JCSE Journal of Computing Science and Engineering

Regular Paper

Journal of Computing Science and Engineering,
Vol. 18, No. 1, March 2024, pp. 19-28

Improving Interpretability of Deep Neural Networks in Medical Diagnosis by Investigating the Individual Units

Ho Kyung Shin and Woo-Jeoung Nam*
School of Computer Science and Engineering, Kyungpook National University, Daegu, Korea
parkland106e@gmail.com, nwj0612@knu.ac.kr

Abstract
Interpretability has emerged as an obstacle to the adoption of deep neural networks (DNNs) in particular domains, which has led to increasing interest in addressing transparency issues to ensure that DNNs can fulfill their impressive potential. In the current paper, we demonstrate the efficiency of various attribution techniques to explain the diagnostic decision of DNNs by visualizing the predicted suspicious region in the image. By utilizing the characteristics of objectness that DNNs have learned, fully decomposing the network prediction enables precise visualization of the targeted lesion. To verify our work, we conduct experiments on chest X-ray diagnosis using publicly accessible datasets. As an intuitive assessment metric for explanations, we present the performance of the intersection of union between visual explanation and the bounding box of lesions. The experimental results show that recently proposed attribution methods can visualize more specific localizations for diagnostic decisions compared to the traditionally used class activation mapping. We also analyze the inconsistency of intentions between humans and DNNs, which is easily obscured by high performance. Visualizing the relevant factors makes it possible to confirm that the criterion for decision is consistent with the training strategy. Our analysis of unmasking machine intelligence demonstrates the need for explainability in medical diagnostic decision-making.

Category: Information Retrieval / Web

Keywords: Deep learning; Explainable computer-aided diagnosis; Explainable AI; Visual explanation; Medical imaging analysis

I. INTRODUCTION

Deep neural networks (DNNs) currently play an important role in improving empirical performance in various computer vision tasks such as image classification [1, 2], object detection [3], human action recognition [4–7], and medical diagnosis [8, 9]. However, a lack of interpretability hinders the applicability of many DNN models in mission-critical systems including medical diagnosis, military, and finance. Despite the remarkable successes that have been achieved in computer-aided detection (CADe) and computer-aided diagnosis (CADx) [10–12], their adoption in the real world remains constrained due to the ambiguity involved in understanding diagnostic decisions. There have been many studies aiming to overcome this limitation by addressing the lack of transparency in DNNs.

In explaining the decisions made by DNNs as a process of decomposition, the contributions of individual neurons are propagated backward through the weights, thereby

Open Access http://dx.doi.org/10.5626/JCSE.2024.18.1.00 http://jcse.kiise.org

Received 06 February 2024; **Accepted** 12 March 2024
*Corresponding Author

 pISSN: 1976-4677 eISSN: 2093-8020

resulting in a redistribution of relevance in the pixel space. The results of sensitivity analysis [13] imply the factors that reduce or increase the evidence for the predicted results. Layer-wise relevance propagation (LRP) [14] is a method for decomposing the output prediction by fully redistributing the relevance throughout the layers. Deep Taylor decomposition [15] is a theoretical extension of LRP that can help interpret the decision by utilizing the Taylor expansion and root point concept. DeepLIFT [16] propagates the differences in contribution scores between the activated neurons and their reference activation. The recently proposed relative attributing propagation (RAP) [17] is a method for decomposing the positive (relevant) and negative (irrelevant) relevance to each neuron, according to its relative influence among the neurons. Changing the perspective from value to influence shows a clear distinction between relevant/irrelevant attributions with a high objectness score. Relative sectional propagation [18] aims to decompose the output predictions of DNNs with class-discriminativeness by setting the hostile activations of neurons with respect to the target class.

The visualization of disease aids radiologists, such as in the form of class activation mapping (CAM) [19], which generates CAMs to highlight the discriminative regions, and this is widely used in the medical domain to localize diverse diseases. Despite its advantages of simple implementation and class discriminations, this method has a limitation in its ability to precisely localize predictions due to the broad area of visualization that results from expanding compressed feature maps. To localize the lesion areas more clearly, image-to-image translation methods based on the generative model have been studied in the medical field [20–23] by visualizing the different factors of inter-domains while maintaining the original subject. Visual attribution generative adversarial networks (VA-GAN) [24] and fixed-point GAN [21] synthesize Alzheimer's disease images using 3D brain magnetic resonance imaging. However, the GAN network also faces limitations in that it itself is not completely explainable, and that it cannot guarantee generality in fields without sufficient localization ground truths of lesion areas. To resolve this problem, we address the efficiency of recent attribution techniques to contribute to the interpretation of diagnostic decisions. These algorithmic approaches investigate the trained network itself without the need for any additional supervision. Fig. 1 shows intuitive examples of the visual explanations used in this work. The main contributions of this work are as follows:

- We demonstrate the methods that can be efficiently used to interpret the diagnostic decisions made by DNNs by utilizing visual explaining techniques: LRP and RAP, which are adaptable to any fully trained networks. Without requiring any supervision of the area of lesions, decomposing the predictions of classification networks makes it possible to precisely localize and illustrate the crucial factors for their diagnosis. Our experiments using chest X-ray datasets show that the recent attribution techniques provide a more sufficient guarantee of performance compared to the existing widely used CAM.
- We demonstrate the experiment of the inconsistency between human intention and DNNs by training binary classification tasks: normal or pneumonia. We utilize general techniques in machine learning fields, and the networks exhibit proper performance. However, the visual explanation shows misalignment between the direction of the training strategy and the actual criteria of the classification. We analyze the phenomenon of inconsistency and add another voice raising the necessity of interpretability.

II. ATTRIBUTION METHODS

In this section, we introduce notations and attribution methods: LRP and RAP, which are closely related to each other but have different perspectives and algorithms. Fig. 2 provides an overview of decomposition and visualization. For input x, we denote the letter $f(x)$ as the value of the network output before it passes through the classification layer, such as the sigmoid and softmax layers. R represents the input relevance for the attributing procedure, which is the same as the value of $f(x)$ of the prediction node. $w_{ij}^{(l,l+1)}$, $b_j^{(l,l+1)}$, and $a(\cdot)$ denote the weight, bias, and activation function between layer $l, l+1$, respectively. $m_j^{(l+1)}$ is the value of the neuron after

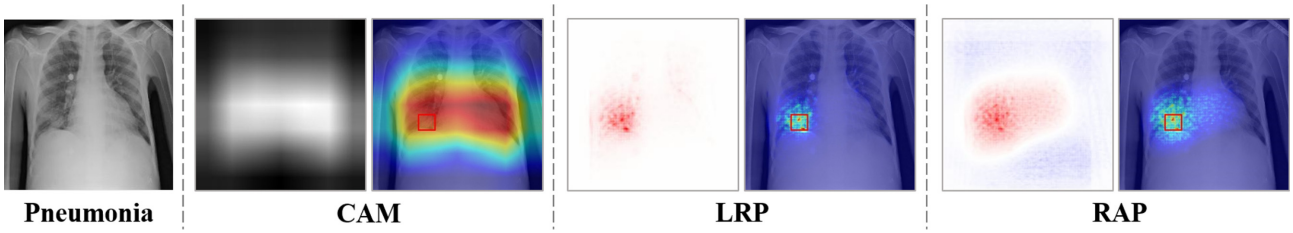

Fig. 1. Intuitive comparison of CAM, LRP, and RAP. For each method, the left and right images show the relevance heatmap and visual explanation, respectively. Red: high relevance, Blue: low relevance.

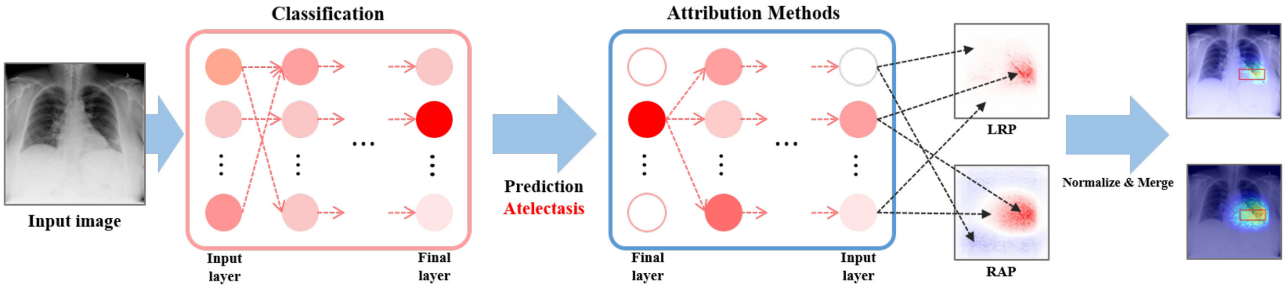

Fig. 2. Overview of the visual explanation procedure. Attribution methods: LRP and RAP are the decomposing procedures in a backward pass after the network is fully trained. Relevance, corresponding to output prediction, is propagated through trained weights and activated neurons.

applying the activation function. The signs of positive and negative values are respectively denoted by + and −.

A. Layer-wise Relevance Propagation

The principle of LRP [14] is to find the parts with high relevance in the input by propagating the result from the back (output) to the front (input). The algorithm is based on the conservation principle, which maintains relevance in all layers: from input to output.

$$\sum_i R_i^{(l)} = \sum_j R_j^{(l+1)}. \tag{1}$$

Among the various LRP versions introduced in [14], we utilize LRP-$\alpha\beta$, which separates the positive and negative activations during the relevance propagation process while maintaining the conservation rule (1).

$$R_i^{(l)} = \sum_j \left(\alpha \cdot \frac{z_{ij}^+}{\sum_i z_{ij}^+} - \beta \cdot \frac{z_{ij}^-}{\sum_i z_{ij}^-} \right) R_j^{(l+1)}. \tag{2}$$

In this rule, $z_{ij}^+ + z_{ij}^- = z_{ij}$ and $\alpha - \beta = 1$. The propagated attributions are allocated to the pixels of the input image in a manner that indicates their relevance to output prediction. In this paper, the function parameters are set as $\alpha = 1, \beta = 0$.

B. Relative Attributing Propagation

RAP [17] decomposes the output predictions of DNNs in terms of relative influence among the neurons, resulting in the relevant and irrelevant attributions being assigned with a bi-polar importance. By changing the perspective from value to influence, the generated visual explanations show the characteristics of strong objectness along with a clear distinction between relevant and irrelevant attributions. The algorithm has three main steps: (i) absolute influence normalization, (ii) decisions regarding the criterion of relevance and propagating in a backward pass, and (iii) uniform shifting to change the irrelevant neurons to negative.

Absolute influence normalization is the process that is only applied in only the first backward propagation for changing the perspective on the neuron from the value to the influence. From the output prediction node j in layer q, the relevance is allocated into the penultimate layer p according to its actual contribution in a forward pass.

$$R_i^{(p)} = \left(\sum_i z_{ij}^+ + \sum_i z_{ij}^- \right) * R_j^{(q)}. \tag{3}$$

For an approach from an influence perspective, the positive or negative relevance values allocated in the penultimate layer are normalized by the ratio of the absolute positive and negative values $\left|R_i^{(p)+}\right| : \left|R_i^{(p)-}\right|$.

$$R_i'^{(p)} = |R_i^{(p)}| * \frac{\sum_i R_i^{(p)}}{\sum_i |R_i^{(p)}|}. \tag{4}$$

This process makes it so that the neurons are allocated in terms of their relative importance to the output prediction, from high influence to low influence. For the next steps, i.e., the attributing procedure from the penultimate layer to the input layer, Eqs. (5) and (6) are repeated in each layer while changing low influence neurons to have negative relevance.

$$\bar{R}_i^{(l)} = \sum_j \left(\frac{z_{ij}^+}{\sum_i \left(z_{ij}^+\right)} R_j^{(l+1)} + \frac{z_{ij}^-}{\sum_i \left(z_{ij}^-\right)} \left(R_j^{(l+1)} * \frac{\sum_i |z_{ij}^-|}{\sum_i \left(|z_{ij}^+| + |z_{ij}^-|\right)} \right) \right), \tag{5}$$

$$\Psi_i^l = \begin{cases} \sum_i \left(\bar{R}_{i \in \mathcal{N}}^{(l)} \right) * \frac{1}{\Gamma}, & m_i^{(l)} \text{ is activated} \\ 0, & \text{otherwise} \end{cases}, \tag{6}$$

$$R_i^{(l)} = \bar{R}_i^{(l)} - \Psi_i^l$$

Here, Γ is the number of activated neurons in each

layer, and $\overline{R}_{i\in N}^{(l)}$ denotes the relevance propagated through the negative weights, i.e., the latter parts of Eq. (5). This procedure makes it possible to assign relatively irrelevant units as negative while emphasizing the important factors as highly positive. RAP also preserves the conservation rule (1).

III. EXPERIMENTAL EVALUATION

A. Data

1) NIH ChestX-ray14

NIH ChestX-ray14 dataset [25] comprises 112,120 frontal-view chest X-ray images in PNG format that have been collected from unique 30,805 patients with corresponding to 14 disease labels (atelectasis, consolidation, infiltration, pneumothorax, edema, emphysema, fibrosis, effusion, pneumonia, pleural thickening, cardiomegaly, nodule, mass, and hernia), each image can be labeled with one or more of these classes. Images where none of these diseases are detected are labeled as no finding. We utilize CheXNet [26], which is based on DenseNet [27] and widely used in radiologist-level chest X-ray analysis. This trained model is publicly available with verified performance. The average area under receiver operating characteristic (AUROC) of this model for 14 classes is 0.843 whereas the AUROC of eight classes that are annotated with bounding box labels, i.e., atelectasis, cardiomegaly, effusion, infiltration, mass nodule, pneumonia, pneumothorax, are presented in Table 1. We utilize 984 images, which are annotated with the bounding box, to evaluate the visual explanations for the target lesions.

2) RSNA Pneumonia Detection

The RSNA Pneumonia Detection Challenge dataset [28] is a subset of 30,000 images from the NIH ChestX-ray14 dataset that is labeled with two classes: normal and pneumonia. The original purpose of this challenge is to detect pneumonia lesions. We exclude the test data that does not have a label and separate the training dataset into training and validation sets in a respective ratio of 9:1. We trained VGG-16 [29], ResNet-50 [2] and DenseNet-121 networks, which have been successfully established to have impressive performance in the machine learning field. While this dataset is designed for localizing the lesions, we train the classification network, which is a much easier level than training the detection network. The purpose of interpreting these models is to analyze whether their criterion of classification is fair compared to human intentions. Therefore, we also make a comparison with the assessment of CheXNet in the same experimental status to verify the analysis. A detailed discussion of this is provided in Section IV.

3) Assessment of Explanation

It is difficult to judge the criterion of a better explanation because each method is designed for slightly different objectives, and because there is not always one commonly accepted measure for evaluating the quality of visualization. In analyzing radiologist-level chest X-ray images, the interpretation of diagnosis could be altered to the localization of lesions, which is crucial evidence for deciding the patient's status. Intersection of union (IOU) is widely used in semantic segmentation or object detection tasks as an evaluation metric by computing the localization scores. The evaluation would be more accurate in the case that the dataset is annotated with a segmentation mask, but there is a limitation to annotating it in the medical domain in practice. Therefore, we utilize IOU to evaluate whether positive attributions are correctly distributed in the area of lesions with a bounding box. The result we report is the localization performance of lesions without any supervision of the bounding box during the training procedure.

B. Results

1) Quantitative Assessment

To validate the efficiency of the attribution methods in visualizing target lesions, we compared them with CAM, which is widely used in medical fields to guarantee reliability. The heatmaps from each method are normalized in {0~1} and the threshold T is applied. For a fair comparison, negative attributions are cast as zero. Table 2 lists the results of the mean IOU per each class on CheXNet. Pixels that have a lower relevance value than the threshold are cast as zero. As can be seen in Table 2, CAM shows low IOU performance compared to LRP and RAP in the low threshold. Since the heatmaps from CAM are generated through resizing from low-dimension feature maps to the original input size, it is difficult to visualize the delicate interpretations for the target lesions. As the threshold value is increased, low attributions that

Table 1. The performance of AUROC for each model we used in our experiment

NIH ChestX-ray14	Atelectasis	Cardiomegaly	Effusion	Infiltration	Mass	Nodule	Pneumonia	Pneumothorax
CheXNet	0.829	0.916	0.887	0.714	0.859	0.787	0.774	0.872
RSNA pneumonia	**DenseNet-121**		**ResNet-50**		**VGG-16**		**CheXNet**	
AUROC	0.858		0.845		0.842		0.827	

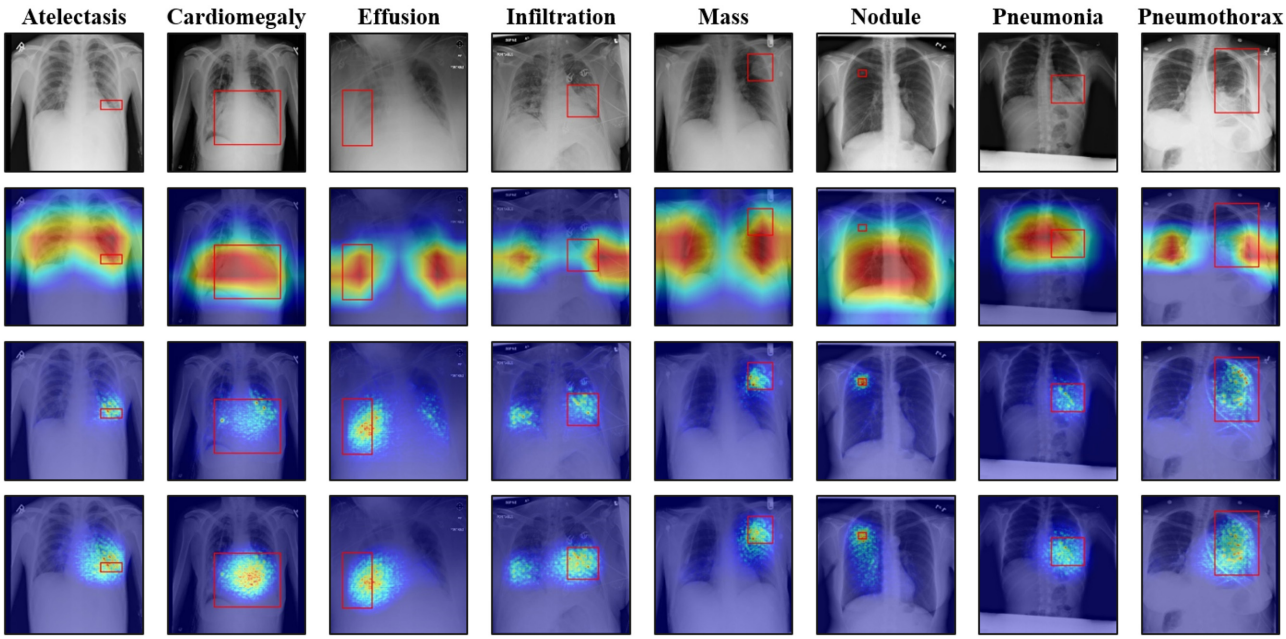

Fig. 3. Qualitative comparison of visual explanations. Each row illustrates, from top to bottom, an X-ray image, CAM, LRP, and RAP. The red square indicates the bounding box for each disease.

Table 2. For each method, the mean IoU between the bounding box and heatmaps

T(IOU)	Method	Atelectasis	Cardiomegaly	Effusion	Infiltration	Mass	Nodule	Pneumonia	Pneumothorax
0.1	CAM	0.243	0.336	0.311	0.309	0.225	0.182	0.295	0.259
	LRP	**0.500**	0.687	**0.503**	0.567	**0.514**	**0.435**	0.568	**0.456**
	RAP	0.487	**0.754**	0.490	**0.576**	0.496	0.416	**0.572**	0.447
0.2	CAM	0.304	0.439	0.368	0.369	0.280	0.238	0.355	0.303
	LRP	**0.560**	0.569	**0.543**	0.603	**0.582**	**0.506**	0.602	**0.494**
	RAP	0.534	**0.703**	0.511	**0.609**	0.540	0.461	**0.606**	0.471
0.3	CAM	0.353	0.525	0.408	0.413	0.326	0.286	0.403	0.339
	LRP	0.563	0.428	**0.526**	0.558	**0.583**	**0.545**	0.564	**0.492**
	RAP	**0.565**	**0.622**	0.518	**0.608**	0.571	0.496	**0.608**	0.479
0.4	CAM	0.396	**0.591**	0.442	0.451	0.370	0.328	0.445	0.370
	LRP	0.543	0.441	0.502	0.508	0.552	**0.561**	0.515	**0.484**
	RAP	**0.569**	0.544	**0.510**	**0.570**	**0.573**	0.521	**0.576**	0.480
0.5	CAM	0.437	**0.635**	0.468	0.483	0.411	0.369	0.479	0.397
	LRP	0.519	0.424	0.485	0.477	0.520	**0.551**	0.482	**0.479**
	RAP	**0.548**	0.480	**0.493**	**0.519**	**0.547**	0.527	**0.525**	0.478

The threshold denotes the criterion for ignoring low relevance: $\{0 \sim T\}$. The performance is the localization result without any supervision of the bounding box. The bold font indicates the best performance achieved for each disease at the specified threshold.

are widely spread with irrelevant parts are deleted, thus resulting in an improvement of the localization performance of CAM. By contrast, the attribution methods of LRP and RAP show a decrement in IOU when the threshold is too high. After the output predictions are fully decomposed and mapped pixel-by-pixel, attributions compose detailed visual explanations with a degree of importance.

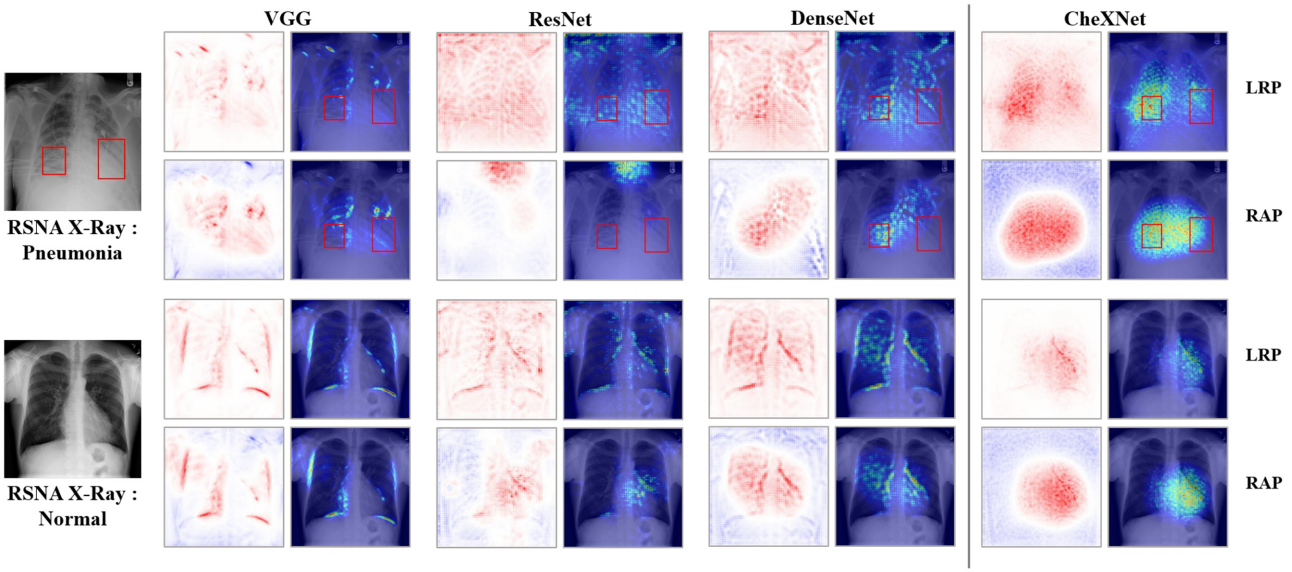

Fig. 4. Depiction of investigation of the inconsistency between human intention and what DNN has learned (see Section IV for details).

2) Qualitative Assessment

For qualitative evaluation of the heatmaps from each method, we compare the results by examining the distribution of the high activated points are distributed in the bounding box. The methods have the same purpose of emphasizing the most important factors, which allows us to assess whether each method is consistent in attributing positive relevance. Fig. 3 presents the heatmaps from each method: CAM, LRP, and RAP for the diagnostic decisions by CheXNet. We qualitatively assessed all images in the test set of the NIH chest X-ray 14 dataset, and most of them appear to show similarly satisfactory results in a human intentions. More qualitative comparisons are included in Figs. 5 and 6.

IV. INCONSISTENCY OF INTENTION

It is not trivial to elucidate the decisions made by DNNs because of the opacity associated with the myriad of linear and nonlinear operations. Their impressive performance makes it easy to believe that the criteria used to make their decisions are the same as those of human intentions. In [30], the authors point out this problem and insists on the need to explain techniques and their evaluation metrics. In the medical field in particular, the identification of causes for diagnosis is a crucial aspect of ensuring reliability.

As described in Section III-A-2, we trained DNN models on RSNA Pneumonia Detection datasets for the binary classification of pneumonia images. The performance of each model based on general learning methods shows fair performance. Figs. 5 and 6 provides visual explanations of what DNNs mainly focus on. The input X-ray images are correctly classified as target labels. For the pneumonia X-ray, the relevance from trained models—VGG, ResNet, and DenseNet—is distributed on irrelevant areas of lesions (bounding box) without regular patterns. However, CheXNet, which is pretrained on the NIH dataset with the certain purpose of classifying various diseases, shows clear visual explanations corresponding to pneumonia. For the normal X-ray image, relevance from the trained model appears in areas that support the normal lung's clear shape. The result of additional normal images shows similar relevance patterns as well. Here, the interesting phenomenon is that DNN models tend to learn the status of normal lung images rather than the characteristics of pneumonia disease. Since we do not provide any supervision of the lesion area, DNN focuses on the lungs in a normal state, which the high performance is due to the lungs being relatively large and prominent within the image. CheXNet classified this input X-ray as Cardiomegaly, which is not closely related to lung diseases, and the visual explanation clearly supports the diagnostic decision by emphasizing the lesion area of the heart.

V. CONCLUSION

In this paper, we demonstrate an efficient method for unmasking the opacity of DNNs and provide an interpretation of diagnostic decisions by utilizing explaining techniques. The introduced methods of LRP and RAP can visualize more accurate and clear parts of lesions than the generally used CAM. Generated heatmaps indicate the important factors affecting the target diseases with intensity from high relevance to low relevance. We

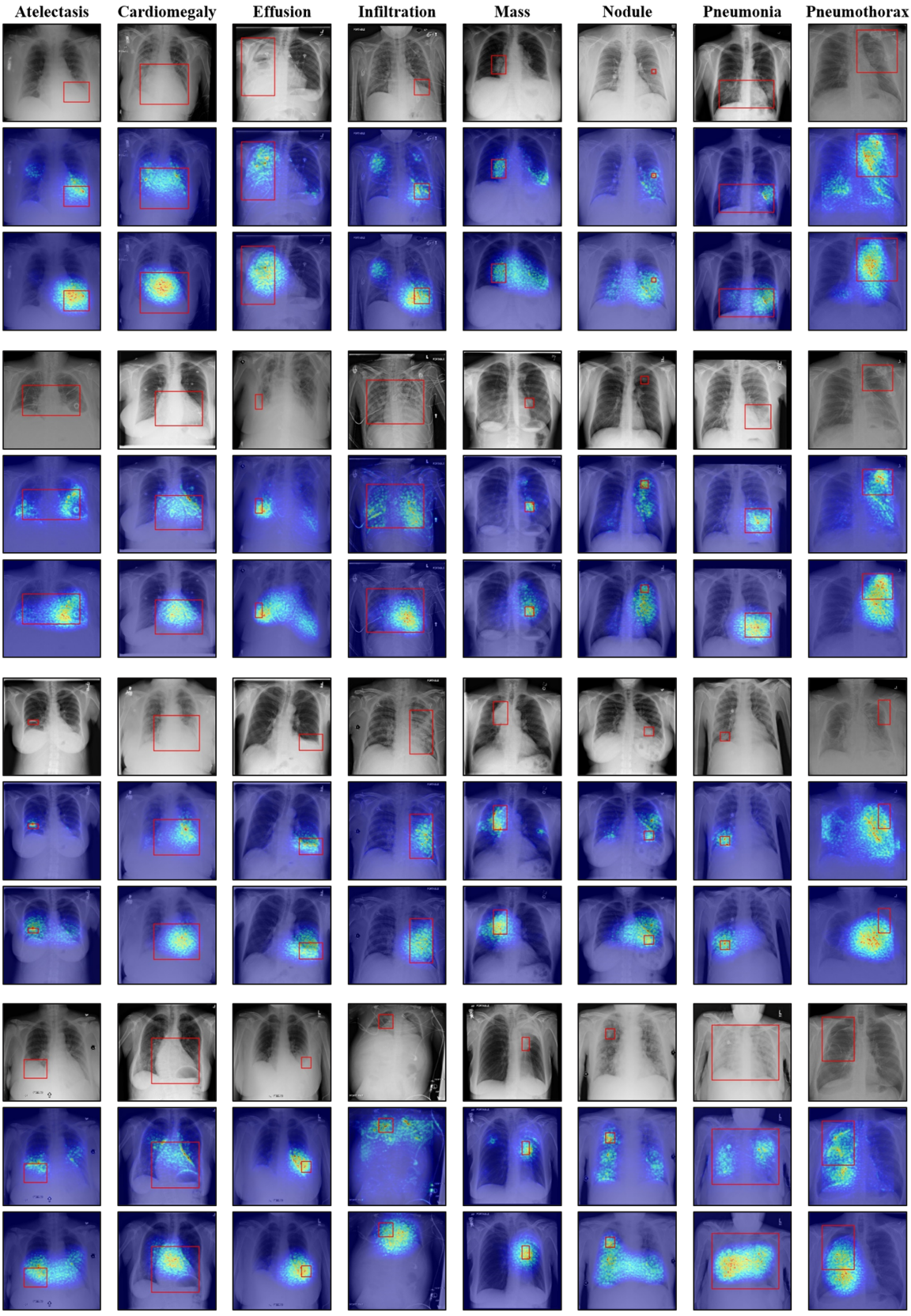

Fig. 5. First additional comparison of visual explanations generated from CheXNet. First, second, and third rows in each tuple denote image, LRP, and RAP, respectively. The visualization is the result without applying threshold.

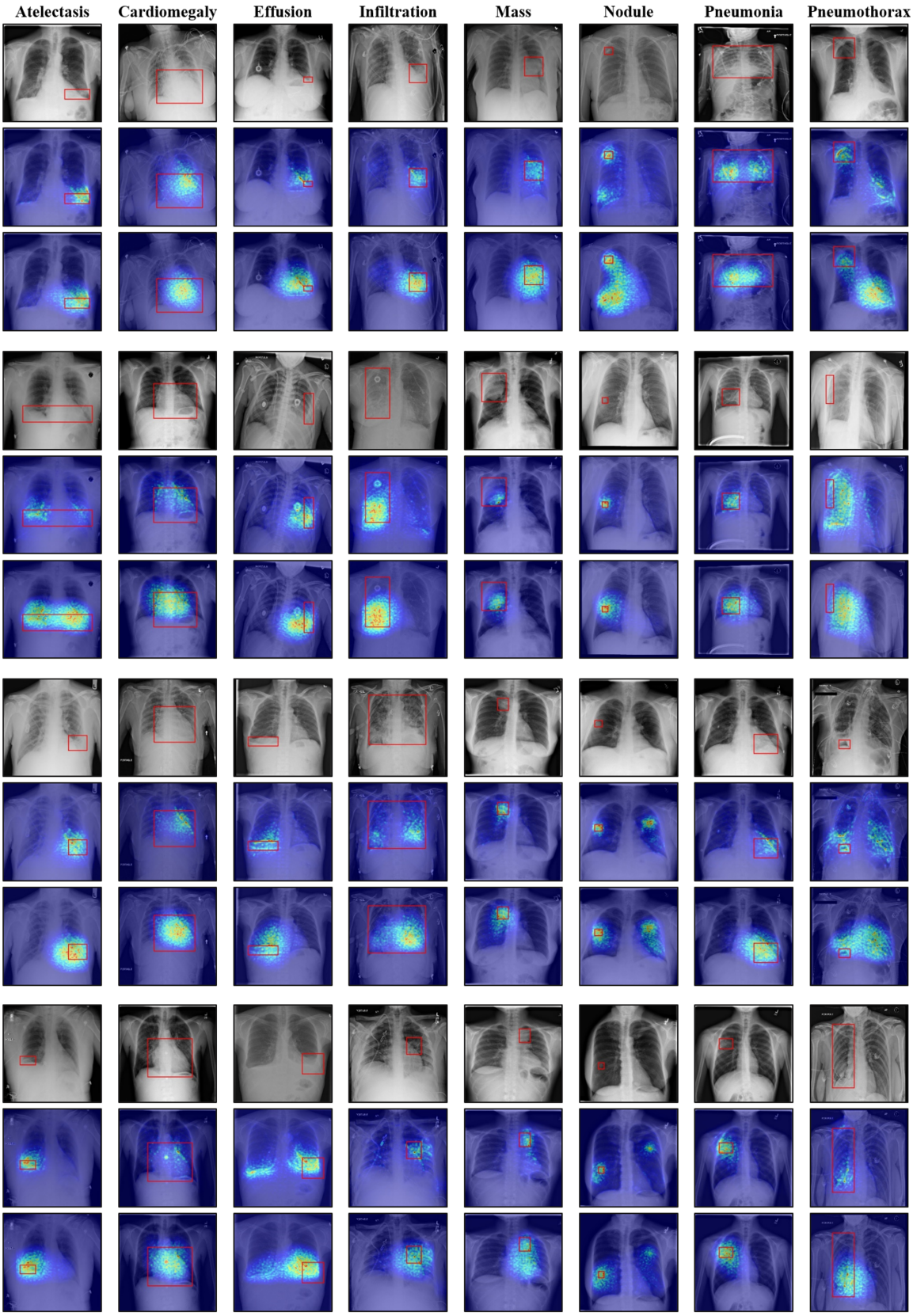

Fig. 6. Second additional comparison of visual explanations generated from CheXNet. First, second, and third rows in each tuple denote image, LRP, and RAP, respectively. The visualization is the result without applying threshold.

utilize chest X-ray datasets: NIH ChestX-ray14 and RSNA pneumonia datasets, to verify how attribution methods could localize the target lesions without any supervision of a bounding box. For the quantitative evaluation, we use the mean Intersection of Union for the three visualization methods: CAM, LRP, and RAP. The results show that fully decomposing the network by investigating the contributions of neurons makes it possible to clearly localize the parts of lesions. We also analyze the inconsistency of human intentions and DNNs by utilizing explainable methods and emphasize the necessity of interpretability for the adoption of machine intelligence in the medical domain. In a future study, we plan to describe a vision transformer-based model to demonstrate its efficiency in the medical domain with various XAI methods.

CONFLICT OF INTEREST

The authors have declared that no competing interests exist.

ACKNOWLEDGMENTS

This research was supported by Kyungpook National University Research Fund, 2023.

REFERENCES

1. A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Advances in Neural Information Processing Systems*, vol. 25, pp. 1106-1114, 2012.
2. K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, NV, USA, 2016, pp. 770-778. https://doi.org/10.1109/cvpr.2016.90
3. C. Szegedy, A. Toshev, and D. Erhan, "Deep neural networks for object detection," *Advances in Neural Information Processing Systems*, vol. 26, pp. 2553-2561, 2013.
4. H. D. Yang and S. W. Lee, "Reconstruction of 3D human body pose from stereo image sequences based on top-down learning," *Pattern Recognition*, vol. 40, no. 11, pp. 3120-3131, 2007. https://doi.org/10.1016/j.patcog.2007.01.033
5. M. C. Roh, T. Y. Kim, J. Park, and S. W. Lee, "Accurate object contour tracking based on boundary edge selection," *Pattern Recognition*, vol. 40, no. 3, pp. 931-943, 2007. https://doi.org/10.1016/j.patcog.2006.06.014
6. S. Ji, W. Xu, M. Yang, and K. Yu, "3D convolutional neural networks for human action recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 1, pp. 221-231, 2013. https://doi.org/10.1109/TPAMI.2012.59
7. M. Ahmad and S. W. Lee, "Human action recognition using multi-view image sequences," in *Proceedings of 7th International Conference on Automatic Face and Gesture Recognition (FGR06)*, Southampton, UK, 2006, pp. 523-528. https://doi.org/10.1109/FGR.2006.65
8. J. Z. Cheng, D. Ni, Y. H. Chou, J. Qin, C. M. Tiu, Y. C. Chang, C. S. Huang, D. Shen, and C. M. Chen, "Computer-aided diagnosis with deep learning architecture: applications to breast lesions in US images and pulmonary nodules in CT scans," *Scientific Reports*, vol. 6, article no. 24454, 2016. https://doi.org/10.1038/srep24454
9. S. Liu, S. Liu, W. Cai, S. Pujol, R. Kikinis, and D. Feng, "Early diagnosis of Alzheimer's disease with deep learning," in *Proceedings of 2014 IEEE 11th International Symposium on Biomedical Imaging (ISBI)*, 2014, pp. 1015-1018. https://doi.org/10.1109/ISBI.2014.6868045
10. G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. W. M. van der Laak, B. van Ginneken, and C. I. Sanchez, "A survey on deep learning in medical image analysis," *Medical Image Analysis*, vol. 42, pp. 60-88, 2017. https://doi.org/10.1016/j.media.2017.07.005
11. T. Kooi, G. Litjens, B. Van Ginneken, A. Gubern-Merida, C. I. Sanchez, R. Mann, A. den Heeten, and N. Karssemeijer, "Large scale deep learning for computer aided detection of mammographic lesions," *Medical Image Analysis*, vol. 35, pp. 303-312, 2017. https://doi.org/10.1016/j.media.2016.07.007
12. H. H. Bulthoff, S. W. Lee, T. Poggio, and C. Wallraven, *Biologically Motivated Computer Vision*. Heidelberg, Germany: Springer, 2002. https://doi.org/10.1007/3-540-36181-2
13. D. Baehrens, T. Schroeter, S. Harmeling, M. Kawanabe, K. Hansen, and K. R. Muller, "How to explain individual classification decisions," *The Journal of Machine Learning Research*, vol. 11, pp. 1803-1831, 2010.
14. S. Bach, A. Binder, G. Montavon, F. Klauschen, K. R. Muller, and W. Samek, "On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation," *PLOS One*, vol. 10, no. 7, article no. e0130140, 2015. https://doi.org/10.1371/journal.pone.0130140
15. G. Montavon, S. Lapuschkin, A. Binder, W. Samek, and K. R. Muller, "Explaining nonlinear classification decisions with deep Taylor decomposition," *Pattern Recognition*, vol. 65, pp. 211-222, 2017. https://doi.org/10.1016/j.patcog.2016.11.008
16. A. Shrikumar, P. Greenside, and A. Kundaje, "Learning important features through propagating activation differences," *Proceedings of Machine Learning Research*, vol. 70, pp. 3145-3153, 2017.
17. W. J. Nam, S. Gur, J. Choi, L. Wolf, and S. W. Lee, "Relative attributing propagation: Interpreting the comparative contributions of individual units in deep neural networks," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 3, pp. 2501-2508, 2020. https://doi.org/10.1609/aaai.v34i03.5632
18. W. J. Nam, J. Choi, and S. W. Lee, "Interpreting deep neural networks with relative sectional propagation by analyzing comparative gradients and hostile activations," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, No. 13, pp. 11604-11612, 2021. https://doi.org/10.1609/aaai.v35i13.17380
19. B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, NV, USA, 2016, pp.

2921-2929. https://doi.org/10.1109/CVPR.2016.319
20. J. C. Seah, J. S. Tang, A. Kitchen, F. Gaillard, and A. F. Dixon, "Chest radiographs in congestive heart failure: visualizing neural network learning," *Radiology*, vol. 290, no. 2, pp. 514-522, 2019. https://doi.org/10.1148/radiol.2018180887
21. M. M. R. Siddiquee, Z. Zhou, N. Tajbakhsh, R. Feng, M. B. Gotway, Y. Bengio, and J. Liang, "Learning fixed points in generative adversarial networks: from image-to-image translation to disease detection and localization," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, Seoul, South Korea, 2019, pp. 191-200. https://doi.org/10.1109%2Ficcv.2019.00028
22. S. A. Taghanaki, M. Havaei, T. Berthier, F. Dutil, L. Di Jorio, G. Hamarneh, and Y. Bengio, "InfoMask: masked variational latent representation to localize chest disease," in *Medical Image Computing and Computer Assisted Intervention-MICCAI 2019*. Cham, Switzerland: Springer, 2019, pp. 739-747. https://doi.org/10.1007/978-3-030-32226-7_82
23. R. Zhang, S. Tan, R. Wang, S. Manivannan, J. Chen, H. Lin, and W. S. Zheng, "Biomarker localization by combining CNN classifier and generative adversarial network," in *Medical Image Computing and Computer Assisted Intervention–MICCAI 2019*. Cham, Switzerland: Springer, 2019, pp. 209-217. https://doi.org/10.1007/978-3-030-32239-7_24
24. C. F. Baumgartner, L. M. Koch, K. C. Tezcan, J. X. Ang, and E. Konukoglu, "Visual feature attribution using Wasserstein GANs," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Salt Lake City, UT, USA, 2018, pp. 8309-8319. https://doi.org/10.1109/CVPR.2018.00867
25. X. Wang, Y. Peng, L. Lu, Z. Lu, M. Bagheri, and R. M. Summers, "ChestX-ray8: hospital-scale chest X-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Honolulu, HI, USA, 2017, pp. 2097-2106. https://doi.org/10.1109/CVPR.2017.369
26. P. Rajpurkar, J. Irvin, K. Zhu, B. Yang, H. Mehta, T. Duan, et al., "CheXNet: radiologist-level pneumonia detection on chest X-rays with deep learning," 2017 [Online]. Available: https://arxiv.org/abs/1711.05225.
27. G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Honolulu, HI, USA, 2017, pp. 2261-2269. https://doi.org/10.1109/CVPR.2017.243
28. Radiological Society of North America, "RSNA pneumonia detection challenge," 2019 [Online]. https://www.kaggle.com/c/rsna-pneumonia-detection-challenge/data.
29. K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014 [Online]. Available: https://arxiv.org/abs/1409.1556.
30. S. Lapuschkin, S. Waldchen, A. Binder, G. Montavon, W. Samek, and K. R. Muller, "Unmasking Clever Hans predictors and assessing what machines really learn," *Nature Communications*, vol. 10, article no. 1096, 2019. https://doi.org/10.1038/s41467-019-08987-4

Ho Kyung Shin https://orcid.org/0000-0002-7778-7265

Ho Kyung Shin received the B.S. degree in computer science and engineering from Kyungpook National University (KNU), Daegu, South Korea, in 2020, and master's degree in computer science and engineering from KNU, Daegu, South Korea, in 2022. He is currently pursuing a doctoral degree at KNU. His research interests are in computer vision and explainable artificial intelligence.

Woo-Jeoung Nam https://orcid.org/0000-0002-6548-4486

Woo-Jeoung Nam received the B.S. degree in computer science and engineering from Hanyang University, Seoul, South Korea, in 2016, and an integrated master's and Ph.D. degrees at the Department of Computer and Radio Communications Engineering, Korea University, Seoul in 2022. He is currently an assistant professor with the School of Computer Science and Engineering, Kyungpook National University, Daegu, Korea. His current research interests include machine learning, computer vision and explainable artificial intelligence.