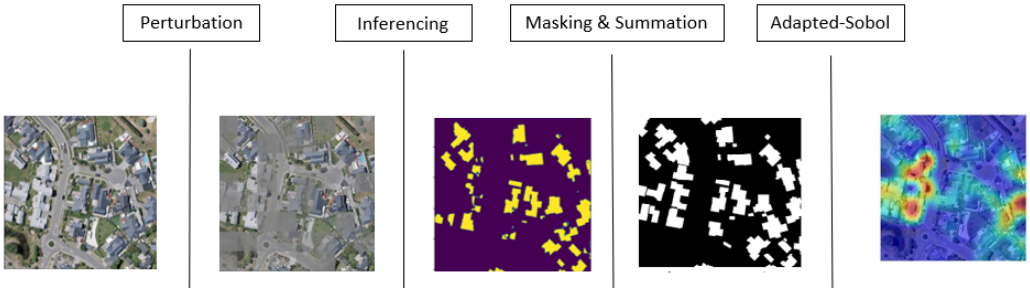

Figure 4. Seg-Sobol: Adaptation of Sobol method from classification to segmentation.

Figure 6 shows the average added noise and the second raw moment of the noise mask for the three XAI methods compared. Starting with a threshold of -0.1, which dictates that no thresholding was performed, the evaluation metrics were calculated on the whole noise mask. We can see that Seg-Grad-CAM++ had the lowest noise average, followed by Seg-Sobol and Seg-Grad-CAM. This is the case too with the second raw moment metric. The same results are also observed for the threshold value of 0. For threshold = 0.1, Seg-Grad-CAM receives the lowest average of noise and thus outperforms the other two methods. For threshold values between 0.2 and 0.8, we can see in Figure 6 that Seg-Sobol is the worst performing method (highest values of the average and second raw moment of noise).

Figure 5. Seg-Sobol results with grid size = 11 using sample from the WHU dataset.

4. Conclusions

In our research, we successfully adapted the Sobol XAI method to better understand image segmentation tasks. To evaluate its effectiveness, we introduced a unique noise model technique. When we compare Seg-Sobol with other methods such as Seg-Grad-CAM and Seg-Grad-CAM++, it showed promising results. Furthermore, using high-resolution satellite images for our tests was a new and important step. These findings are crucial because they make AI-driven image processing more transparent and easier to understand, paving the way for safer and more reliable real-world applications.

Author Contributions: Conceptualization, Hossein Shreimm Abdul Karim Gizzini, and Ali J. Ghandour; Data curation, Hossein Shreim; Formal analysis, Hossein Shreim, Abdul Karim Gizzini, and Ali J. Ghandour; Investigation, Hossein Shreim; Methodology, Abdul Karim Gizzini, Hossein Shreim; Project administration, Ali J. Ghandour; Resources, Ali J. Ghandour; Software, Hossein Shreim; Supervision, Abdul Karim Gizzini and Ali J. Ghandour; Validation Hossein Shreim, Abdul Karim Gizzini, and Ali J. Ghandour; Visualization,Hossein Shreim; Writing - original draft, Hossein Shreim; Writing - review & editing, Abdul Karim Gizzini, and Ali J. Ghandour;

Funding: "This research received no external funding".

Conflicts of Interest: "The authors declare no conflict of interest."

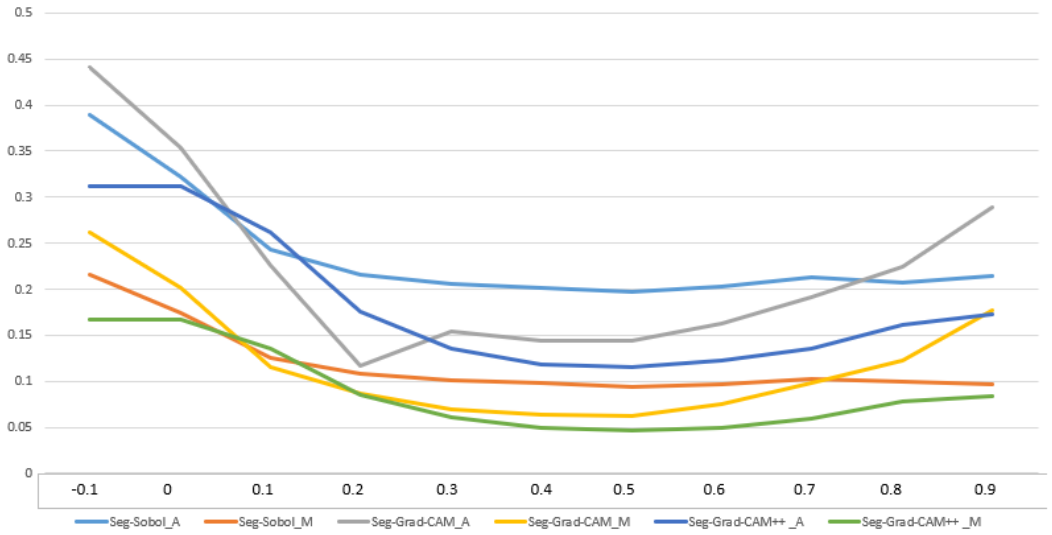

Figure 6. Quantitative metrics results for the benchmarked XAI methods using Method 1 ($\boldsymbol{I}'_{\text{mul}} = \boldsymbol{I} \times \boldsymbol{L}_c$) with different threshold values. Seg-Sobol_A and Seg-Sobol_M are the average and the second raw moment of the added noise added on Seg-Sobol, respectively. Seg-Grad-CAM_A and Seg-Grad-CAM_M are the average and the second raw moment of the added noise added on Seg-Grad-CAM, respectively. Seg-Grad-CAM++_A and Seg-Grad-CAM++_M are the average and the second raw moment of the added noise added on Seg-Grad-CAM++, respectively.

References

1. Jung, H.; Oh, Y. Towards better explanations of class activation mapping. In Proceedings of the Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 1336–1344.
2. Fel, T.; Cadène, R.; Chalvidal, M.; Cord, M.; Vigouroux, D.; Serre, T. Look at the variance! efficient black-box explanations with sobol-based sensitivity analysis. *Advances in Neural Information Processing Systems* **2021**, *34*, 26005–26014.
3. Koker, T.; Mireshghallah, F.; Titcombe, T.; Kaissis, G. U-noise: Learnable noise masks for interpretable image segmentation. In Proceedings of the 2021 IEEE International Conference on Image Processing (ICIP). IEEE, 2021, pp. 394–398.
4. Nielsen, I.E.; Dera, D.; Rasool, G.; Ramachandran, R.P.; Bouaynaya, N.C. Robust explainability: A tutorial on gradient-based attribution methods for deep neural networks. *IEEE Signal Processing Magazine* **2022**, *39*, 73–84.
5. Vinogradova, K.; Dibrov, A.; Myers, G. Towards interpretable semantic segmentation via gradient-weighted class activation mapping (student abstract). In Proceedings of the Proceedings of the AAAI conference on artificial intelligence, 2020, Vol. 34, pp. 13943–13944.
6. Nasrallah, H.; Samhat, A.E.; Shi, Y.; Zhu, X.X.; Faour, G.; Ghandour, A.J. Lebanon Solar Rooftop Potential Assessment Using Buildings Segmentation From Aerial Images. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* **2022**, *15*, 4909–4918.
7. Chattopadhay, A.; Sarkar, A.; Howlader, P.; Balasubramanian, V.N. Grad-CAM++: Generalized Gradient-Based Visual Explanations for Deep Convolutional Networks. In Proceedings of the 2018 IEEE Winter Conference on Applications of Computer Vision (WACV), 2018, pp. 839–847. https://doi.org/10.1109/WACV.2018.00097.