SLRP: Improved heatmap generation via selective layer-wise relevance propagation

Yeon-Jee Jung, Seung-Ho Han, and Ho-Jin Choi
School of Computing, Korea Advanced Institute of Science and Technology, South Korea

Deep learning has been recently applied to various areas of artificial intelligence, where it has displayed excellent performance. However, many deep-learning models are a black box, which makes it difficult to interpret the models and understand the predictions. Explainability is crucial for critical real-world systems (in the fields such as defense, aerospace, and security). To solve this problem, the concept of explainable artificial intelligence has emerged. For image classification, various approaches have been proposed to visually explain the model's prediction. A typical approach is layer-wise relevance propagation, which generates a heatmap, where each pixel value represents the contributions to the model's predictions. However, even advanced versions of layer-wise relevance propagation (such as contrastive layer-wise relevance propagation and softmax-gradient layer-wise relevance propagation) have some limitations. Here, selective layer-wise relevance propagation, which generates a clearer heatmap than the existing methods by combining relevance-based methods and gradient-based methods is proposed. To evaluate the proposed method and verify its effectiveness, we conduct comparative experiments. Qualitative and quantitative results show that selective layer-wise relevance propagation produces less noisy, class-discriminative, and object-preserving results. The proposed method can be used to improve the explainability of deep-learning models in image classification.

Introduction: Over the decades, machine learning (ML) has been applied to various artificial intelligence (AI) tasks, such as image classification, machine translation, and speech recognition. In particular, deep learning has shown excellent performance in recent years. For modelling complex and high-dimensional data using deep learning, non-linearity is an essential property. Deep-learning models include many hidden layers, where the output of each neuron is processed by a non-linear activation function. By passing multiple layers, the deep-learning models can understand given complicated data appropriately for the target domain.

The non-linearity property in deep learning is a powerful tool for modelling the data and increasing performance. However, non-linearity can make the interpretation of models difficult. For this reason, deep-learning models are considered as complex black-box models. This problem is known as a black-box problem in the deep-learning areas, representing a trade-off between performance and interpretability. The difficulty of a model's interpretation reduces its reliability. Therefore, despite the high performance of the model, it limits its deep-learning applications in critical domains (such as aerospace, defense, and security). To solve this problem, the explainable artificial intelligence (XAI) technique emerged, which attempts to interpret the deep-learning model and explain the model's prediction. Thus, the XAI strives to improve the interpretability of the deep-learning model without causing a performance drop. As a result, it aims to provide a trustable model with high performance.

Recently, some papers have been published to summarize the concepts of XAI and categorize the XAI approaches. In [1], the following two XAI approaches for ML are considered: transparent models and post hoc explainability. A model is considered to be transparent if it is understandable by design. The post hoc explainability aims to communicate understandable information about how an already developed model produces its predictions for any given input. This approach is further split into model-agnostic and model-specific techniques, and model-specific techniques include explainability in deep learning. Our approach targets image classification using a deep-learning algorithm.

From the point of XAI taxonomy in [1], our approach corresponds to feature relevance explanation and visual explanation approaches, which are model-specific techniques. Recent studies on feature relevance explanation approach (referred to as relevance-based method) include layer-wise relevance propagation (LRP) [2], CLRP [3], and SGLRP [4]. The goal of these studies is to generate a heatmap (or sensitivity map, salience map), where each pixel represents its contribution (or importance) to the model's prediction. Each pixel value in the heatmap is called a relevance score. If the relevance score for a pixel is higher than that of other pixels, the pixel contributed more to the model's prediction.

Although relevance-based approaches have been actively studied, as mentioned above, the existing studies have some limitations. In the case of the original LRP [2], the generated heatmap is noisy and non-class discriminative, and the heatmaps for different predictions are similar. More recent methods, such as CLRP [3] and SGLRP [4], aimed to solve these problems; however, some heatmaps generated by these methods show poor results when the target object encompasses the entire image. To overcome these limitations of the existing methods, we propose a novel heatmap generation method, selective layer-wise relevance propagation (SLRP), by utilizing the gradient value for each activation when calculating the relevance scores. The activations with a positive gradient for the target class contain the pixels of the target object with a high value, whereas the activations with negative gradients contain high pixel values for the background. With this concept, our approach is to select activations with a positive gradient, compute the weight of each gradient in each layer, and apply the weight to each relevance score.

Figure 1 shows an overview of the proposed SLRP. To generate the heatmap using our approach, first, the trained model predicts the class for the input image. Second, we initialize the initial relevance score using the value of the target class. With the initial relevance score, the SLRP computes the relevance scores of each layer from the output layer to the input layer. The final relevance scores represent the heatmap as a visual explanation for the model's prediction.

The remaining is structured as follows. In the next section, we briefly introduce the original LRP algorithm and describe our proposed method in detail. Subsequently, we show and discuss the experimental results qualitatively and quantitatively. Finally, we conclude our paper.

Relevance-based methods: Image classification using deep learning is conducted as follows. The input image is passed through each layer in the trained model by computing the activation values. Using the last activation values, the model classifies a proper class in the output layer. LRP [2] is a representative relevance-based method. The LRP algorithm assigns the relevance scores that represent the degree of contribution for the model's prediction to all the neurons in each layer. In particular, if the relevance score assigned to a neuron is close to zero, the neuron did not contribute to the prediction result. On the contrary, if the relevance score assigned to a neuron is large, it means that the neuron has significantly contributed to the prediction.

The relevance score is calculated by propagating the scores backward from the output layer to the input layer. The initial relevance value uses the output value for the target class. To explain the calculation process of
are computed as follows by using the $z^t$ (before softmax) for target class $t$ from a trained network with $L$ layers, where $1, \ldots, n, \ldots N$ are nodes in layer $l$, and $1, \ldots, m, \ldots M$ are nodes in layer $l + 1$. The relevance scores of the nodes in each layer are computed as follows by using the $z^t$ rule, which is the most used LRP rule.

$$R_n^t = \frac{\sum_{m} x_m^t w_{nm}^t \frac{\delta_{n}^{t+1}}{\delta_{m}^{t+1}}} {\sum_{m} x_m^t w_{nm}^t \frac{\delta_{n}^{t+1}}{\delta_{m}^{t+1}}},$$

where $R_n^t$ is a relevance score of $n$th node in layer $l$, $x_m^t$ is an input value of layer $l$ (i.e. activation value in layer $l$) and $w_{nm}^t$ is a positive weight value between layer $l$ and $l + 1$. The initial relevance score $R_n^L$ is defined as follows:

$$R_n^L = \begin{cases} z_n^L, & n = t, \\ 0, & \text{otherwise}. \end{cases}$$

where $t$ is the target class, and $z_n^t$ is an output value (before softmax) for class $t$. Starting from this value, the LRP algorithm calculates the relevance scores for each layer and last relevance score $R^0$, which has the same dimension of input image, produces a final heatmap, which is a visual explanation for class $t$.

The LRP approach has two limitations. First, the heatmap generated by LRP is noisy. The relevance score in the heatmap represents the corresponding contribution to the target class. If a pixel in the heatmap is located in the background, the relevance score for the pixel should be close to zero. However, in LRP-based heatmaps, the relevance scores in the background positions are quite high. Another problem of the LRP is its non-class discriminativeness: the LRP cannot generate different heatmaps for the same input image depending on the class. These problems are caused owing to the fact that the first relevance scores for non-target classes are initialized to zero, as represented in Equation (2). To solve these problems, more advanced LRP algorithms were proposed, such as CLRP [3] and SGLRP [4]. These approaches changed two major parts in the original LRP. First, they changed the initial relevance scores for non-target classes. The original LRP initializes the first relevance score to zero for non-target classes, whereas CLRP and SGLRP initialize it with a non-zero value. Second, they generated the final heatmap by subtracting the heatmap for the target class from the heatmap for the non-target class. Consequently, the sum of the heatmap is zero, because of which the final heatmap is class discriminative and clearer. Although the two methods above can generate a better heatmap than the original LRP, a critical problem is faced: the important portion of the image could be erased when generating a heatmap. Our approach aims to overcome the limitation of relevance-based methods. More details are provided in the next section.

Proposed method: SLRP: To overcome the limitations of the existing relevance-based methods described above, we propose a novel method called SLRP. LRP performs calculations using all the neuron activation values that are created during the forward pass process of the network. Unlike LRP, SLRP selects activations to include them in the relevance propagation. We visualized the activations with the positive gradient and the negative gradient, respectively. As a result, we found the following tendency: activations with a positive gradient were more relevant to the target object. In contrast, activations with a negative gradient were less relevant to the target object. Therefore, only activations with a positive gradient would be used in the relevance propagation in SLRP. In other words, the selection process includes calculating the gradient of the input of each layer for the target class and selecting the activations with the positive gradients. The selected activations are weighted by the corresponding gradient, so that activations with a larger gradient can have a greater impact on the final heatmap. After the selecting and weighting phase, relevance propagation is applied exactly as in LRP. The formulation used for relevance propagation in SLRP is given in Equation (3).

$$R_n^L = \frac{\sum_{m} R_m^L g_{nm} \frac{x_m^L}{w_{nm}} \frac{\delta_n^{L+1}}{\delta_{m}^{L+1}}} {\sum_{m} R_m^L g_{nm} \frac{x_m^L}{w_{nm}} \frac{\delta_n^{L+1}}{\delta_{m}^{L+1}}},$$

where $g_{nm}$ refers to the gradient of $n$th activation in $l$th layer for target class output. If the gradient is 0 or less, $g_{nm}$ becomes 0, and activation $x_m^L$ is not considered in the calculation. When applying SLRP in CNN, it is applied only in the convolutional layer, and not in all layers. Since it selects channel units, and not neuron units, a channel-level gradient is needed. To obtain a channel-level gradient, global average pooling (GAP) is applied to the pixel-level gradient. GAP outputs the spatial average of the activation of each unit; therefore, the gradient per channel can be obtained. Subsequently, channels with positive gradients are selected, and each channel is weighted with a gradient corresponding to the channel. This process is similar to that of Grad-CAM which is the other approach for visual explanation in [1]. However, there is a difference: it is applied in each layer and used during the LRP process.

The advantage of using activation selectively is that only the activation related to the target class is selected and included in the relevance propagation process, resulting in reduced noise compared to LRP that uses all activation. In other words, activations with negative gradients are not considered in the relevance propagation. Hence, the part with little relevance to the target class in the heatmap is less likely to have a high relevance score. In addition, LRP posed a problem that similar heatmaps were derived even if the target was changed. This was resolved in CLRP and SGLRP. Because SLRP calculates the gradient according to the target class, the gradient value is changed when the target class changes. For this reason, the heatmap produced using SLRP is class discriminative. Lastly, the sum of the heatmap is 0 in CLRP and SGLRP, which inevitably erases a part of the target object when the object encompasses the same image. Unlike the above methods, there is no constraint on the sum of the heatmap in SLRP, and therefore the target object does not appear to be erased. In other words, SLRP has the advantage that the entire target can be preserved in the heatmap.

Experiment: We performed qualitative and quantitative evaluations to compare our method with other methods. All experiments were conducted with the VGG16 model pre-trained in TensorFlow and the validation set of the ImageNet 2012 dataset [5], which includes 50,000 images.

For qualitative evaluation, we compared our results with that of LRP, CLRP, SGLRP, and Guided Grad-CAM. As Figure 2 shows, the heatmap using LRP, Guided Grad-CAM, and SGLRP captures the entire object well. Meanwhile, the heatmap using CLRP and SGLRP causes a part of the object to be erased. For example, consider the heatmap in the first row with “owl” as the target class in Figure 2; the results of CLRP and SGLRP show that only the face part of an owl has large values. On the other hand, the results of LRP, Guided Grad-CAM, and SGLRP show that the whole owl region has large values, among which SLRP has less noise around the owl compared with LRP. Similar aspects can be seen for other images. However, in LRP, sometimes the background or even the part irrelevant to the target object is included in the result, as shown in Figure 3. This problem does not occur in CLRP and SGLRP because of the property that the sum of the heatmap is 0. With CLRP, the results also show that unnecessary parts do not appear in the heatmap. For example, the heatmap in the first row with “kimono” as the target class is shown in Figure 3. The results of the LRP show that even windows unrelated to “kimono” were included in the heatmap. The results of CLRP, SGLRP, Guided Grad-CAM, and SGLRP show that only the part that represents “kimono” has high values. Similar aspects can be seen for other
value with the patch would decrease the probability of the target class. Therefore, removing the region around the highest value is based on the premise that the most relevant part of the heatmap proposed in [3], and different patch sizes were used in [4]. This experiment was conducted for qualitative evaluation. Maximal patch masking was better at explaining the model than LRP in the class-discriminativeness view.

Another strength of SLRP is class discriminativeness. Because heatmap is a tool for explaining a model, if the model comes up with a prediction that is different from the answer, it should inform you which part of the answer it has come up with. Using LRP produces almost similar heatmaps, even if the target class changes, so it remains unknown which part has led to the wrong prediction. In contrast, CLRP, SGLRP, and Guided Grad-CAM are class discriminative and show different results depending on the target class. SLRP also shows different results depending on the target class; hence, it can distinguish the features of the target class. The first and second rows in Figure 4 show the result of the target class of ‘palace’ and ‘monastery’, respectively. The ground-truth label for this image is ‘monastery’, and the model’s prediction is ‘palace’. Heatmaps in the second rows in Figure 4 show the result of the target class of ‘palace’ and ‘monastery’, respectively. The ground-truth label for this image is ‘monastery’, and the model’s prediction is ‘palace’. Heatmaps in the second rows in Figure 4 show the result of the target class of ‘palace’ and ‘monastery’, respectively. The ground-truth label for this image is ‘monastery’, and the model’s prediction is ‘palace’.

Two experiments—maximal patch masking and pointing game—were conducted for qualitative evaluation. Maximal patch masking was proposed in [3], and different patch sizes were used in [4]. This experiment is based on the premise that the most relevant part of the heatmap has the highest value; therefore, removing the region around the highest value with the patch would decrease the probability of the target class. A decreased probability indicates a more important removed part. The steps of this experiment are as follows. First, we classify the input image using the trained model. Second, we generate a heatmap with the aforementioned methods and find the maximal point of the heatmap. Third, we cover the maximal point with a patch and re-classify the patched image using the trained model to see how much the probability of the target class has decreased. The effects of removing the most relevant region are determined by each method with different patch sizes, \( p = 1, 3, 5, 7, 9 \), as in [4].

The pointing game was first proposed in [6], where hit and miss were determined using only the maximal point. In [3], the game was extended to include all relevant points, because the original pointing game was too forgiving. We used this extended pointing game for evaluation. First, the threshold heatmap to the foreground area covers \( p \) percent energy out of the whole energy, where the energy is the sum of pixel values in the heatmap. A hit is counted if the remaining area lies in the ground-truth bounding box of the target object. Otherwise, a miss is counted. Accuracy is calculated by the number of hits among the whole images for each remaining percent of energy. A higher pointing accuracy means that there is no value in the area except for the bounding box, so it can be seen as less noise. However, if the pointing accuracy is too high, we may suspect that a part of the object would also be erased. In a pointing game, the smaller the number of points, the higher the probability of being in the bounding box; therefore, the fewer values remaining in the heatmap, the higher the accuracy. The results are compared in Figure 6. The pointing accuracy of SLRP is higher than that of LRP with any percent of energy. The accuracies of CLRP, SGLRP, and SLRP are similar when there is less energy left. However, higher energy retention (when the remaining energy is 20% or more) means a higher pointing accuracy of CLRP and SGLRP over SLRP. Since the graph of SLRP is located between the graph of LRP and SGLRP, it can be interpreted that...
it produces a performance that erases noise well and does not erase the target object. In addition, as the pointing accuracy of SLRP and SGLRP is similar when the energy is low, we conclude that SLRP performs well for points with large relevance scores. However, Guided Grad-CAM is superior to all other methods with any percent of energy.

**Conclusion:** In summary, we proposed a novel relevance-based method called SLRP. LRP has the disadvantage of being noisy and class discriminative, and CLR and SGLRP have the disadvantage of erasing the part of target objects. The method we propose is a modified LRP, but unlike LRP, which uses all activations for propagation, activations are selectively used. The method of selection is used to calculate the gradient of each activation for the target class and include only the activations with positive gradients in the calculation. SLRP produces less noisy, class-discriminative, and object-preserving results. We performed qualitative and quantitative evaluations to compare our method with other methods, and it showed similar or better performance than other methods.

**Acknowledgement:** This research was supported by the Korea Electric Power Corporation (grant number: R18XA05).

© 2021 The Authors. *Electronics Letters* published by John Wiley & Sons Ltd on behalf of The Institution of Engineering and Technology

This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

Received: 7 September 2020  Accepted: 4 November 2020
doi: 10.1049/ell2.12061

**References**

1. Arrieta, A., et al.: Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Inf. Fusion* **58**, (2020)
2. Bach, S., et al.: On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PLoS One* **10**(7), 130140 (2015)
3. Gu, J., Yang, Y., Tresp, V.: Understanding individual decisions of CNNs via contrastive backpropagation. *Computer Vision ACCV 2018*, pp. 119–134. Springer International Publishing, New York (2019)
4. Iwana, B., Kuroki, R., Uchida, S.: Explaining convolutional neural networks using softmax gradient layer-wise relevance propagation. In: 2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW), pp. 4176–4185 (2019)
5. Deng, J., et al.: Imagenet large scale visual recognition competition 2012 (ILSVRC2012), p. 41 (2012)
6. Zhang, J., et al.: Top-down neural attention by excitation backprop. *Int. J. Comput. Vision* **126**(10), 1084–1102 (2017)