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

1. Comparison with Score-CAM

With the proposition of the shattered gradient problem, we are not the only ones who have doubts about making CAM with gradient weights. Score-CAM[] gets rid of the gradient shattered problem through obtaining weights by performing a perturbation on its input image with a feature map, rather than by backpropagating the class specific information from logit. This is intuitive and effectively eliminates the shattered gradient problem. But the operating time increases in proportion to the channel of the target layer, which is usually more than 500 times slower than our proposed Relevance-CAM because Score-CAM can be obtained by propagation as many as the number of channels of the target layer while Relevance-CAM can be obtained with 1 propagation and 1 backpropagation, Table 1. For applications in various tasks such as attention mechanism or weakly supervised localization, the operating time is crutial.

Looking at the evaluation tables of main paper, it can be seen that even though Score-CAM is not affected by the gradient shattered problem, the faithfulness and localization performance decrease as the layer becomes shallower. This results from the process of Score-CAM calculating the weighting component. Score-CAM is obtained by Hadamard Product of the normalized activation map to the input in order to measure the importance of the activation map. The weighting components obtained by this way are the local importance calculated only using one activation map. But, in the case of Relevance-CAM, the weighting component is obtained by equation 9, including the value of the other activation maps. Therefore, since Relevance-CAM uses the overall information of the layer output, the weights can be assigned appropriately. Therefore, Relevance-CAM shows clear localization even in shallow layers.

2. Comparision between CAM-based methods and LRP

Unlike CAM based methods, LRP is not derived by combining feature maps of the convolutional layer, but rather computed by assigning the relevance score on each pixel.

*Corresponding author.

Method	Score-CAM	Relevance-CAM
Time	16.67s	0.031s

Table 1. Average Time for generating saliency map by Score-CAM and Relevance-CAM. We use the last convolutional layer of Resnet50 with GPU, Titan X pascal

Model	Thres -hold	Grad-CAM	R-CAM	CLRP
ResNet152 (maxpool3)	25%	0.28	**0.42**	0.39
	75%	0.44	**0.52**	0.49
GoogleNet (maxpool3)	25%	0.26	**0.38**	0.37
	75%	0.47	**0.51**	0.47

In other words, while LRP intervenes in spatial information of attention map, CAM based methods only combinate the feature maps which is extracted from the layer without any spatial effects. Therefore, CAM based methods are suitable for analyzing the feature extraction capability of a particular layer. But from a localization perspective, LRP should also be compared. Therefore, we include the performance of CLRP in view of IoU. As the table 2 shows. our Relevance-CAM outperforms other methods in localization evaluation.

3. Weakly supervised segmentation

We also include results of various attention map and segmentation map in which we use same threshold value with section 4.3. We use ResNet50 model and extract attention map at layer 2 of the model.