Supplementary Material

Kwang Hee Lee[1,*,**], Chaewon Park[1,*], Junghyun Oh[1,2,*] and Nojun Kwak[2]

[1]Boeing Korea Engineering and Technology Center(BKETC)
[2]Seoul National University

1. Network Architecture

LFI-CAM is composed of the attention branch and perception branch and both branches are connected by the attention mechanism. Conventional baseline models such as ResNet [1], DenseNet [3], ResNeXt [4] and SENet [2] or the customized classifier can be used as the perception branch. The Feature Importance Network(FIN) contains multiple convolution layers for extracting feature importance of the feature map. The architecture details of FIN are shown in Table 1. Notations are as follows: h and w: height and width of the input image, N: the number of output channels, K: kernel size, S: stride size, P: padding size, BN: batch normalization.

2. Additional Experimental Results

2.1. Visual Explanation Results

In addition to the results presented in the paper, we show supplementary visual explanation results of ABN and LFI-CAM for CIFAR100, STL10 and ImageNet in Figs. 2, Figs. 3 and Figs. 4

2.2. Visualization of the Feature Importance Network Effectiveness

To evaluate effectiveness of the proposed Feature Importance Network, we visualize the pixel-wise mean feature map from the last convolutional layer of the LFI-CAM model trained without and with the FIN. Then we compare them against the CAM generated from LFI-CAM model trained with FIN. Fig. 1 shows the additional visual explanation results for FIN effectiveness. After the FIN's feature importance is incorporated, our $L_{LFI-CAM}$ successfully focuses on the most distinguishable region of the target object. For example, as shown in the first and third row, the attention focuses more on the desk area after applying FIN because LFI-CAM classifies the input image as 'desk'.

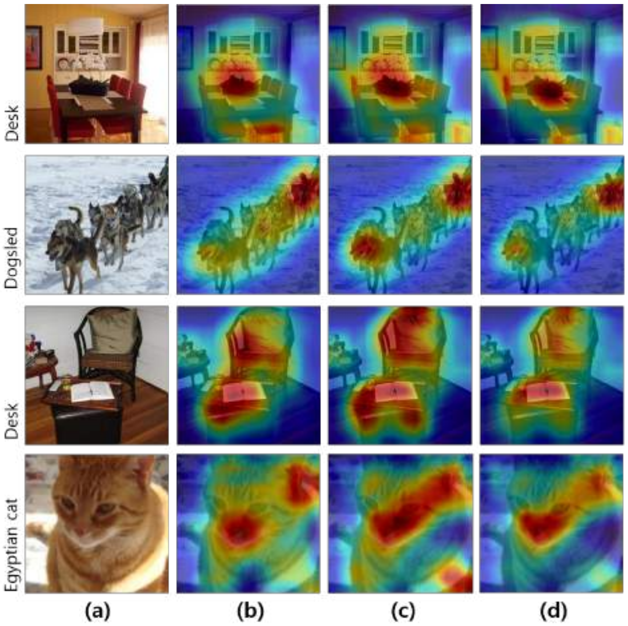

Figure 1. Visualization of the Feature Importance Network Effectiveness. (a) Input image, (b) Pixel-wise mean feature map from the last convolutional layer of LFI-CAM trained without FIN, (c) Pixel-wise mean feature map from the last convolutional layer of LFI-CAM trained with FIN, (d) CAM generated from LFI-CAM.

2.3. Stability Evaluation on Visual Explanation for ABN and LFI-CAM

We have observed that ABN outputs unreliable and inconsistent attention maps through several experiments. We trained several ABN models with various hyper-parameters on the Cat&Dog and STL10, and then compared CAMs of the same image from several models with similar accuracy. Fig. 5 and Fig. 6 show the additional stability test results on visual explanation for the STL10 and Cat&Dog. CAM results for the exactly same test images are unreliable and inconsistent although the trained ABN models have similar

[*] indicates equal contribution.
[**] indicates corresponding author.

accuracy. On the other hand, the results of LFI-CAM can be confirmed to be reliable and stable.

References

[1] 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. 1

[2] 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. 1

[3] 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. 1

[4] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1492–1500, 2017. 1

Table 1. Architecture of Feature Importance Network (FIN).

Part	Input → Output Shape	Layer Information
CONV Layer	$(\frac{h}{16}, \frac{w}{16}, 2048) \rightarrow (\frac{h}{16}, \frac{w}{16}, 2048)$	CONV-(N2048, K3, S0, P1), BN, ReLU
	$(\frac{h}{16}, \frac{w}{16}, 2048) \rightarrow (\frac{h}{16}, \frac{w}{16}, 2048)$	CONV-(N2048, K3, S0, P1), BN, ReLU
	$(\frac{h}{16}, \frac{w}{16}, 2048) \rightarrow (\frac{h}{16}, \frac{w}{16}, 2048)$	CONV-(N2048, K3, S0, P1), BN, ReLU
	$(\frac{h}{16}, \frac{w}{16}, 2048) \rightarrow (\frac{h}{16}, \frac{w}{16}, 2048)$	CONV-(N2048, K3, S0, P1), BN, ReLU
Output Layer	$(\frac{h}{16}, \frac{w}{16}, 2048) \rightarrow (2048)$	Global Average Pooling & SoftMax

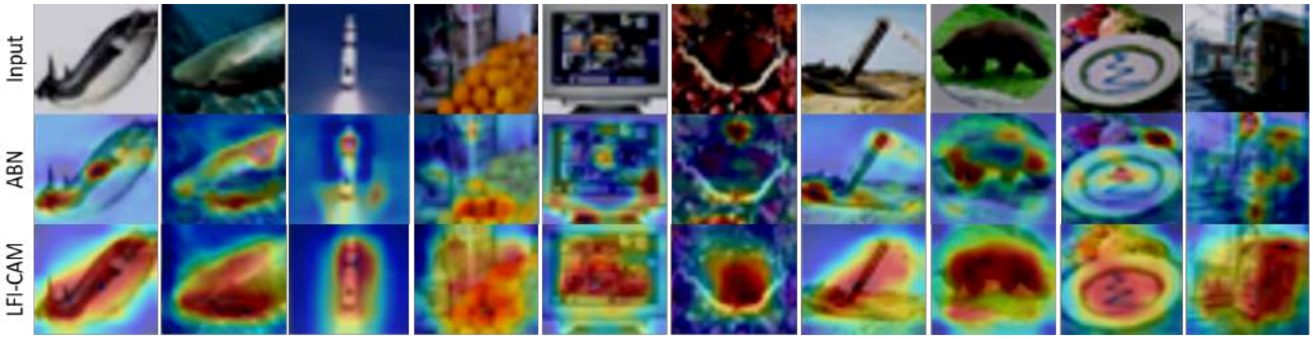

Figure 2. Visual Explanation Results of ABN and LFI-CAM for CIFAR100

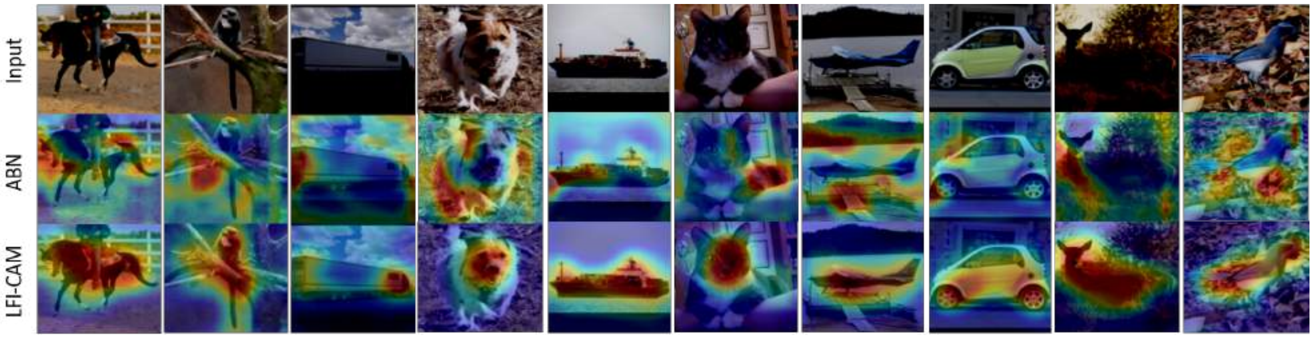

Figure 3. Visual Explanation Results of ABN and LFI-CAM for STL10

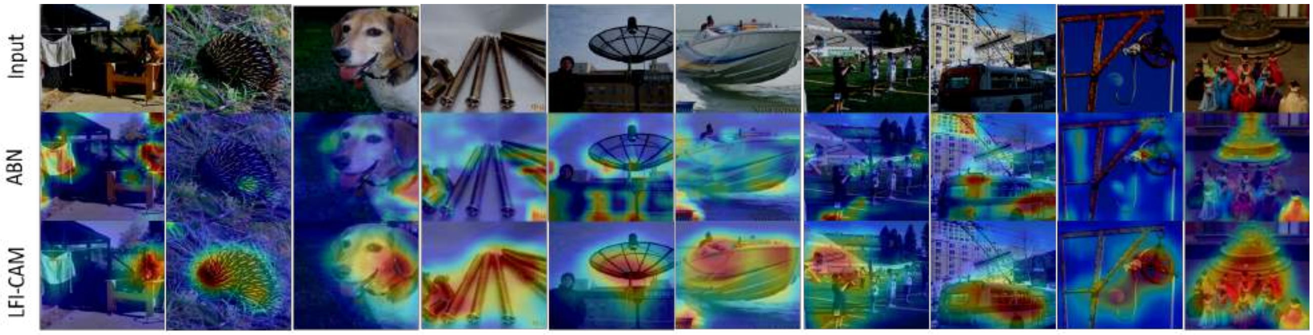

Figure 4. Visual Explanation Results of ABN and LFI-CAM for ImageNet

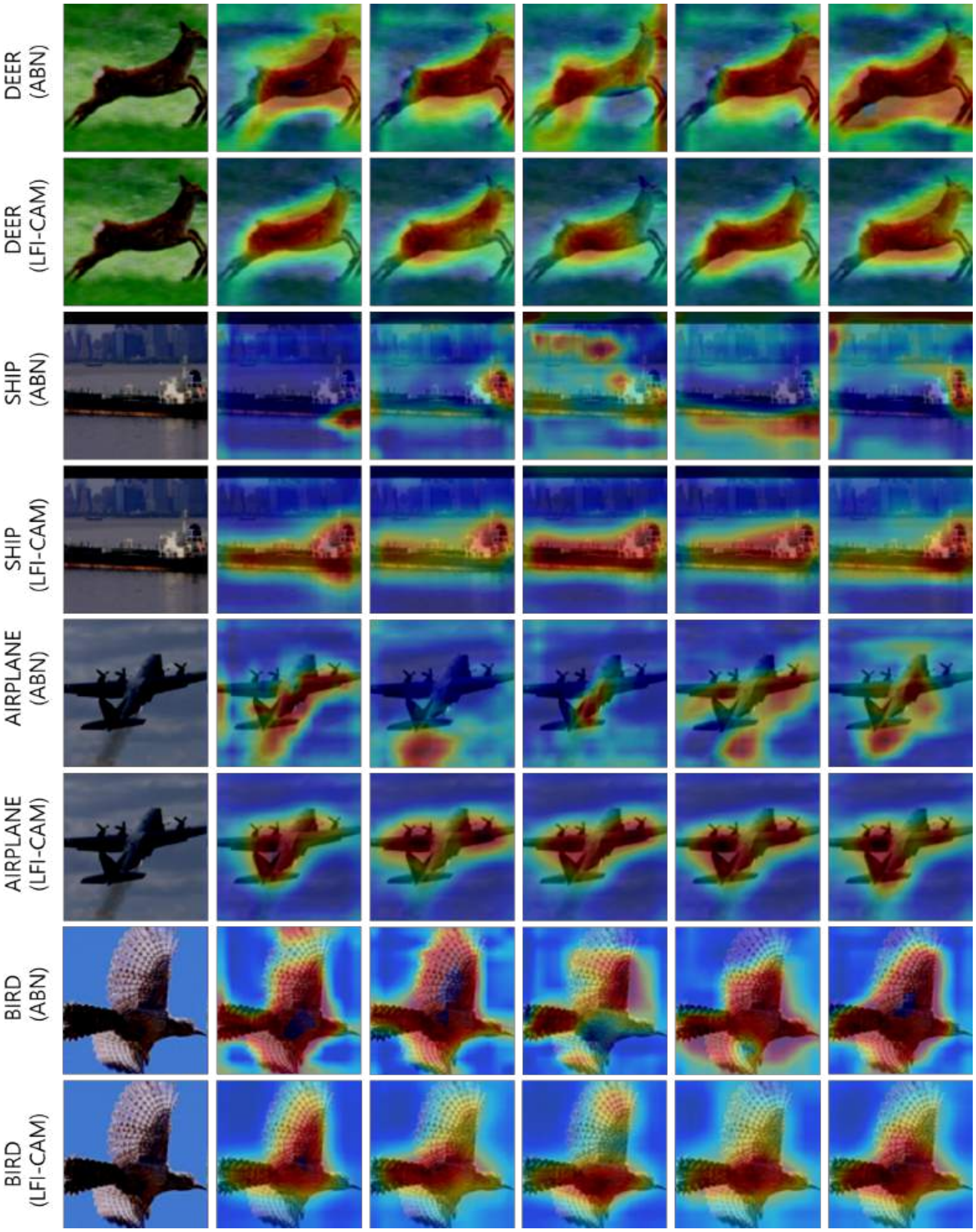

Figure 5. Examples of stability test on visual explanation. Each row displays CAM results of ABN or LFI-CAM models that were trained with various (5) hyper-parameters on the STL10 dataset.

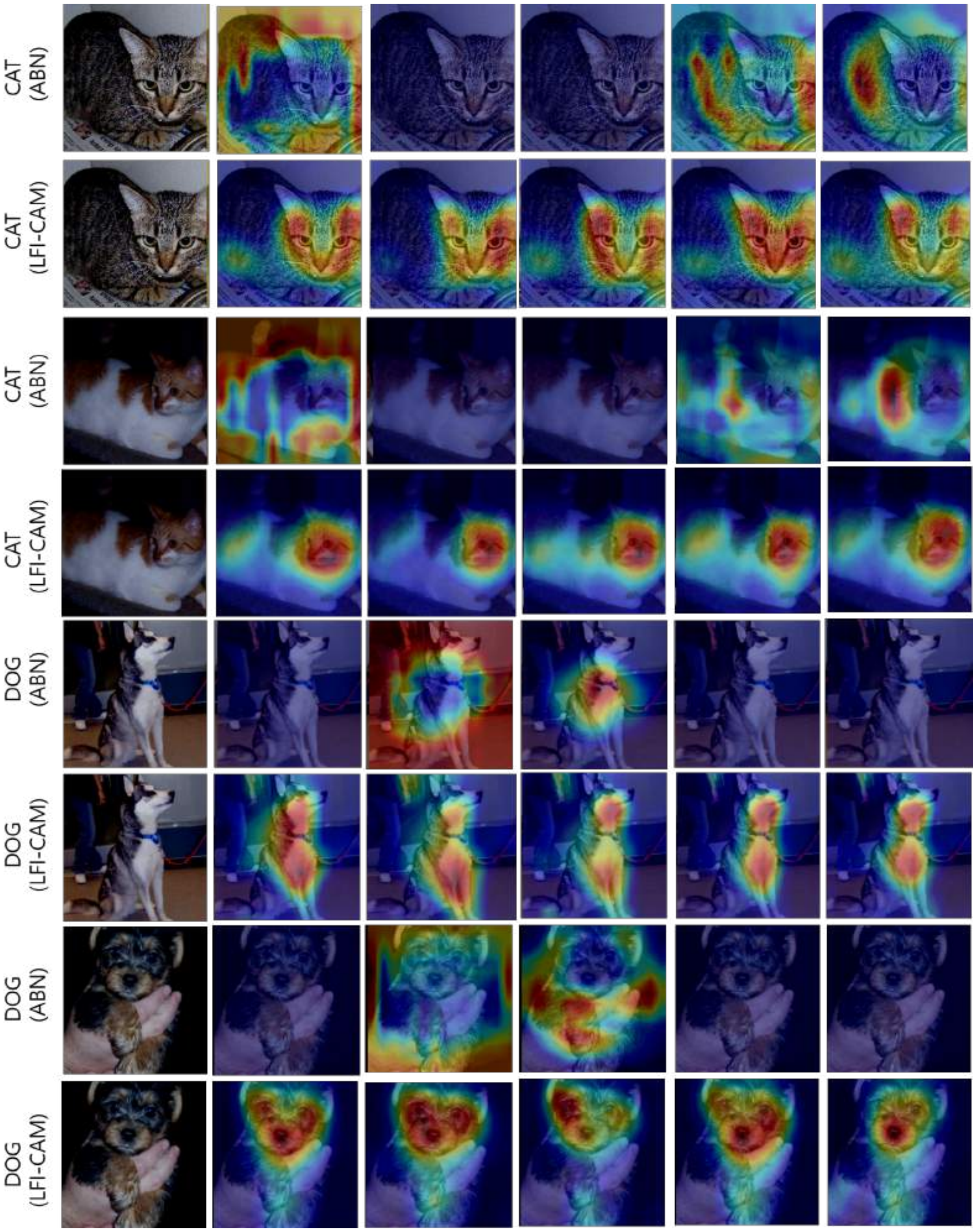

Figure 6. Examples of stability test on visual explanation. Each row displays CAM results of ABN or LFI-CAM models that were trained with various (5) hyper-parameters on the Cat&Dog dataset.