# Relevance-CAM: Your Model Already Knows Where to Look

Jeong Ryong Lee Sewon Kim Inyong Park Taejoon Eo Dosik Hwang*
School of Electrical and Electronic Engineering, Yonsei University

## Abstract

*With increasing fields of application for neural networks and the development of neural networks, the ability to explain deep learning models is also becoming increasingly important. Especially, prior to practical applications, it is crucial to analyze a model's inference and the process of generating the results. A common explanation method is Class Activation Mapping(CAM) based method where it is often used to understand the last layer of the convolutional neural networks popular in the field of Computer Vision. In this paper, we propose a novel CAM method named Relevance-weighted Class Activation Mapping(Relevance-CAM) that utilizes Layer-wise Relevance Propagation to obtain the weighting components. This allows the explanation map to be faithful and robust to the shattered gradient problem, a shared problem of the gradient based CAM methods that causes noisy saliency maps for intermediate layers. Therefore, our proposed method can better explain a model by correctly analyzing the intermediate layers as well as the last convolutional layer. In this paper, we visualize how each layer of the popular image processing models extracts class specific features using Relevance-CAM, evaluate the localization ability, and show why the gradient based CAM cannot be used to explain the intermediate layers, proven by experimenting the weighting component. Relevance-CAM outperforms other CAM-based methods in recognition and localization evaluation in layers of any depth. The source code is available at:* https://github.com/mongeoroo/Relevance-CAM

## 1. Introduction

Recently, deep learning is producing remarkable results with the development in GPUs and the advancements of new neural net architectures[10, 12, 13]. In this context, the field of interpretable deep learning is also being ardently studied[1, 24, 36, 4], especially in medical image processing [15, 7]. There are many methods for analyzing models in Computer Vision, and such as Class Activation Map(CAM) based methods [36, 25, 6, 30] and the decomposition based methods [27, 33, 9, 14, 19, 34, 4, 21, 3, 26]. CAM-based method calculates the weighted linear summation of the last convolutional feature map to visualize a model decision. The weights of fully connected layer between global average pooling output and target class output nodes are multiplied to the activation maps of the corresponding channels, and the sum of the weighted activation maps along the channel axis is the Class Activation Map. It can be understood as generating heatmaps describing the model decision by localizing where the model looks at. However, although CAM[36] can localize a target class, it is heavily constrained to the model architecture where the model must consist of Global Average Pooling(GAP) and one fully connected layer as its classifier. Grad-CAM[25] and Grad-CAM++[6] explain a model without the constraint to the model architecture. These gradient based CAMs use gradients of the target class output scores with respect to the last convolutional layer as the weighting components of activation maps. These methods are based on the

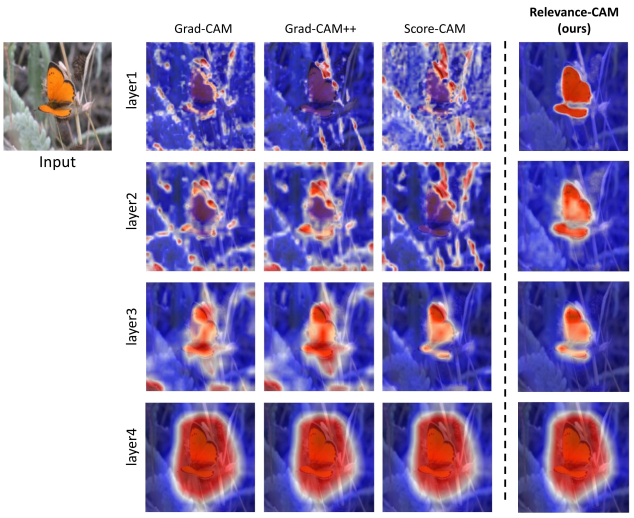


Figure 1. Visualization results of Grad-CAM[25], Grad-CAM++[6], Score-CAM[30] and proposed Relevance-CAM. The label of the input is a admiral butterfly. Relevance-CAM makes the high resolution heatmaps even in the shallow layer.

*Corresponding author.

idea that the sensitivity of activation to a target class can be understood as the importance of the activation map to the class.

Layer-wise Relevance Propagation(LRP)[4] is a decomposition-based method. LRP redistributes the model class output scores into input image through specific relevance propagation rule. It has a theoretical background on Deep Taylor Decomposition[19], and with this theoretical base, it is proved that LRP is robust to the shattered gradient problem[5] and show great performances[4, 20].

In this paper, we propose the novel explanation method which can analyze the model not only at the last convolutional layer but also at intermediate layers. Specifically, we propose the Relevance-weighted Class Activation Map (Relevance-CAM). As shown in Fig 1,Relevance-CAM outperforms other methods in visualizing target objects.

The contributions of this paper are as follows.

1. We propose the novel CAM-based method which is faithful and robust to shattered gradient problem. Therefore, it can operate well even at the intermediate layers. Relevance-CAM helps you to analyze the model in detail that other CAM-based method cannot. With Relevance-CAM, we find the surprising fact that the shallow layers which have small receptive fields can even extract the class specific features in some networks.

2. Through visualization using heatmaps, we show that our Relevance-CAM works effectively at any layer. Relevance-CAM especially outperforms other methods in localizing target objects in shallow layers.

3. We objectively evaluate faithfulness and localization ability of Relevance-CAM through Average Drop, Average Increase, and Intersection over Union. The proposed method outperforms the other CAM-based methods especially at the intermediate layers.

4. We demonstrate class sensitivity not only in deep layers but also in shallow layers. Relevance-CAM shows that even shallow layers can extract class specific information.

# 2. Background

## 2.1. CAM

Class Activation Mapping(CAM)[36]is an explanation method for visualizing class specific regions through a linearly weighted combination of the last convolutional layer output before the global pooling layer. However, this restricts the application of CAM to the models with specific architectures.

## 2.2. Grad-CAM

Grad-CAM[25] is similar to CAM except for the method of calculating the weight values. It is designed to generalize CAM and can be applicable to all CNN models. The motivation of Grad-CAM is that activation maps are the feature maps extracted by a certain convolutional layer, and the importance of each activation map to a class can be defined as the gradients of the activation maps. Grad-CAM, $L^c_{Grad-CAM}$, is defined as:

$$L^c_{Grad-CAM} = \sum_k \alpha^c_k A_k \tag{1}$$

*where*

$$\alpha^c_k = GP(\frac{\partial y^c}{\partial A_k}) \tag{2}$$

where $A_k$ denotes the activation map in the $k$-th channel of the last convolutional layer, $y^c$ denotes the model output for the class c, $\alpha^c_k$ denotes the weighting component of Grad-CAM and $GP(\cdot)$ represents Global Pooling function.

## 2.3. Layer-wise Relevance Propagation(LRP)

LRP[4] incorporates divide & conquer strategy. In general, it can be difficult to define the relevance between a model's input pixels and output scores. But the task becomes easy in cases of single layer models. LRP explains a model through layer wise decomposition of its structure[19], propagating the relevance score from the output to the input in layerwise manner. The propagation of the score is proceeded while meeting the following definitions.

**Definition 1** A relevance score is conservative if the sum of assigned relevance in the pixel space corresponds to the total relevance detected by the model:

$$\forall x : f_c(x) = \sum_p R^l_p(x). \tag{3}$$

**Definition 2** A relevance score is positive if all values forming the heatmap are greater or equal to zero:

$$\forall x, p : R_p(x) \geq 0 \tag{4}$$

where $R^l_p(x)$ denotes the relevance score of the pixel $p$ on $l$-th layer.

Based on the conservative definition 1, the total redistributed relevance is always same to $f_c(x)$ which denotes the output score for the target class $c$. The relevance score is interpreted directly as the value of contribution to the output score.

The general relevance propagation rule is the z-rule, which has theoretical basis on the Deep Taylor Decomposition[19]. When propagating the relevance score from the j-th layer to the i-th layer, the z-rule can be written as:

$$R_i = \sum_j \frac{z_{ij}^+}{\sum_i z_{ij}^+} R_j \tag{5}$$

*where*

$$z_{ij}^+ = x_i w_{ij}^+ \tag{6}$$

where $R_i$, $R_j$ denote the $i$-th layer relevance and the $j$-th layer relevance, respectively, $x_i$ denotes the activation output of the $i$-th layer, and $w_{ij}^+$ denotes the positive part of weight between the $i$-th and the $j$-th layer. While the z-rule performs successfully on model explanation tasks[4], the drawback of the z-rule is that it is less sensitive to the target class in multi object image[9, 14].

### 2.4. Contrastive Layer-wise Relevance Propagation(CLRP)

Contrastive Layer-wise Relevance Propagation (CLRP) [9] was proposed to resolve LRP's drawback: low sensitivity to the target class. CLRP subtracts the relevance of the non-target classes from the relevance of the target class. As a result, the heatmaps generated by CLRP become more sensitive to the target class.

The relevance score of the final layer in CLRP is:

$$R_n^{(L)} = \begin{cases} z_t^{(L)} & n = t \\ -\frac{z_t^{(L)}}{N-1} & otherwise \end{cases} \tag{7}$$

where $z_t^{(L)}$ denotes the model output value for the target class index $t$ on $L$-th layer and $N$ denotes the number of class. By assigning the relevance in this way, the relevance pixels of the non-target class are removed from the generated saliency map.

### 2.5. Gradient Issue

**Noisiness and discontinuity:** There are many studies raising questions to the faithfulness of gradients as a model explanation tool[30, 20, 5, 16]. As a network deepens, gradients become noisy and discontinuous. This problem is called the shattered gradient problem[5]. The noisiness of gradients comes from the saturation of gradients in activation functions; when passing through activation functions such as ReLU or Sigmoid, gradients can become saturated, vanishing or exploding in value as a result. Moreover, the piece-wise linearity of gradients causes discontinuity. Because gradients are calculated by the weights of the connected layers, relations to the next pixels are lost.

**Explanation to sensitivity:** In Grad-CAM, the importance of an activation map is measured by the gradient of output w.r.t the activation map. This indicates that Grad-CAM does not take account for the activation value in assigning the importance. Thus, Grad-CAM measures the sensitivity of an activation map towards the model output. But what we want to explain is how much an activation map contributes to a target class output and not how sensitive an activation map is. This issue is called False Confidence[30].

**Relevance Score:** On the contrary, the LRP is robust to the gradient issue, as proven by experiments [20]. LRP shows continuity and less noisy characteristics as a target class score changes. For these reasons, we consider the relevance score of LRP as the weighting components of class activation mapping.

## 3. Relevance-weighted Class Activation Map

Before Relevance-CAM, the other previously mentioned methods concentrated on analyzing activation maps of the last convolutional layer as it is known that the last convolutional layer has high-level semantics. Also, the fact that the heatmaps of intermediate layers created by the previous methods are noisy and show non-class specific result further encouraged such approach.

However, it is not only the last convolutional layer that affects a model output. Effects of deep and intermediate layers to a model outputs are to be analyzed as well. But as gradients become noisy and discontinuous in deep layered models, the quality of gradients as a weighting component become questionable. Therefore, we devise the new CAM-based method namely Relevance- CAM, and as a result, we are able to obtain meaningful information from the model even from layers of shallower depth that was not formally found in the previous methods. What we found with Relevance-CAM is that the intermediate layers also can extract class specific information as well as the last convolutional layer. Our findings are detailed in section 4.

Considering the gradient issue mentioned in section 2, the relevance score obtained through LRP are used as the weighting component of our proposed method. In addition, when performing this procedure, CLRP is applied to get the class sensitivity. The pipeline of Relevance CAM is shown in fig 2.

Relevance-CAM equation is as follows:

$$L_{Relevance-CAM}^{(c,i)} = \sum_k \alpha_k^{(c,i)} A_k^c \tag{8}$$

*where*

$$\alpha_k^{(c,i)} = \sum_{x,y} R_k^{(c,i)}(x,y) \tag{9}$$

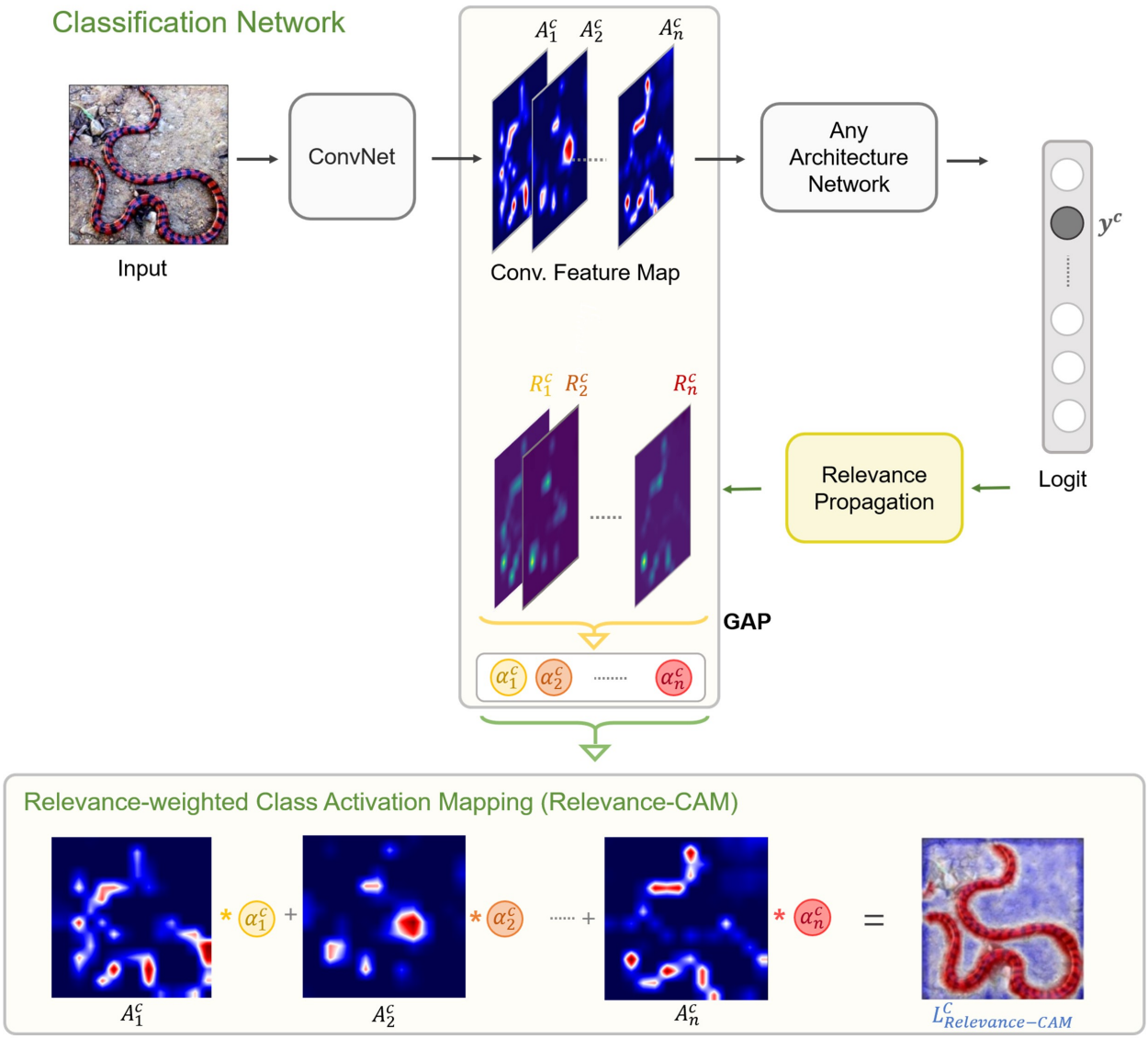


Figure 2. Relevance-CAM pipeline. Activation maps, $A_k^c$, are extracted during forward propagation and Relevance Maps, $R^{(c,i)}$, are calculated by Relevance propagation process. And the weighting components are obtained by global average pooling of relevance map. Finally, Relevance-CAM is obtained by weighted linear summation of activation maps

$R_k^{(c,i)}$ denotes the relevance map of the $i$-th layer feature map $k$ for the target class $c$ that can be obtained through the LRP process. $\alpha_k^{(c,i)}$, namely the weighting component, is calculated by global average pooling of the relevance map, $R_k^{(c,i)}$. Since Relevance-CAM can be obtained through activation maps and relevance maps, it can be calculated with only one forward propagation and one backpropagation.

Also, looking at the meaning of the relevance weight, since LRP is a deep taylor decomposition of output scores, the relevance value itself can be interpreted as the contribution to a target class output. Therefore, the sum of the relevance, $\alpha_k^{(c,i)}$, represents the importance or contribution of the $k$-th channel activation map to the target class output score.

## 4. Experiment

In section 4, we evaluate our post-hoc attention method through various experiments. First, we visualize the depth-wise heatmap made by various CAM-based method to compare the effectiveness of our method and measure the faithfulness of generated heatmap through Average Drop(A.D.) and Average Increase(A.I.). Second, we demonstrate the difference between Grad-CAM and Relevance-CAM. Third, we evaluate the localization ability through Intersec-

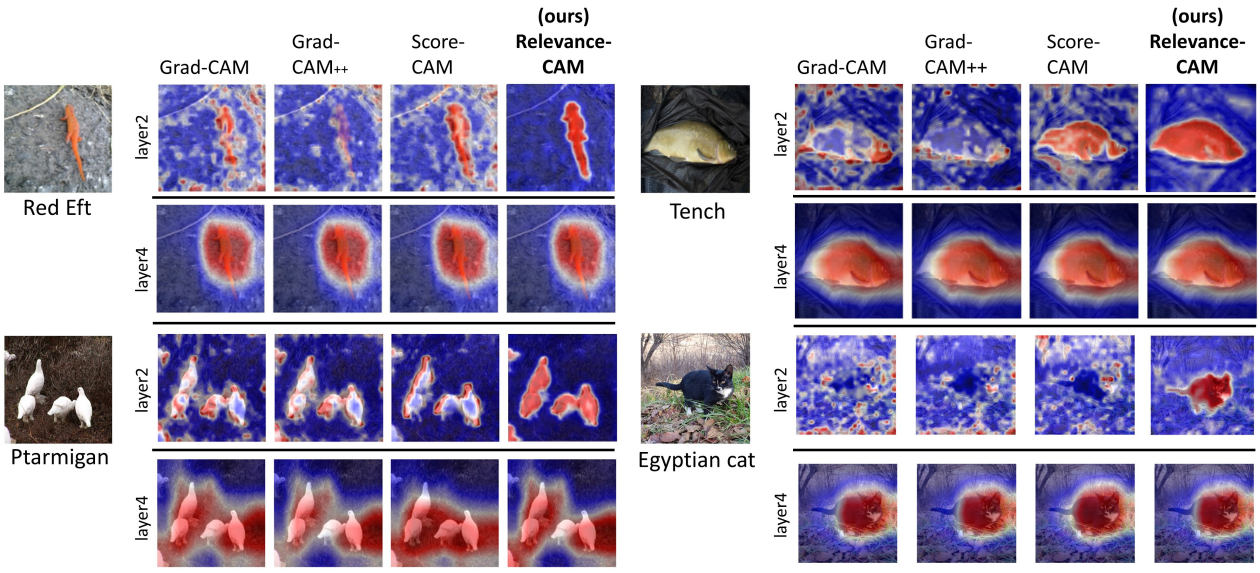


Figure 3. Comparison of various methods. The columns are divided by the explanation methods. The rows are divided along the layer depth. layer 2 is the intermediate layer and layer 4 is the last convolutional layer. In the deep layer, layer4, the heatmaps are similar for the various methods. But in the shallow layer, layer2, the quality of Relevance-CAM is better than that of the other methods in localizing the target objects. And Relevance-CAM shows high resolution heatmaps at low level layer.

tion over Union(IoU). Finally, we assess the class sensitivity of our proposed method.

### 4.1. Depth-wise visualization

We visualize the depth-wise heatmap through the various explanation methods on fig 3. All of the explanation methods localize the target object well in layer 4. But in layer 2, gradient-based methods such as Grad-CAM and Grad-CAM++ cannot localize class specific regions even though Score-CAM and Relevance-CAM localize well. Since the gradients becomes noisy as it pass through the layers, the gradient-based CAM cannot accurately assign weights to the activation map. On the other hand, Score-CAM and Relevance-CAM that are robust to the gradient problems can work well even in shallow layers, although the heatmaps of Relevance-CAM are clearer than the heatmaps of Score-CAM. When using gradient-based CAM to analyze intermediate layers, it may be judged that the intermediate layers cannot extract class semantic information since the generated saliency maps cannot localize the class object[25]. However, the saliency maps generated through Relevance-CAM localize the target class well even from the intermediate layers. This indicates that the shallow layers also can extract class specific information.

We further evaluate the objective faithfulness of our method through Average Drop(A.D.) and Average Increase( A.I.) as adopted in [6, 30]. The two metrics measure how well an attention map explains a model through observing the change of a target class score. In this experiment, top 50% pixels of the attention map are used as the mask. Average Drop is expressed as:

$$\sum_{i=1}^{N} \frac{max(0, Y_i^c - O_i^c)}{Y_i^c} \times \frac{100}{N} \tag{10}$$

Average Increase is expressed as:

$$\sum_{i=1}^{N} \frac{Sign(Y_i^c < O_i^c)}{N} \tag{11}$$

where N is the number of dataset, $Y_i^c$ is the output softmax value for class $c$ on image I and $O_i^c$ is the output softmax value for class $c$ with the attention masked image as the input. $Sign(\cdot)$ denotes the indicator function that return 1 if input is positive. we experiment the explanation methods on 2000 randomly selected images from the ImageNet(ILSVRC2012) validation set, but only the cases which the labels and the model predictions match are considered to measure the contribution exactly.

As shown in Table 1 and Table 2, in both ResNet 50 and VGG16, the values of A.D. and A.I. have almost similar values in the deepest layers of the models. However, the differences in the evaluation values between each explanation methods increase in the lower level layers.

For ResNet 50, in the case of Relevance- CAM, the difference in evaluation between layer 2 and layer 4 is not

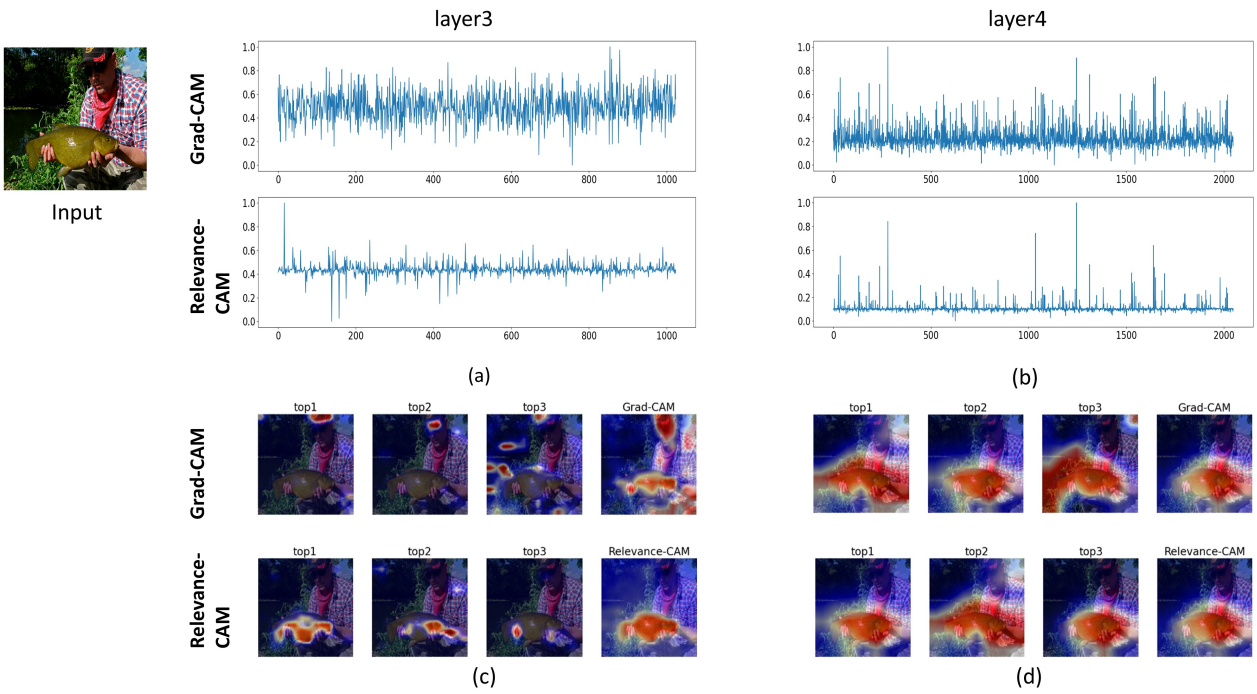


Figure 4. Evaluation of the selectivity. First column and second column show the results for layer3 and layer4 respectively. (a), (b) show the weighting component for each method. (c), (d) show the activation maps of the top 3 weighted channel for each method and the generated attention maps at the end.

| | Layer2 | | Layer4 | |
|---|---|---|---|---|
| Method | A.D. | A.I. | A.D. | A.I. |
| Grad-CAM | 74.91 | 4.45 | 23.13 | 24.05 |
| Grad-CAM++ | 71.15 | 4.85 | 22.03 | 25.35 |
| Score-CAM | 56.59 | 8.8 | 21.89 | 24.65 |
| Relelvance-CAM | **39.02** | **16.6** | **21.53** | **25.7** |

Table 1. Lower Average Drop(A.D.) and higher Average Increase(A.I.) indicate better performance. Evaluation for ResNet 50. Layer 2 is the low level layer and, Layer 4 is the last convolutional layer.

| | layer23 | | layer43 | |
|---|---|---|---|---|
| Method | A.D. | A.I. | A.D. | A.I. |
| Grad-CAM | 85.43 | 1.5 | 23.15 | 22.35 |
| Grad-CAM++ | 86.77 | 1.3 | 22.98 | 22.35 |
| Score-CAM | 76.11 | 3.2 | **21.22** | 23.5 |
| Relelvance-CAM | **72.25** | **3.9** | 22.42 | **24.95** |

Table 2. Lower Average Drop(A.D.) and higher Average Increase(A.I.) indicate better performance. Evaluation for VGG 16 with batch normalization. Layer 23 is the 3-th maxpooling layer and layer 43 is the 5-th and last maxpooling layer.

large, which shows that layer 2 of ResNet 50 also can extract class specific features.

Table 2, VGG 16, also shows similar aspects to Table 1. In deeper layers, the evaluation values are similar, and in the shallower layers, Relevance-CAM outperforms other methods. However, it should be noted that the differences in the evaluation values of Relevance-CAM between layers in VGG 16 is larger than that of ResNet 50. This shows that in the case of ResNet 50, class specific information can be extracted from the intermediate layers, whereas in VGG 16, the ability to extract class specific features from the intermediate layers are insufficient.

### 4.2. Evaluation for selectivity

In this section we visualize the weighting components to show how the shattered gradient problem damages the localization ability of Grad-CAM along with the top 3 weighted activation maps and to compare the selectivity of Grad-CAM to that of Relevance-CAM. The results of the two models in layer 3 and layer 4 are demonstrated for comparision.

In fig 4, (a), (b) shows the weighting component for each method along the channels. Here, we normalize the weighting components for a constant scale. The shattered gradient problem occurs at the gradient weight in layer 3. The gradient weights become noisy and flatten in layer 3 com-

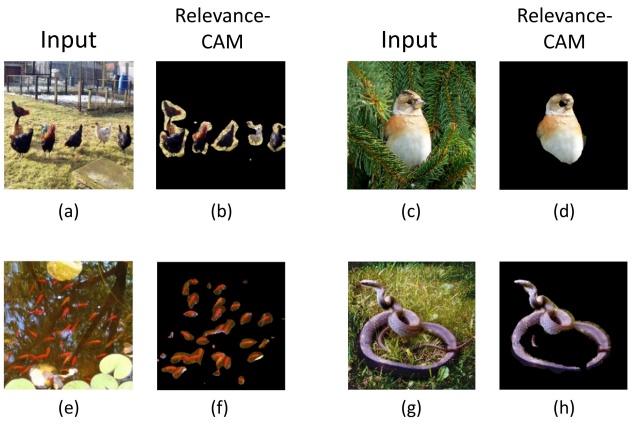


Figure 5. Visualization of weakly supervised localization. (a), (c), (e), and (f) are images for hen, brambling bird, gold fish and ring-neck snake, respectively. (b), (d), (f), and (h) are segmentation image through Relevance-CAM

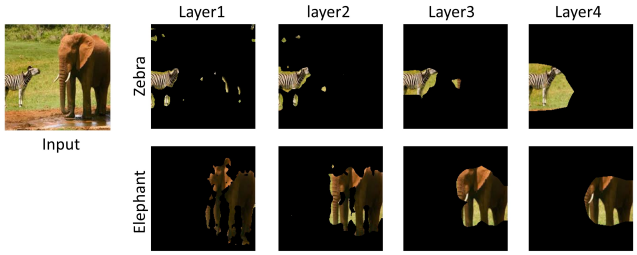


Figure 6. Class sensitivity test of Relevance-CAM on ResNet 50 for multi objects

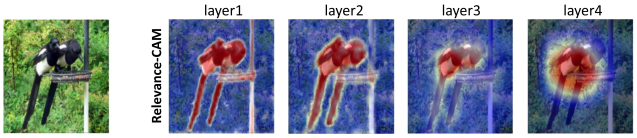


Figure 7. Class sensitivity test of Relevance-CAM on ResNet 50 for single object. The label of the input is a magpie.

pare to that of layer 4. On the other hand, the relevance weights prove its robustness to the shattered gradient problem, showing sparsity in both layer 3 and layer 4.

And we demonstrate the selectivity of Grad-CAM and Relevance-CAM in (c), (d) of fig 4 where the top 3 weighted activation maps and generated saliency maps of each method are displayed. The top 3 weighted activation maps of both Grad-CAM and Relevance-CAM localize the target class object, a tench, well in layer 4. But in layer 3, Grad-CAM shows a noisy heatmap. Even the most weighted activation map cannot extract the target class features. This means that the noisy weighting component of Grad-CAM cannot select the important feature maps. On the other hand, Relevance-CAM is clearly localizing the tench, and the top 3 weighted activation maps are also highlighting the target object in higher resolution. This indicates that Relevance-CAM provides good weights for the important channels. Through this demonstration, we show that Relevance-CAM is robust to gradient shattered problem and shows high selectivity in all layers.

### 4.3. Evaluation for Localization

The localization ability of attention map is important because the saliency map can be applied to localization tasks, such as the attention mechanism[8, 18, 29, 32] or the self erasing system[11, 35]. For this need, we evaluate the localization ability of Relevance-CAM in this section. ImageNet (ILSVRC2012) validation set and ResNet 50 pretrained on ImageNet classification task are used for this experiment. We conduct segmentation with Relevance-CAM using a mask, where its area consists of pixels of the generated saliency map higher than average + 1*standard deviation of the saliency map, on the segmented input image at the layer 2 of ResNet 50, in fig 5. Relevance-CAM separates the objects from the background well, and even small objects like goldfish are localized with high precision.

Next, we experiment the localization ability quantitatively. The performance is evaluated using Intersection over Union (IoU) metric, which is the ratio of the intersection area to union area between the pixels of generated attention map for a class and bounding box pixels. If the saliency map localizes the target object tightly, the value of IoU would be high. Experiment is conducted on the randomly selected 2000 images that the model predicted correctly, Table 3. Almost similar performance results from the last convolutional layer. But in the shallow layer, Relevance-CAM outperforms the other methods. And it should be noted that Relevance-CAM has fewer changes in IoU even when the target layer becomes shallower.

### 4.4. Class Sensitivity Test

In fig 6, we visualize the masked image for each class to qualitatively test the class sensitivity of our proposed method. First row shows the masked images for class 'zebra' along different layer depths. Second row shows the masked images for class 'elephant' along different layer depths. Here, we use the identical mask described in sec 4.3

The masked images show amazing class sensitivities. As the layer becomes shallower, it can be seen that the localization performance is slightly degraded but still creates sufficiently class sensitive high resolution heatmaps. It should be noted that the model already separates the elephant and the zebra from the layer 1.

Another phenomenon can be seen in fig 7 - where only the magpie is localized in layer 3 and 4, and the bar next to the magpie starts to be localized in layer 2. In layer 1, the

| | layer 1 | layer 2 | layer 3 | layer 4 |
|---|---|---|---|---|
| Grad-CAM | 0.12 | 0.18 | 0.22 | 0.34 |
| Grad-CAM++ | 0.13 | 0.19 | 0.22 | 0.34 |
| Score-CAM | 0.21 | 0.25 | 0.28 | 0.34 |
| Relevance-CAM | **0.30** | **0.32** | **0.32** | 0.34 |

Table 3. IoU of the various explanation methods along the layer depth of ResNet 50

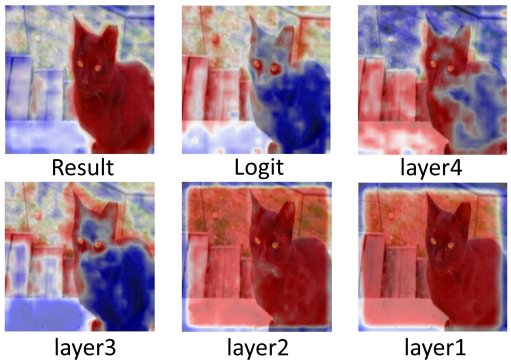


Figure 8. Sanity Check for Relevance-CAM of layer 2 of ResNet50 model

magpie and the rod on which the magpie are sitting are localized together very clearly. Recognizing the bar together at the low level layer can be understood as localizing the rods similar to the branches, which the magpies in the imagenet dataset are observed to sit on often. In other words, magpie and rod are considered as the same class in the shallow layer, and as the layer deepens, the magpie and rod are separately extracted as different features.

To summarize the findings in fig 6 and fig 7. Relevance-CAM creates a class specific heatmap on layer 1 of ResNet 50. Through this, it can be seen that not only general features or local features are extracted from layer 1, but also class specific information is extracted. In addition, as the layer deepens, the channel of the activation map increases, and accordingly, the features are further subdivided within the scope initially extracted from the shallow layers.

### 4.5. Sanity check for Relevance-CAM

We evaluate our method with cascading randomization test which is proposed in [2]. The experiment is a very important work when it comes to the explainable attention map. Fig 8 is Relevance-CAM results for layer 2 of ResNet50 model obtained by progressively randomizing the parameters from logit to layer 1. As we can see, the saliency map is destroyed with the parameter randomization. Thus, our method is sensitive to model parameters.

## 5. Evidence that class specific information is extracted from shallow layers

Someone can raise a question for the argument that the class specific information is extracted from shallow layers. When obtaining Relevance-CAM, the forward path information of shallower layers and the backward path information of the deeper layers are required. However, since Relevance-CAM uses LRP only for channel-wise weights, there is no effect on spatial-wise weights. For example, if the shallow layer only extracts general features such as edge or texture, there would be no feature map that localizes only objects of a particular class. In that situation no matter the channel-wise weighting done through LRP, we would not be able to create a class sensitive attention map. In fig 6 of our paper, Relevance-CAM results of ResNet50 show that objects of different classes are localized separately even in shallow layers. We can infer that there are feature maps that separately localize objects of different classes and those feature maps have a great influence on the class output score. However, Relevance-CAM results of deeper layers can extract class specific features more explicitly than that of the shallow layer as shown in fig 7. Thus, we argue that class specific features can be extracted from the shallow layers, but higher level features are extracted as the layer deepens.

## 6. Conclusion

In this paper, we proposed a novel Class Activation Mapping method called Relevance-CAM for faithful and accurate explanation of deep learning models and its layers of various depths. Our proposed Relevance-CAM is robust to the problems other explanation methods share, such as the shattered gradient problem and False Confidence. Due to these advantages, Our Relevance-CAM enables analysis of shallow layers, and we find that class specific features can be extracted even in shallow layers which have small receptive fields. This insight can be used in various fields such as transfer learning [22, 28, 23], model pruning[31], and weakly supervised segmentation [35, 17]. As an example, in the case of transfer learning, rather than selecting layers to be fine-tuned empirically, it is possible to select a layer through layer-wise analysis. We believe that our proposed Relevance-CAM allows other researchers to deepen their analysis of deep learning models.

## Acknowledgements

This work was supported by Samsung Research Funding Center of Samsung Electronics under Project Number SRFC-IT1901-08.

## References


[1] Kjersti Aas, Martin Jullum, and Anders Løland. Explaining individual predictions when features are dependent: More accurate approximations to shapley values. *arXiv preprint arXiv:1903.10464*, 2019.

[2] Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. Sanity checks for saliency maps. In *Advances in Neural Information Processing Systems*, pages 9505–9515, 2018.

[3] Leila Arras, José Arjona-Medina, Michael Widrich, Grégoire Montavon, Michael Gillhofer, Klaus-Robert Müller, Sepp Hochreiter, and Wojciech Samek. Explaining and interpreting lstms. In *Explainable ai: Interpreting, explaining and visualizing deep learning*, pages 211–238. Springer, 2019.

[4] Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller, and Wojciech Samek. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PloS one*, 10(7):e0130140, 2015.

[5] David Balduzzi, Marcus Frean, Lennox Leary, JP Lewis, Kurt Wan-Duo Ma, and Brian McWilliams. The shattered gradients problem: If resnets are the answer, then what is the question? *arXiv preprint arXiv:1702.08591*, 2017.

[6] Aditya Chattopadhay, Anirban Sarkar, Prantik Howlader, and Vineeth N Balasubramanian. Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 839–847. IEEE, 2018.

[7] Chi-Tung Cheng, Tsung-Ying Ho, Tao-Yi Lee, Chih-Chen Chang, Ching-Cheng Chou, Chih-Chi Chen, I-Fang Chung, and Chien-Hung Liao. Application of a deep learning algorithm for detection and visualization of hip fractures on plain pelvic radiographs. *European radiology*, 29(10):5469–5477, 2019.

[8] Hiroshi Fukui, Tsubasa Hirakawa, Takayoshi Yamashita, and Hironobu Fujiyoshi. Attention branch network: Learning of attention mechanism for visual explanation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[9] Jindong Gu, Yinchong Yang, and Volker Tresp. Understanding individual decisions of cnns via contrastive backpropagation. In *Asian Conference on Computer Vision*, pages 119–134. Springer, 2018.

[10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[11] Qibin Hou, PengTao Jiang, Yunchao Wei, and Ming-Ming Cheng. Self-erasing network for integral object attention. In *Advances in Neural Information Processing Systems*, pages 549–559, 2018.

[12] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141, 2018.

[13] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017.

[14] Brian Kenji Iwana, Ryohei Kuroki, and Seiichi Uchida. Explaining convolutional neural networks using softmax gradient layer-wise relevance propagation. In *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, pages 4176–4185. IEEE, 2019.

[15] Yohan Jun, Taejoon Eo, Taeseong Kim, Hyungseob Shin, Dosik Hwang, So Hi Bae, Yae Won Park, Ho-Joon Lee, Byoung Wook Choi, and Sung Soo Ahn. Deep-learned 3d black-blood imaging using automatic labelling technique and 3d convolutional neural networks for detecting metastatic brain tumors. *Scientific reports*, 8(1):1–11, 2018.

[16] Beomsu Kim, Junghoon Seo, Seunghyeon Jeon, Jamyoung Koo, Jeongyeol Choe, and Taegyun Jeon. Why are saliency maps noisy? cause of and solution to noisy saliency maps. In *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, pages 4149–4157. IEEE, 2019.

[17] Jungbeom Lee, Eunji Kim, Sungmin Lee, Jangho Lee, and Sungroh Yoon. Ficklenet: Weakly and semi-supervised semantic image segmentation using stochastic inference. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5267–5276, 2019.

[18] Kunpeng Li, Ziyan Wu, Kuan-Chuan Peng, Jan Ernst, and Yun Fu. Guided attention inference network. *IEEE transactions on pattern analysis and machine intelligence*, 2019.

[19] Grégoire Montavon, Sebastian Lapuschkin, Alexander Binder, Wojciech Samek, and Klaus-Robert Müller. Explaining nonlinear classification decisions with deep taylor decomposition. *Pattern Recognition*, 65:211–222, 2017.

[20] Grégoire Montavon, Wojciech Samek, and Klaus-Robert Müller. Methods for interpreting and understanding deep neural networks. *Digital Signal Processing*, 73:1–15, 2018.

[21] Woo-Jeoung Nam, Shir Gur, Jaesik Choi, Lior Wolf, and Seong-Whan Lee. Relative attributing propagation: Interpreting the comparative contributions of individual units in deep neural networks. In *AAAI*, pages 2501–2508, 2020.

[22] Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10):1345–1359, 2009.

[23] Rajat Raina, Alexis Battle, Honglak Lee, Benjamin Packer, and Andrew Y Ng. Self-taught learning: transfer learning from unlabeled data. In *Proceedings of the 24th international conference on Machine learning*, pages 759–766, 2007.

[24] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144, 2016.

[25] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017.

[26] Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. Learning important features through propagating activation differences. *arXiv preprint arXiv:1704.02685*, 2017.

[27] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*, 2013.

[28] Chuanqi Tan, Fuchun Sun, Tao Kong, Wenchang Zhang, Chao Yang, and Chunfang Liu. A survey on deep transfer learning. In *International conference on artificial neural networks*, pages 270–279. Springer, 2018.

[29] Fei Wang, Mengqing Jiang, Chen Qian, Shuo Yang, Cheng Li, Honggang Zhang, Xiaogang Wang, and Xiaoou Tang. Residual attention network for image classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3156–3164, 2017.

[30] Haofan Wang, Zifan Wang, Mengnan Du, Fan Yang, Zijian Zhang, Sirui Ding, Piotr Mardziel, and Xia Hu. Score-cam: Score-weighted visual explanations for convolutional neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 24–25, 2020.

[31] Kafeng Wang, Xitong Gao, Yiren Zhao, Xingjian Li, Dejing Dou, and Cheng-Zhong Xu. Pay attention to features, transfer learn faster cnns. In *International Conference on Learning Representations*, 2019.

[32] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. Cbam: Convolutional block attention module. In *Proceedings of the European conference on computer vision (ECCV)*, pages 3–19, 2018.

[33] Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *European conference on computer vision*, pages 818–833. Springer, 2014.

[34] Jianming Zhang, Sarah Adel Bargal, Zhe Lin, Jonathan Brandt, Xiaohui Shen, and Stan Sclaroff. Top-down neural attention by excitation backprop. *International Journal of Computer Vision*, 126(10):1084–1102, 2018.

[35] Xiaolin Zhang, Yunchao Wei, Jiashi Feng, Yi Yang, and Thomas S Huang. Adversarial complementary learning for weakly supervised object localization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1325–1334, 2018.

[36] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2921–2929, 2016.