C^3-LRP: Visual Explanation Generation based on Layer-Wise Relevance Propagation for ResNet

Félix Doublet Seitaro Otsuki Iida Tsumugi Komei Sugiura

Keio University

In this paper, we focus on the task of visualizing important regions in an image as high-quality visual explanations of the model's decisions with a clear theoretical background. We introduce a novel calculation method for Layer-wise Relevance Propagation (LRP) specifically tailored to models featuring skip connections such as ResNet. This method's strength lies in its adaptability, as the backpropagation technique is distinctly defined for each layer, enhancing its extensibility. To validate our method, we conduct an experiment on the CUB-200-2011 dataset. The proposed method successfully generates appropriate explanations and, based on the Insertion-Deletion score, outperforms the baseline methods.

1. Introduction

The widespread adoption of neural networks underscores the critical importance of explainability of these models [Shrikumar 17] [Ribeiro 16]. The European Parliament has even proclaimed that AI systems must be safe and transparent in its AI Act [Madiaga 23] promulgated in December 2023. This strengthens the needs of accurate and meaningful explanations in neural network models.

However, current methods often lack transparency, leading the interpretation of the results to be a non-trivial task [Molnar 20]. Additionally, the black-box essence of neural network architectures tends to veil the underlying logic of their decision-making processes. This lack of transparency poses significant challenges in verifying the validity of the models' classifications, necessitating a rigorous assessment to determine if they are based on pertinent or irrelevant factors. To address this issue, research in eXplainable Artificial Intelligence (XAI) is aiming to delve into the inner workings of neural networks and to develop trustworthy models.

The generation of visual explanations within neural networks poses a significant challenge, necessitating the precise extraction of critical areas [Jacovi 23]. For example, the current Layer-wise Relevance Propagation (LRP) implementation encounters notable challenges when applied to ResNet architectures. ResNet's residual connections create non-linear and multiple relevance pathways, which are not handled by the typical relevance attribution process of LRP. Another difficulty in this process is the frequent absence of a clear and definitive ground truth, which serves as a benchmark for validating these explanations. This task demands a meticulous balance in identifying important areas, ensuring that the focus is neither excessive nor insufficient. Consequently, the explanations generated by LRP for ResNet models often do not show reliable or insightful results, as shown in Section 6. Therefore, while LRP offers a framework for relevance attribution in neural networks, its limitations, particularly in handling architectures like ResNet, presents a challenge, underscoring the pressing need for advanced methodologies in this domain.

For models based on Convolutional Neural Networks (CNNs), a multitude of studies have put forward methodologies for generating visual explanations, with notable examples including GradCAM [Selvaraju 17], LIME [Ribeiro 16], and RISE [Petsiuk 18]. These techniques predominantly rely on predefined computational approaches to formulate explanations. While these methods are generally agnostic to the specific architecture of the model, they encounter limitations in generating tailored explanations for ResNet architectures, often resulting in attending irrelevant areas, as shown in Section 6.

Another notable method is LRP [Bach 15], which employs backpropagation from the output for explanation generation. Nevertheless, this method has only been applied to models without residual connections, and its extension to models incorporating such connections such as ResNet remains an area yet to be explored.

In this study, we extend the conventional method of generating explanations for ResNet models by integrating LRP. LRP is distinguished by its well-established theoretical framework and transparent computational processes. Thus, using rule-based explainers, we are aiming to generate high-quality explanations. Our approach leverages the clarity and transparency inherent in LRP to elucidate the decision-making processes within ResNet models, thereby contributing to a deeper understanding and improved interpretability of these neural networks.

Our study distinguishes itself from existing research through two primary innovations.

- We introduce a novel calculation method for LRP, specifically tailored to models featuring skip connections. This novel approach is designed to adapt LRP's methodology to the unique architectural characteristics of such models, enhancing its applicability and effectiveness.
- We implement the Contour Component Choice (C^3) mechanism. This mechanism significantly elevates the quality of the generated explanations by strategically selecting the most pertinent area.

Contact: 3-14-1, Hiyoshi, Kohoku Ward, Yokohama, Kanagawa, felixdoublet@keio.jp

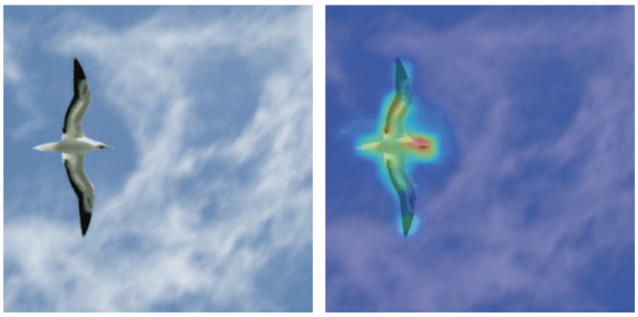


Figure 1: Example of an input image (left) and the visual explanation (right).

2. Problem Statement

In this paper, we focus on the task of visualizing important regions in an image as a visual explanation of the model's decisions. The pixels that contributed to the model's prediction should be attended.

Figure 1 shows an example of an CUB-200-2011 image. The left and right figures show the input image and the visual explanation, respectively.

The input is an image $\boldsymbol{x} \in \mathbb{R}^{c\times w\times h}$, where c, w, and h denote the number of channels, width and height of the input image, respectively. The output $p(\hat{\boldsymbol{y}}) \in \mathbb{R}^{c}$ denotes the predicted probability for each class, where C is the number of classes. Additionally, the importance of each pixel is obtained as a heatmap $\boldsymbol{\alpha} \in \mathbb{R}^{w\times h}$ which is used as a visual explanation. In this paper, we assume that the model is based on a ResNet architecture. The Insertion-Deletion score [Petsiuk 18] is used as an evaluation metric for explanation generation.

3. Proposed Method

3.1 Relevance backpropagation

We extend Layer-wise Relevance Propagation (LRP) [Bach 15] for use in models with residual connections, specifically developing LRP for ResNet [He 15] models. The extension implemented in our method defines the calculation method of LRP for models with residual structures. Consequently, this approach is generally applicable to models possessing residual blocks.

We propose a calculation method for LRP in models featuring skip connections. We also introduce C^3-LRP (Contour and Component Choice), that take the most significant regions from the generated attention areas based on the greatest contour.

ResNet50 consists of Convolution layers, Batch Normalization layers, Max Pooling layers, 16 Bottleneck layers, a Global Average Pooling (GAP) layer, and a Linear layer. Each Bottleneck consists of three layers – a 1×1 convolution layer for dimensionality reduction, a 3×3 convolution for processing, and another 1×1 convolution to restore dimensions. We treat each bottleneck block as a single dense layer with its unique LRP rule. The Relevance score can then be computed via backpropagation by applying an LRP rule for each encountered layer.

Figure 2 shows a schematic diagram of the calculation method for Relevance $\mathcal{R}(\boldsymbol{x})$ for ResNet50. The input is an image $\boldsymbol{x} \in \mathbb{R}^{3\times h\times w}$ where h and w denote the height and width of the image, respectively.

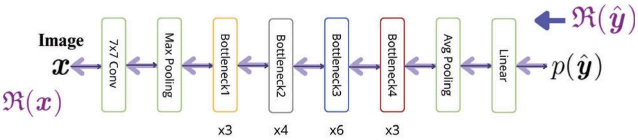


Figure 2: Schematic diagram of the calculation method for Relevance $\mathcal{R}(\boldsymbol{x})$ for ResNet50. The black path represents the forward propagation, where $p(\hat{\boldsymbol{y}})$ is computed based on the input image $\boldsymbol{x}$. The purple path illustrates how Relevance $\mathcal{R}(\boldsymbol{x})$ is calculated by back-propagating $\mathcal{R}(\hat{\boldsymbol{y}})$ using our method.

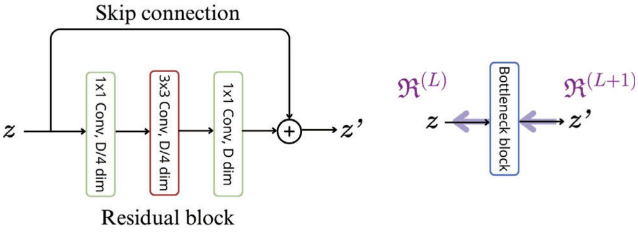


Figure 3: Schematic diagram of a Bottleneck layer from ResNet model (left) and how it is considered when computing the Relevance.

Figure 3 shows a schematic of the calculation method for the Relevance $\mathcal{R}^{(L)}$ of the layer L. For computing $\mathcal{R}^{(L)}$ based on $\mathcal{R}^{(L+1)}$, we consider a Bottleneck block as a single layer with $\boldsymbol{z}$ as an input and $\boldsymbol{z}'$ as an output.

This allows the backpropagation to be calculated in the same manner as for other layers. We consider $\boldsymbol{z} \in \mathbb{R}^{c_\mathrm{I}\times h_\mathrm{I}\times w_\mathrm{I}}$ and $\boldsymbol{z}' \in \mathbb{R}^{C\times U\times V}$ where C, U, and V represent the number of output channels, height, and width, respectively.

The backpropagation of $\mathcal{R}^{(L+1)}$ is represented as follows, with ε ensuring numerical stability by avoiding a zero division :

$$\mathcal{R}^L(z_{ij}) = \sum_u^U \sum_v^V \frac{z_{ij}\frac{\partial z'_{uv}}{\partial z_{ij}}}{z'_{uv}+\varepsilon}\mathcal{R}^{L+1}(z'_{uv}) \qquad (1)$$

For Convolution layers, Batch Normalization layers, Max Pooling layers and the GAP, the same formula is applied. For the Linear layer, the Relevance of a unit j in this layer L is computed as follows:

$$\mathcal{R}^L(h_j) = \sum_k \frac{w_{kj}h_j}{\sum_{j'} w_{kj'}h_{j'}+\varepsilon}\mathcal{R}^{L+1}(h'_k) \qquad (2)$$

where h_j, h'_k and w_{kj} denote activation of the j-th unit, activation of the k-th unit and the weight connecting unit j in the Linear layer to the unit k of the layer $L+1$, respectively.

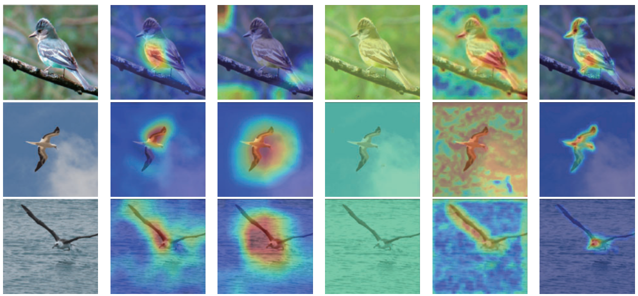
3.2 Obtaining the final heatmap with C^3

To prevent the relevance map to focus on inappropriate areas such as the background, we introduce C^3 to extract the most noteworthy area in order to obtain the finale heatmap $\boldsymbol{\alpha}_{C^3}$.

First, $\mathcal{R}(\boldsymbol{x})$ is reduced to a 28×28 $\boldsymbol{\alpha}_{C_R}$ to remove fine noise and unnecessary information. Then, based on the largest contour, a binary mask $\boldsymbol{\alpha}_{C_1}$ is created. As with C1C

Table 1: Quantitative comparison results between methods.

Method	Acc ↑	Insertion ↑	Deletion ↓	ID score ↑
RISE [Petsiuk 18]	**0.815 ± 0.001**	0.371 ± 0.015	0.043 ± 0.004	0.328 ± 0.004
GradCAM [Selvaraju 17]	**0.815 ± 0.001**	0.466 ± 0.019	0.156 ± 0.008	0.310 ± 0.020
LRP [Bach 15]	**0.815 ± 0.001**	0.063 ± 0.007	0.051 ± 0.006	0.011 ± 0.001
ABN [Fukui 19]	0.642 ± 0.009	0.282 ± 0.052	0.075 ± 0.011	0.207 ± 0.054
Ours	**0.815 ± 0.001**	**0.685 ± 0.015**	**0.017 ± 0.001**	**0.668 ± 0.015**



(a) Original (b) RISE [Petsiuk 18] (c) GradCAM [Selvaraju 17] (d) LRP [Bach 15] (e) ABN [Fukui 19] (f) Ours

Figure 4: Qualitative Results.

[Iida 23], the largest connected component is calculated and another binary mask $\boldsymbol{\alpha}_{C_2}$ is created. These two masks are then combined by taking the bitwise AND operation of the two masks to obtain $\boldsymbol{\alpha}_{C_3}$:

$$\boldsymbol{\alpha}_{C_3} = \boldsymbol{\alpha}_{C_1} \wedge \boldsymbol{\alpha}_{C_2} \tag{3}$$

The final mask $\boldsymbol{\alpha}_{C^3}$ is obtained by multiplying the original mask with the combined mask converted to a boolean array to keep the original pixel values where both the largest contour or the largest connected component are present:

$$\boldsymbol{\alpha}_{C^3} = \boldsymbol{\alpha}_{C_R} \odot \boldsymbol{\alpha}_{C_3} \tag{4}$$

Finally, $\boldsymbol{\alpha}_{C^3}$ is expanded to $w \times h$ to obtain the final relevance map $\boldsymbol{\alpha}$.

4. Experiment

4.1 Dataset

The Caltech-UCSD Birds-200-2011 (CUB-200-2011) dataset [Wah 11] was used for experimental evaluation. According to [Wah 11], a list of 278 bird species was compiled from an online field guide. Next, all images were downloaded from the corresponding Wikipedia page or fed to Flickr as query terms, and up to 40 images were downloaded for each species. We used the CUB-200-2011 dataset because it is a standard dataset for visual explanation generation tasks.

It contains 11,788 images of 200 subcategories belonging to birds. According to the standard setting [Wah 11], the dataset was divided with 5,794 images for training, 200 for validation and 5,794 for testing.

4.2 Experiment settings

The input images were resized to 224 × 224. During training, we flipped, rotated, cropped and changed brightness to images for data augmentation. The training, validation, and test sets were used for parameter training, hyperparameter validation, and evaluation, respectively. We used the SGD optimizer with a learning rate of 1.0×10^{-3} and a batch size of 32.

The number of parameters and the number of multiply-accumulate operations for the proposed method were 23.9M and 65.4G, respectively. For training, a GeForce3090 GPU and an Intel Core i9 processor were used. It took approximately 2 hours to train a ResNet50 model on the CUB-200-2011 dataset. The inference time was approximately 0.1s. We stopped the training if the loss on the validation set did not improve for six consecutive epochs.

4.3 Quantitative results

Table 1 presents the quantitative results of the comparison between the baseline methods and the proposed method. Five experiments were conducted for each method. Additionally, the bold values in Table 1 represent the best values. As the baseline methods, we used RISE [Petsiuk 18], GradCAM [Selvaraju 17], Layer-wise Relevance Propagation (LRP) [Bach 15], and Attention Branch Network (ABN) [Fukui 19]. The reason for choosing ABN as a baseline method is because it is a standard method that uses ResNet as the backbone network. On the other hand, RISE, GradCAM, and LRP were chosen as baseline methods because they are standard among methods applicable to generic models.

In this experiment, the evaluation metrics used were Accuracy, Insertion score, Deletion score, and Insertion-Deletion score (ID score). Additionally, as it is the most standard metric, we used ID score as the primary evaluation metric. The Insertion score and Deletion score are calculated as the Area Under Curve (AUC) of the Insertion and Deletion curves, respectively. Additionally, the ID score is defined as the difference between the Insertion score and Deletion score. The Insertion and Deletion curves represent the changes in prediction when important regions based on the final relevance map $\boldsymbol{\alpha}$ are inserted or deleted, respectively. The details are defined as follows. First, sort the elements of $\boldsymbol{\alpha}$ in descending order as $\alpha_{i_1,j_1}, \alpha_{i_2,j_2}, \cdots, \alpha_{i_w,i_h}$, and define the sets A_n, $\boldsymbol{i}_n$, $\boldsymbol{d}_n$ as follows:

$$A_n = \{(i_k, j_k) \mid k \leq n\} \tag{5}$$

$$(\boldsymbol{i}_n, \boldsymbol{d}_n) = \begin{cases} (x_{ij}, 0) & \text{if } (i,j) \in A_n \\ (0, x_{ij}) & \text{if } (i,j) \notin A_n \end{cases} \tag{6}$$

Here, n represents the number of pixels to insert or delete. When $\boldsymbol{i}_n$ and $\boldsymbol{d}_n$ are input into the model, we denote the outputs as $\boldsymbol{y}^{(\mathrm{ins},n)}$ and $\boldsymbol{y}^{(\mathrm{del},n)}$, respectively. The curves plotted for n, $\boldsymbol{y}_C^{(\mathrm{ins},n)}$ and $\boldsymbol{y}_C^{(\mathrm{del},n)}$ are the Insertion and Dele-

Table 2: Quantitative results of the Ablation Study.

Model	Method	Insertion ↑	Deletion ↓	ID score ↑
(i)	None	0.250 ± 0.017	0.018 ± 0.001	0.232 ± 0.017
(ii)	C1C [Iida 23]	0.642 ± 0.009	0.042 ± 0.004	0.600 ± 0.006
(iii)	C^3	**0.685 ± 0.015**	**0.017 ± 0.001**	**0.668 ± 0.015**

tion curves, respectively, where C represents the class to which $\boldsymbol{x}$ belongs.

From Table 1, in the primary metric, the ID score, RISE, GradCAM, LRP, ABN, and the proposed method scored 0.328, 0.310, 0.011, 0.140, and 0.668 respectively. The proposed method exceeded the highest among baselines, RISE, by 0.340 and achieved both the best performances for Insertion and Deletion scores. The performance difference in the primary metric, the ID score, and the Insertion score was statistically significant ($p < 0.05$).

4.4 Qualitative results

Figure 4 shows the qualitative results. Column (a) displays the original images, while columns (b)-(e) show the explanations generated by the baseline methods overlaid on the original images, and column (f) represents the results generated by the proposed method. From column (b) in Figure 4, the explanations generated by RISE strongly focus on specific parts like the bird's eyes and feathers, but fail to focus on the bird as a whole. Furthermore, column (c) shows that the explanations generated by GradCAM have attention regions that encompass the whole bird, but also focus on the background surrounding the bird. Next, from column (d), the explanations generated by LRP are inappropriate, as most regions have equal attention. Column (e) shows that the explanations generated by ABN are spotty and of low quality. On the other hand, column (f) shows that the proposed method focuses in detail on the entire bird, especially on the eyes, and has a low degree of attention to the background, thus generating appropriate explanations.

4.5 Ablation study

We investigated the contribution of C^3 by removing it (Model i) or replacing it with C1C [Iida 23] (Model ii).

From Table 2, the ID score in model (i) was 0.600, which is a decrease of 0.436 compared to model (iii). Also, the ID score in model (ii) was 0.600, which is a decrease of 0.068 compared to model (iii). This suggests that C^3 was effective to exclude the background and areas not directly related to classification, thereby promoting the generation of high-quality explanations.

5. Conclusion

This study dealt with the generation of visual explanations for the rationale in the multi-class classification problem of predicting the bird species from an image. The contributions of this research are as follows:

- We proposed a method for calculating Layer-wise Relevance Propagation (LRP) in models with residual connections, taking the example of ResNet.
- We introduced the C^3-LRP, which improves the quality of explanations by selecting the most noteworthy area based on the generated attention regions.

In the standard evaluation metrics for this task, such as the Insertion score and ID score, the proposed method outperformed the baseline methods.

Acknowledgment

This work was partially supported by JSPS KAKENHI Grant Number 23H03478, JST CREST, and NEDO.

References

[Bach 15] Bach, S., et al.: On Pixel-Wise Explanations for Non-Linear Classifier Decisions by Layer-Wise Relevance Propagation, *PLOS ONE*, Vol. 10, No. 7, pp. 1–46 (2015)

[Fukui 19] Fukui, H., Hirakawa, T., et al.: Attention Branch Network: Learning of Attention Mechanism for Visual Explanation, in *CVPR*, pp. 10705–10714 (2019)

[He 15] He, K., Zhang, X., Ren, S., and Sun, J.: Deep Residual Learning for Image Recognition, in *CVPR*, pp. 770–778 (2015)

[Iida 23] Iida, T., Otsuki, S., et al.: Visual Explanation Generation for Road Damage Classification by Using Layer-wise Relevance Propagation for Branch Networks, in *SIG-AM* (2023)

[Jacovi 23] Jacovi, A., Schuff, H., et al.: Neighboring Words Affect Human Interpretation of Saliency Explanations, *arXiv preprint arXiv:2305.02679* (2023)

[Madiaga 23] Madiaga, : Artificial intelligence act (2023)

[Molnar 20] Molnar, C., Casalicchio, G., et al.: Interpretable Machine Learning - A Brief History, State-of-the-Art, and Challenges, in *ECML PKDD*, pp. 417–431 (2020)

[Petsiuk 18] Petsiuk, V., Das, A., and Saenko, K.: RISE: Randomized Input Sampling for Explanation of Black-box Models, in *BMVC*, pp. 151–164 (2018)

[Ribeiro 16] Ribeiro, M., Singh, S., et al.: ”Why Should I Trust You?”: Explaining the Predictions of Any Classifier, in *KDD*, pp. 1135–1144 (2016)

[Selvaraju 17] Selvaraju, R., et al.: Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization, in *ICCV*, pp. 618–626 (2017)

[Shrikumar 17] Shrikumar, A., et al.: Learning Important Features Through Propagating Activation Differences, in *PMLR*, Vol. 70, pp. 3145–3153 (2017)

[Wah 11] Wah, C., Branson, S., Welinder, P., and others., : The Caltech-UCSD Birds-200-2011 Dataset, Technical Report CNS-TR-2011-001, California Institute of Technology (2011)