Layer-Wise Relevance Propagation with Conservation Property for ResNet (Supplementary Material)

Seitaro Otsuki[1], Tsumugi Iida[1], Félix Doublet[1], Tsubasa Hirakawa[2], Takayoshi Yamashita[2], Hironobu Fujiyoshi[2], and Komei Sugiura[1]

[1] Keio University, Japan
{otsu8sei14, tiida, felixdoublet, komei.sugiura}@keio.jp
[2] Chubu University, Japan
hirakawa@mprg.cs.chubu.ac.jp, {takayoshi, fujiyoshi}@isc.chubu.ac.jp

A Additional Details of Experimental Setup

Datasets. We used the Caltech-UCSD Birds-200-2011 (CUB) dataset [10] and the validation set of ImageNet [2] (ILSVRC) 2012 to evaluate our method. These datasets were chosen because they are standard datasets for visual explanation generation tasks. The CUB dataset contains 11,788 images from 200 classes of bird species. The validation set of ImageNet consists of 50,000 images from 1,000 classes.

Models. For the experiments on the CUB dataset, we employed ResNet50 [3] trained on it. To train the model, we divided the CUB dataset into training, validation, and test sets comprising 5,394, 600, and 5,794 samples, respectively. The official split of the CUB dataset provides only the training and test sets. Therefore, we used the official split for the test set and partitioned the official training set into our training and validation sets. We used the training set to train the model and the validation set to tune the hyperparameters. We evaluated the model on the test set. As part of the preprocessing stage, the input images were resized to a uniform dimension of 224×224 pixels. Moreover, during the training process, we applied flipping and cropping to the input images as data augmentation techniques. We used a ResNet50 model pretrained on ImageNet, which is publicly available.

Table A summarizes the experimental setups. We applied a learning rate decay such that the learning rate decreases by a factor of 10 every 30 epochs. We employ the cross-entropy loss as our loss function during the training of ResNet models. In our experiments on the CUB dataset, We stopped the training when the loss on the validation set did not improve for six consecutive epochs. We trained the models on a GeForce RTX 3080 with 16GB of memory and an Intel Core i9-11980HK with 64GB of memory. It took approximately 1 h to train a ResNet50 model on the CUB dataset. The inference time for a single sample using ResNet50 was approximately 4.66 ms. Our implementation required 40.7 ms to calculate the forward path and generate the attribution for each sample.

Table A: Experimental setup for the experiment on the CUB dataset.

Optimizer	SGD w/ momentum 0.9
Learning rate	1.0×10^{-3}
Weight decay	1.0×10^{-4}
#Epoch	300
Batch size	32
Image size	224

Theoretically, the proposed explanation method has the same computational cost as backpropagation.

Evaluation metrics. To quantitatively evaluate our method, we used the Insertion, Deletion, and Insertion-Deletion (ID) scores [5]. The Insertion and Deletion curves represent the changes in prediction when important regions based on the final attribution $\boldsymbol{\alpha}$ are inserted or deleted, respectively. The details are as follows. First, the elements of $\boldsymbol{\alpha}$ are sorted in descending order as $\alpha_{i_1 j_1}, \alpha_{i_2 j_2}, \ldots, \alpha_{i_N j_N}$ where N denotes the total number of pixels. The sets A_n, $\boldsymbol{i}_n$, $\boldsymbol{d}_n$ are defined as:

$$A_n = \{(i_k, j_k) \mid k \leq n\}, \tag{1}$$

$$(\boldsymbol{i}_n, \boldsymbol{d}_n) = \begin{cases} (x_{ij}, 0) & \text{if } (i,j) \in A_n, \\ (0, x_{ij}) & \text{if } (i,j) \notin A_n. \end{cases} \tag{2}$$

Here, n represents the number of pixels to insert or delete. Given inputs $\boldsymbol{i}_n$ and $\boldsymbol{d}_n$, the model outputs $\boldsymbol{y}^{(\mathrm{ins},n)}$ and $\boldsymbol{y}^{(\mathrm{del},n)}$, respectively. The Insertion and Deletion curves are defined as the curves plotted for n against the c-th elements of $\boldsymbol{y}^{(\mathrm{ins},n)}$ and $\boldsymbol{y}^{(\mathrm{del},n)}$, respectively, where c represents the class to which $\boldsymbol{x}$ belongs.

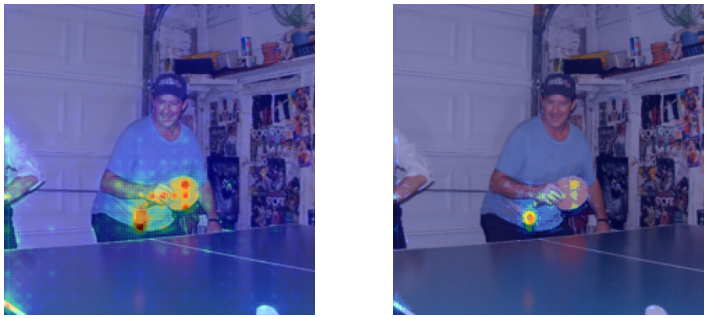
B Full Quantitative Results

Table B presents the full quantitative comparison between the proposed method and the several baseline methods: Layer-wise Relevance Propagation (LRP) [1], Integrated Gradients [9], Guided BackPropagation [8], DeepLIFT [7], Grad-CAM [6], and Score-CAM [11].

As listed in Table B, our method achieved an ID score of 0.582 in the experiments on the CUB dataset. The corresponding ID scores of LRP, Integrated Gradients, Guided BackPropagation, DeepLIFT, Grad-CAM, and Score-CAM are 0.011, 0.006, 0.028, 0.006, 0.453 and 0.457, respectively. The proposed method outperformed the best baseline method, Score-CAM, by 0.125 in terms of the ID score and achieved the best performance in terms of both the Insertion and Deletion scores.

Table B: Full quantitative results on ImageNet and the CUB dataset. IG and Guided BP denote Integrated Gradients and Guided BackPropagation, respectively. Ins. and Del. denote Insertion and Deletion score, respectively. The best results are marked in bold.

Dataset	CUB [10]			ImageNet [2]		
Metric [%]	Ins. (↑)	Del. (↓)	ID score (↑)	Ins. (↑)	Del. (↓)	ID score (↑)
Method						
LRP [1]	5.8 ± 0.2	4.7 ± 0.1	1.1 ± 0.0	9.5	8.3	1.1
IG [9]	2.0 ± 0.1	1.5 ± 0.1	0.6 ± 0.0	5.2	6.2	-1.1
Guided BP [8]	4.2 ± 0.2	1.4 ± 0.1	2.8 ± 0.2	11.5	5.7	5.7
DeepLIFT [7]	2.4 ± 0.2	1.8 ± 0.1	0.6 ± 0.1	5.6	6.3	-0.7
Grad-CAM [6]	50.8 ± 1.5	5.5 ± 0.4	45.3 ± 1.1	49.7	12.6	37.1
Score-CAM [11]	51.1 ± 1.7	5.4 ± 0.4	45.7 ± 1.4	48.8	13.3	35.5
Ours	**59.5 ± 1.0**	**1.4 ± 0.0**	**58.2 ± 1.0**	**56.3**	**1.8**	**54.5**



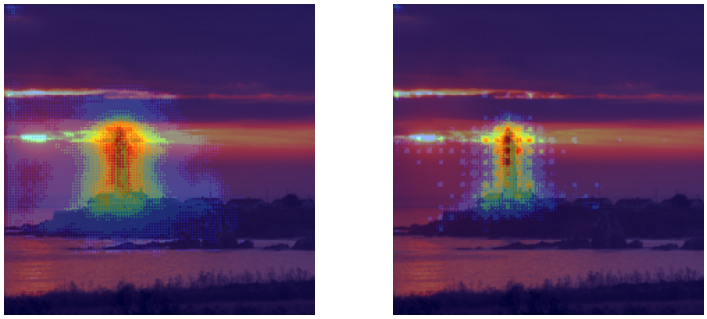
W/o Heat Quantization W/ Heat Quantization

Fig. A: The attribution map with respect to the ground-truth class: "Ping Pong Ball" without (left) and with (right) Heat Quantization.

Furthermore, as listed in Table B, in the ImageNet experiments, our method outperformed all the baselines on the ID score. Specifically, it exceeded the highest-scoring baseline method, Grad-CAM, by 0.174 on the ID score, and again achieved the best performances in terms of both the Insertion and Deletion scores.

C Qualitative Analysis of Heat Quantization

Fig. A shows the qualitative result of Heat Quantization. As shown in Fig. A, Heat Quantization removes the excessive attribution allocated to the irrelevant regions around the ping pong ball, such as the arm and the racket, while keeping the meaningful features.



Symmetric Splitting Ratio-based Splitting

Fig. B: The attribution map with respect to the ground-truth class: "Beacon" with Symmetric Splitting (left) and Ratio-based Splitting (right).

D Further Discussions

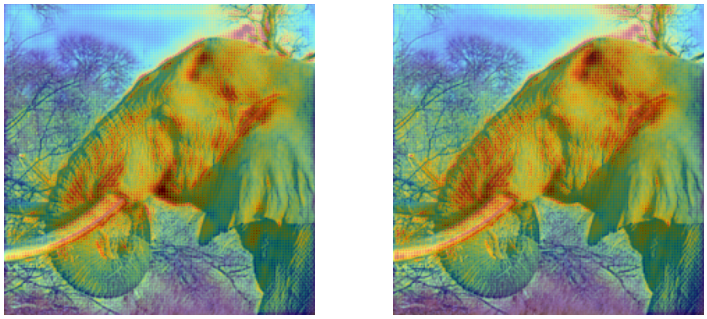
D.1 Discussion on splitting approaches based on qualitative comparison

We investigated the two types of splitting approaches—Symmetric Splitting and Ratio-Based Splitting—and their performance. Through the preliminary experiments and ablation study, Ratio-Based Splitting outperforms Symmetric Splitting in ID score. In this section, we briefly conduct a qualitative comparison between these two approaches.

As shown in Fig. B, we have observed that Symmetric Splitting results in dispersed attribution. In contrast, such dispersion does not occur with Ratio-Based Splitting. This difference matches the results of quantitative comparison. However, simultaneously, we have observed that Ratio-Based Splitting produced the grid patterns in the attribution map, though it can reduce the dispersion of attribution to non-relevant backgrounds. We leave the theoretical validation of the relationship between the formula of both approaches and these resulting attributions for future study.

D.2 Small performance downgrade on the deletion task from Method (i) to (ii) in Table 2

In table 2, while Method (ii) outperforms Method (i) in the primary metric (IDscore), there is a performance downgrade from Method (i) to (ii) for the deletion task. To elucidate the cause of this performance downgrade, we analyzed a sample in which Method (i) outperformed Method (ii) on the deletion score. As shown in Fig. C, in such samples, Method (ii) generates a slightly larger high-attribution region, indicated in red, than Method (i). This leads to a delayed drop in classification accuracy when higher attribution regions are deleted, thereby slightly downgrading the deletion score.



Method (i) Method (ii)

Fig. C: The attribution map with respect to the ground-truth class: "African Elephant" generated by Method (i) (left) and Method (ii) (right).

Table C: Comparison of different LRP rules on the ImageNet. Ins. and Del. refer to Insertion and Deletion scores, respectively.

Method [%]	Ins. (↑)	Del. (↓)	ID Score (↑)
LRP (ϵ rule)	9.5	8.3	1.1
LRP (Mixture of rules)	8.6	7.5	1.1
Ours (Mixture of rules)	64.6	0.2	64.4
Ours (z^+ rule)	56.3	1.8	54.5

D.3 Discussion on the LRP rules

We used the ϵ rule to obtain the baseline results. To investigate the impact of the choice of LRP rules for the baseline results, we conducted additional experiments using the mixture of rules. The mixture of rules employs the ϵ rule for layers after the eighth bottleneck block and the $\alpha 1\beta 0$ rule for the other layers. The $\alpha 1\beta 0$ rule is equivalent to the $z+$ rule [4], which we have adopted in our proposed method. As shown in Table C, we observed that our mixture of rules performs comparably to the ϵ rule.

Furthermore, to examine the impact of LRP rule selection on our proposed method, we applied the same mixture of rules to our approach. As shown in Table C, we did not observe any critical influence on the results by changing the rules; rather, the quantitative results slightly improved.

D.4 Error Analysis

We conducted an error analysis on ImageNet. We defined a failure of explanation generation as cases where the ID score was not greater than 0.371, the highest score achieved by the baseline methods (specifically by Grad-CAM [6]). There were 311 instances in which the explanation generation process failed. These failures represent 31.1% of the total generated attribution maps.

"Solar Collector" (a) Insufficiently Attended | "Oboe" (b) Over-Attended | "Bubble" (c) Wrongly Attended

Fig. D: Examples of failure categories in relevance attribution. (a) Insufficiently Attended where the area of focus is too narrow and fails to cover all relevant regions. (b) Over-Attended where an excessively large area is deemed relevant, indicating a lack of precision in identifying relevant features. (c) Wrongly Attended where the relevance is attributed to areas that do not directly contribute to the classification decision.

Table D: Number of failure cases in each failure category. IA, OA and WA are the three failure categories: Insufficiently Attended, Over-Attended, and Wrongly Attended, respectively. The number of the most frequent failures is highlighted in bold.

Failure category	IA	OA	WA
#Failure	**40**	25	35

For the error analysis, we investigated 100 samples selected in ascending order based on their ID scores. The causes of failure in these samples can be broadly categorized into three groups:

- IA (Insufficiently Attended): This category includes cases where the area of attribution is too small, indicating insufficient focus on relevant regions. Fig. D (a) illustrates an IA example.
- OA (Over-Attended): This category represents cases where the area of relevance is excessively large, suggesting that too much of the image is being considered relevant. Fig. D (b) presents an OA example.
- WA (Wrongly Attended): This category comprises cases where relevance is assigned to pixels that do not directly contribute to the classification. An example of WA is shown in Fig. D (c).

Table D indicates that the most common cause of errors is focusing on an insufficient area of the image. This observation suggests that our method might, in some instances, excessively restrict the distribution of attribution. This finding is consistent with our method's tendency to minimize the dispersion of attribution to non-relevant regions, as observed in Section 5.4 in the main text, which could contribute to this issue. This highlights a potential area for refinement in

our relevance score splitting rule. Adjusting the splitting rule accordingly could reduce the number of errors.

E Additional Qualitative Results

Figs. E and F show additional qualitative comparisons of our proposed method against baseline methods. In order to provide a clearer view, the panels show larger images than those presented in the main text. In each figure, the top row displays the original image, while the bottom row presents the attribution maps generated by our method. The rows in the middle show the results from baseline methods.

Attribution maps generated by the original Layer-wise Relevance Propagation (LRP) [1] tended to be noisy and often failed to sufficiently highlight relevant regions, as demonstrated in the second rows of Figs. E and F. The third and fourth rows of Figs. E and F present explanations generated by Grad-CAM [6] and Score-CAM [11], respectively. Results from both methods have attention regions that encompass the whole of the relevant objects, but also include the background surrounding them. In contrast, attribution maps generated by our proposed method sharply focus on the relevant objects and demonstrate minimal attention to irrelevant regions such as background, thus yielding more appropriate explanations.

References

1. Bach, S., et al.: On Pixel-Wise Explanations for Non-Linear Classifier Decisions by Layer-Wise Relevance Propagation. PLOS ONE **10**(7), 1–46 (2015)
2. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: ImageNet: A Large-Scale Hierarchical Image Database. In: CVPR. pp. 248–255 (2009)
3. He, K., Zhang, X., Ren, S., Sun, J.: Deep Residual Learning for Image Recognition. In: CVPR. pp. 770–778 (2016)
4. Montavon, G., Binder, A., Lapuschkin, S., Samek, W., Müller, K.R.: Layer-Wise Relevance Propagation: An Overview, pp. 193–209. Springer International Publishing (2019)
5. Petsiuk, V., Das, A., Saenko, K.: RISE: Randomized Input Sampling for Explanation of Black-box Models. In: BMVC. pp. 151–164 (2018)
6. Selvaraju, R., et al.: Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. In: ICCV. pp. 618–626 (2017)
7. Shrikumar, A., et al.: Learning Important Features Through Propagating Activation Differences. In: ICML. vol. 70, pp. 3145–3153 (2017)
8. Springenberg, J., Dosovitskiy, A., Brox, T., Riedmiller, M.: Striving for Simplicity: The All Convolutional Net. In: ICLR (workshop track) (2015)
9. Sundararajan, M., Taly, A., Yan, Q.: Axiomatic Attribution for Deep Networks. In: ICML. pp. 3319–3328 (2017)
10. Wah, C., Branson, S., et al.: The Caltech-UCSD Birds-200-2011 Dataset. Tech. Rep. CNS-TR-2011-001, California Institute of Technology (2011)
11. Wang, H., Wang, Z., et al.: Score-CAM: Score-Weighted Visual Explanations for Convolutional Neural Networks. In: CVPR. pp. 24–25 (2020)

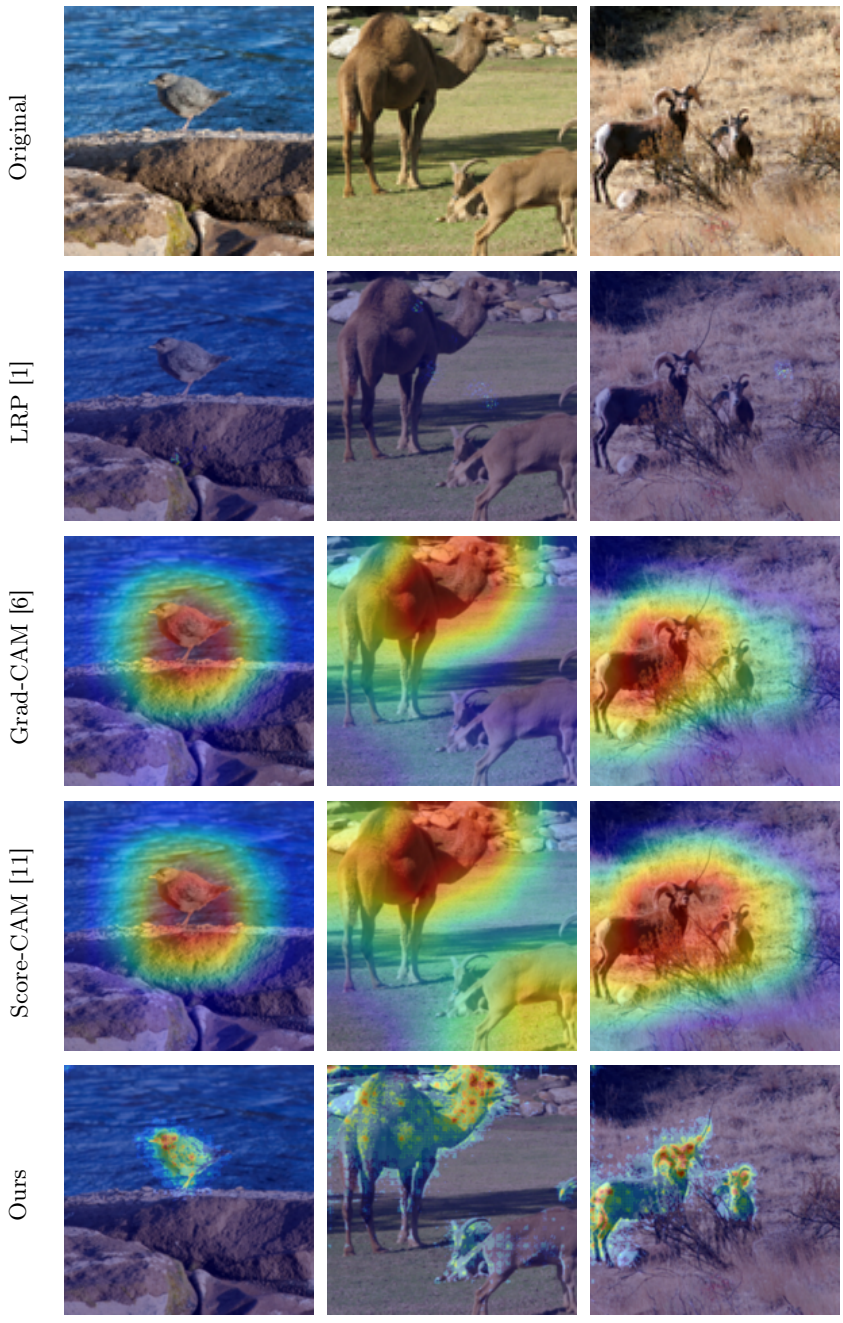


Fig. E: Attribution maps produced by each explanation method for the prediction of ResNet50 with respect to the ground-truth classes (left to right): "Water Ouzel," "Arabian Camel," "Ram."

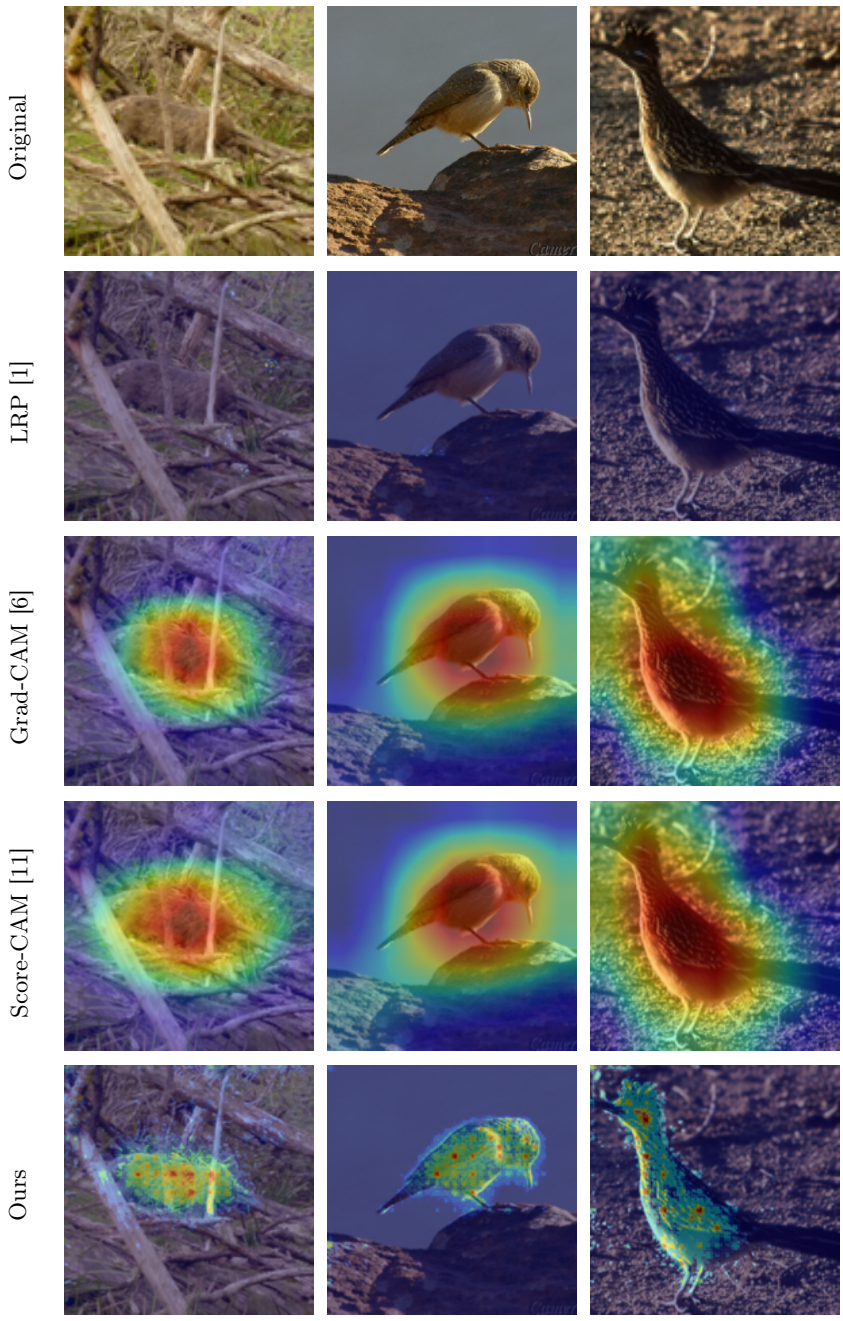


Fig. F: Attribution maps produced by each explanation method for the prediction of ResNet50 with respect to the ground-truth classes (left to right): "Wombat," "Rock Wren," "Geococcyx."