Category-Independent Visual Explanation for Medical Deep Network Understanding

Yiming Qian[1], Liangzhi Li[6], Huazhu Fu[1], Meng Wang[1], Qingsheng Peng[2,5], Yih Chung Tham[2,3,4,5], Chingyu Cheng[2,3,4,5], Yong Liu[1], Rick Siow Mong Goh[1], and Xinxing Xu[1(✉)]

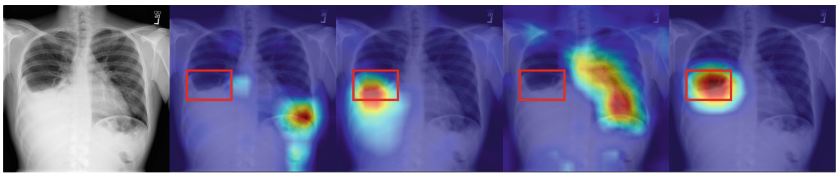
[1] Institute of High Performance Computing (IHPC), Agency for Science, Technology and Research (A*STAR), 1 Fusionopolis Way, 16-16 Connexis, Singapore 138632, Republic of Singapore
xuxinx@ihpc.a-star.edu.sg
[2] Ocular Epidemiology and Data Sciences, Singapore Eye Research Institute, Singapore, Singapore
[3] Centre for Innovation and Precision Eye Health, Yong Loo Ling School of Medicine, National University of Singapore, Singapore, Singapore
[4] Department of Ophthalmology, Yong Loo Ling School of Medicine, National University of Singapore, Singapore, Singapore
[5] Duke-NUS Medical School, National University of Singapore, Singapore, Singapore
[6] Meetyou AI Lab, Xiamen, China

Abstract. Visual explanations have the potential to improve our understanding of deep learning models and their decision-making process, which is critical for building transparent, reliable, and trustworthy AI systems. However, existing visualization methods have limitations, including their reliance on categorical labels to identify regions of interest, which may be inaccessible during model deployment and lead to incorrect diagnoses if an incorrect label is provided. To address this issue, we propose a novel category-independent visual explanation method called Hessian-CIAM. Our algorithm uses the Hessian matrix, which is the second-order derivative of the activation function, to weigh the activation weight in the last convolutional layer and generate a region of interest heatmap at inference time. We then apply an SVD-based post-process to create a smoothed version of the heatmap. By doing so, our algorithm eliminates the need for categorical labels and modifications to the deep learning model. To evaluate the effectiveness of our proposed method, we compared it to seven state-of-the-art algorithms using the Chestx-ray8 dataset. Our approach achieved a 55% higher IoU measurement than classical GradCAM and a 17% higher IoU measurement than EigenCAM. Moreover, our algorithm obtained a Judd AUC score of 0.70 on the glaucoma retinal image database, demonstrating its potential applicability in various medical applications. In summary, our category-independent visual explanation method, Hessian-CIAM, generates high-quality region of interest heatmaps that are not dependent on

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/978-3-031-43895-0_17.

H. Greenspan et al. (Eds.): MICCAI 2023, LNCS 14221, pp. 181–191, 2023.
https://doi.org/10.1007/978-3-031-43895-0_17

categorical labels, making it a promising tool for improving our understanding of deep learning models and their decision-making process, particularly in medical applications.



(a) Original (b) Mass (c) Effusion (d) Infiltration (e) Ours

Fig. 1. Example of GradCAM (b-d) supplied different labels vs. our method (e). Three different categorical labels lead GradCAM to generate different distinguishable heatmaps. By contrast, our Hessian-CIAM generates a stable ROI without the categorical label.

1 Introduction

Medical application is a field that has high requirements of model reliability, trustworthiness, and interpretability. According to the act proposed by the European Commission on AI system regulation [4], medical AI systems are categorized as high-risk systems. Five sets of requirements are listed: (1) high quality of data, (2) traceability, (3) transparency, (4) human oversight, (5) robustness, accuracy, and cybersecurity. These requirements impose a potential challenge for deep learning models where such a model is often used as a black-box system. To increase a model's explainability, many visualization methods are proposed to generate the region of interest (ROI) heatmap based on the output of the deep learning model [7,9]. This ROI heatmap highlights the region that deep learning algorithms focus on. This region often contains cues for researchers to investigate the algorithm's decision making process which would help doctors to gain confidence in the AI assisted products. For example, when doctors see a model make a correct prediction and at the same time highlight the right ROI, then it would help this model to gain more trust from doctors.

The state-of-art algorithms mostly focus on providing visualization during training where the categorical label is available. It becomes problematic at product deployment stage when no label is available. Without supplying the ground truth categorical labels, the false categorical labels would mislead the visualization algorithm to highlight wrong regions for cues. A sample is shown in Fig. 1. GradCAM [23] visualization is used widely on a deep learning network that classifies multiple diseases. Three different categorical labels are supplied (Fig. 1 (b-d)) which leads GradCAM to generate three distinguishable ROI heatmaps. To address this issue, we propose a method called **Hessian-C**ategory **I**ndependent **A**ctivation **M**aps (Hessian-CIAM), which utilizes the Hessian matrix as an activation weighting function to eliminate the need for categorical labels to compute

the ROI heatmap. Then a polarity checking process is added to the post process which corrects the polarity error from the SVD based smoothing function. Figure 1 (e) shows the visualization from our category-independent method. We benchmark our algorithm against seven state-of-art algorithms on the Chestx-ray8 dataset which demonstrated the superior performance of our algorithm. Additionally, we demonstrate a clinical use case in glaucoma detection from retinal images which shows the flexibility of our algorithm.

2 Related Works

The visual explanation for deep networks is an essential task for researchers to interpret and debug deep networks where an ROI heat map is one of the most popular tools. This field is pioneered by Oquab et al. [18] which additional Global Max Pooling (GMP) layers are added to extract the attention region from a trained convolutional network. It is later improved by CAM [27] by attaching a Global Average Pooling (GAP) layer to the existing model. The GAP identifies the extent of the object while GMP only finds one discriminative part. One drawback of Oquab's method and CAM is the requirement of modifying the original network to output visualizations. This requirement is eliminated by a gradient-based approach GradCam [23]. In this algorithm, the activation weights from the last convolutional layer of the deep network are extracted and weighed by a gradient from the back-propagation to generate the ROI heat map. This method is later improved by GradCAM++ [3] and LayerCAM [12]. An alternative way to generate an ROI heatmap is perturbation-based methods. It removes the requirement of the gradient calculation by iteratively perturbating different parts of the activations weight [22] or image [20,24] to identify the region on the image that has the highest impact on the prediction result. One major drawback of such an approach is its speed as it requires iteratively running the deep learning model. The gradient-based and perturbation-based methods deliver high-quality ROI heatmap when a categorical label is supplied. It is a useful visualization tool to help researchers interpret the deep network during the development stage. It becomes a different story when it comes to deployment. During the deployment, there is no such luxury of having a ground truth categorical label that is supplied to the visualization algorithm. One solution to relax this problem is using the prediction result as a target label, but this solution often generates a wrong visualization as when the deep learning algorithm outputs incorrect prediction. Muhammad [17] proposed a method to eliminate the dependence on the ground truth categorical label by directly applying SVD on the 2D activations and using its first principle component as the ROI heat map. The first principle component's polarity is bi-directional which could potentially highlight the non-interest region instead. Visualization techniques such as slot attention [15], SCOUTER [13], and SHAP [16] require modification on the original network and training to generate an ROI heatmap. It is not the main scope of our paper and we will not further discuss it here.

Medical applications have high requirements for model reliability, trustworthiness, and interpretability. The visualization tools such as GradCAM and

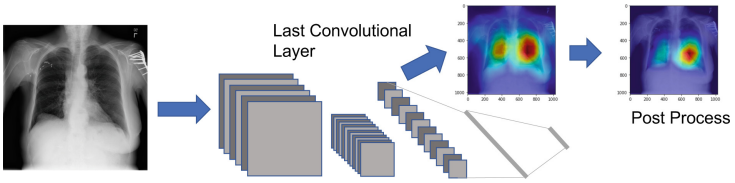


Fig. 2. Overview of our algorithm, the Hessian matrix, and activation weight from the last convolutional layer is used to create an ROI heatmap followed by a post process.

GradCAM++ are widely applied to medical applications such as retina imaging [21], X-ray [10], CT [6], MRI [26], and ultrasound [11]. However, those visualization algorithms require categorical labels to generate visual explanations. This requirement limits the usage of algorithms to the training stage where the ground truth category label is available. Generating high quality visual explanations without relying on the category label at the deployment stage remains a challenge. In this work, we propose a category-independent visual explanation method to solve this problem.

3 Method

Our algorithm generates an ROI heatmap to indicate the region on the image that the deep learning algorithms focus on when making classification decisions. Our method does not require any modification or additional training on target deep networks. The overview flow of our algorithm is illustrated in Fig. 2. Input images feed into the deep network where the activation weights from the last convolution layer are weighted by the Hessian matrix followed by a post-process to output a clean ROI heatmap.

It is well known that the Hessian matrix appears in the expansion of gradient about a point in parameter space [19], as:

$$\bigtriangledown_\omega (\omega + \Delta\omega) = \bigtriangledown_\omega(\omega) + H\Delta\omega + O(\|\Delta\omega\|^2), \tag{1}$$

where ω is a point in parameter space, $\Delta\omega$ is a perturbation of ω, $\Delta\omega$ is the gradient and H is the hessian matrix. In order to approximate the Hessian matrix H, we let $\Delta\omega = rv$, where v is the identity matrix, and r is a small number which leads the $O(r)$ term to become insignificant. So we can further simplify the equation into:

$$Hv = \frac{\bigtriangledown_\omega(\omega + rv) - \bigtriangledown_\omega(\omega)}{r} + O(r) = \frac{\bigtriangledown_\omega(\omega + rv) - \bigtriangledown_\omega(\omega)}{r}. \tag{2}$$

Our goal is to apply the Hessian matrix as a weighting function to indicate the significance of each activation function output in the CNN. So we applied an L2 normalization on the Hv, here v is an identity matrix, so we can get the

normalized Hessian matrix $\hat{H} = \frac{|Hv|}{\|Hv\|_2}$. In the CNN we denote A^k as the feature activation map from the kth convolution layer. $\hat{H}^k$ denotes the normalized Hessian matrix in the kth layer. We calculate the Hadamard product between $\hat{H}^k$ and A^k, then apply ReLU to obtain the new activation map. n is the depth of the activation map. The ROI heatmap $L_H = \sum_{k=1}^{n} ReLu(\hat{H}^k \odot A^k)$.

The ROI heatmap L_H can be noisy, we follow Muhammad's approach [17] to smooth out the L_H which applies SVD on $A_H^k = ReLu(\hat{H}^k \odot A^k) = U\Sigma V^T$ where U denotes a $M \times M$ matrix. Σ denotes a diagonal matrix with size of $M \times N$. V denotes a $N \times N$ matrix. The column of U and V are the left singular vectors. The V_1 denotes the first component in V which is a weight function to create a smoothed ROI heatmap $L_{HS} = A_H^k V_1$. One drawback of Muhammad's approach [17] is the polarity of V_1 is not considered as the Eigenvectors from SVD are bidirectional. It could lead the algorithm to output non-ROI regions. To solve this problem, we revise the algorithm to calculate the correlation between the smoothed version L_{HS} and the original ROI heatmap L_H. If the correlation appears negative, we will reverse the ROI heatmap, as:

$$L_{HS} = \begin{cases} ReLu(A_H^k V_1), & \text{if } corr(A_H^k V_1, L_H) > 0, \\ ReLu(-A_H^k V_1), & \text{otherwise.} \end{cases} \tag{3}$$

4 Experiment

4.1 Experiment Setup

We conduct experiments on lung disease classification Chestx-ray8 [25] to evaluate the performance of our algorithm. The Chestx-ray8 dataset contains 100,000 x-ray images with 19 disease labels. It is a significantly imbalanced dataset with some categories having as few as 7 images. To demonstrate our visualization techniques, we simplified the dataset by selecting 6 diseases with a higher number of images. After the selection, our training set contains images from atelectasis (3135 images), effusion (2875 images), infiltration (6941 images), mass (1665 images), nodule (2036 images), and pneumothorax (1485 images). Additionally, we randomly selected 7000 images from healthy people. 20% of images in the training set were set aside as validation sets for parameter tuning. This dataset contains 881 test images with bounding boxes that indicate the location of the diseases which 644 images were in the 6 diseases we selected.

We utilize the pre-trained ResNet50 [8] as the backbone. The cross-entropy loss is used as a loss function; the learning rate is set to 0.00001; the batch size is 64. The training cycle is set to 100 epochs. Our workstation is equipped with 2 Nvidia 3090 GPU (24 GB RAM), Intel Xeon CPU (3.30 GHz), and 128 GB RAM.

4.2 Quantitative Evaluation

The algorithm is evaluated following the method proposed by Cao et al. [2]. The union of intersection (IoU) between the bounding box and ROI is measured. The

Table 1. Quantitative evaluation of visualization methods on Chestx-ray8 dataset. The IoU using prediction as the label is shown here. The value in the bracket is the IoU that uses ground truth as the label.

	IoU on different thresholds				
	0.95	0.90	0.85	0.80	0.75
Gradient based approaches					
GradCAM [23]	0.127 (0.136)	0.139 (0.151)	0.150 (0.163)	0.159 (0.175)	0.175 (0.185)
GradCAM++ [3]	0.103 (0.108)	0.117 (0.124)	0.131 (0.139)	0.144 (0.154)	0.155 (0.167)
HiResCAM [5]	0.104 (0.120)	0.112 (0.133)	0.118 (0.143)	0.124 (0.153)	0.129 (0.162)
Perturbation based approaches					
AblationCAM [22]	0.092 (0.090)	0.097 (0.094)	0.102 (0.098)	0.107 (0.103)	0.113 (0.109)
ScoreCAM [24]	0.134 (N/A)	0.141 (N/A)	0.149 (N/A)	0.158 (N/A)	0.168 (N/A)
RISE [20]	0.095 (0.097)	0.096 (0.096)	0.097 (0.097)	0.098 (0.098)	0.099 (0.099)
Category-independent approaches					
EigenCAM [17]	0.213	0.222	0.227	0.231	0.232
Ours	**0.240**	**0.253**	**0.262**	**0.267**	**0.271**

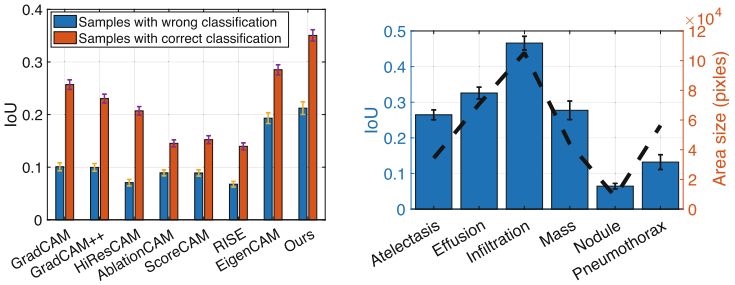


Fig. 3. (left) IoU on samples with a wrong and correct prediction on methods. (right) IoU for our method on different diseases in the bar chart (left y-axis) and the ground truth bounding box size (dashed line, right y-axis).

foreground of ROI is extracted based on applying thresholds to find the area that covers 95%, 90%, 85%, 80%, and 75% of energy from the heatmap. The gradient and perturbation-based methods require ground truth labels to generate an ROI heatmap but, at the inference time, the ground truth label is not available. To simulate the deployment scenario, we conduct two sets of evaluations. In the first evaluation, the prediction results (our ResNet model delivers 42.6% prediction accuracy) from the deep learning model are used as a label feed into the visualization methods. One drawback of this approach is the prediction result is not always reliable and the incorrect prediction could mislead the algorithm to output the wrong ROI. As a comparison, in the second evaluation, we supply ground truth labels to visualization methods. The quantitative evaluation of different visualization methods is shown in Table 1.

Three groups of visualization algorithms are evaluated in our experiment. The gradient based group contains the algorithm that relies on the gradient from the label to generate the ROI. In this group, GradCAM achieved the highest at 0.175 IoU at the 75% threshold. The pertubation based group makes small perturbations in the input image or activation weight to find the ROI that has the highest impact. In this group, the ScoreCAM achieved the highest 0.168 IoU at the 75% threshold. The category independent group contains algorithms that do not require a label to generate ROI. Our method scored the highest IoU at 0.271 IoU at the 75% threshold. When ground truth labels are supplied, the IoU for gradient based methods is improved in the range of 5% to 20%. For perturbation based methods, supplying ground truth data reduced the performance of AblationCAM and had minimal impact on RISE.

Next, we split the test set into two categories which are samples with wrong and correct predictions (shown in Fig. 3 (left)). The 75% threshold is used to calculate IoU. The samples with correct prediction consistently scored higher IoU across all visualization methods. Our method shows the highest performance in both wrong and correct prediction categories. The perturbation based methods consistently scored lower than other methods indicating this group is not suitable for X-ray image classification applications.

To further investigate the efficiency of our algorithm, we extract the IoU on each disease (Fig. 3 (right)) where a 75% threshold is applied to calculate the mean IoU. The evaluation shows our algorithm is positively correlated with the size of the ground truth bounding box. It indicates the disease with a larger infection area is easier to visualize by our algorithm.

4.3 Qualitative Evaluation

Sample images comparing our algorithm with five state-of-art algorithms are shown in Fig. 4. Our algorithm has a cleaner heatmap. The gradient methods generate a heatmap that contains a higher level of noise that covers a large area of the non-lung regions such as the shoulder. The perturbation based methods deliver the worst visualization in our evaluation. The AlbationCAM and ScoreCAM are only able to highlight the whole lung area but it does not provide any clinical value to pinpoint the disease locations. The RISE [20] method delivers multiple clusters of highlight regions that are not feasible to provide human-readable information. The last row of Fig. 4 shows the worst case in our evaluation which is a representative case to illustrate the failure mode of our algorithm. The deep learning algorithm may fail to detect the small size lesions which leads to the wrong ROI for visualization methods. More comparison is available in the supplementary material.

4.4 Clinical Application

Our algorithm has the potential to apply to many clinical applications. We conducted an additional experiment on the glaucoma retinal image database [14]

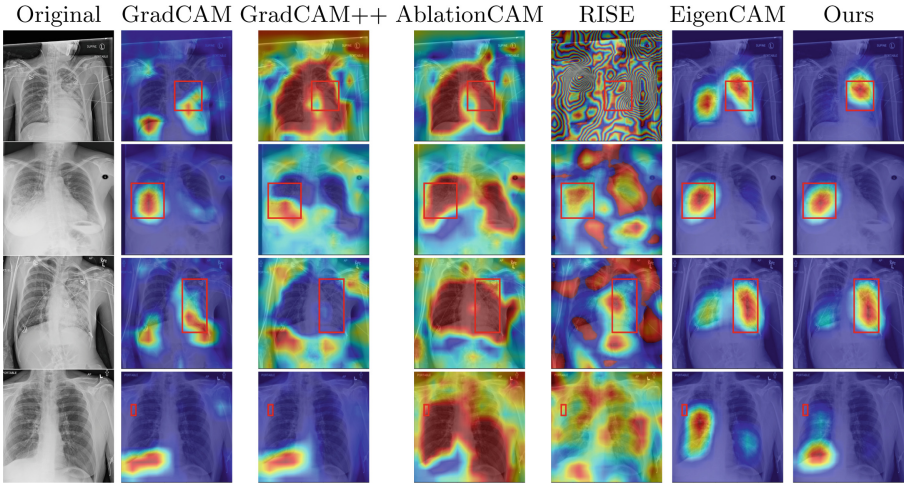


Fig. 4. Comparison of visualization methods on Chestx-ray8 dataset. The ground truth bounding box drawn by clinicians overlays on the heatmaps.

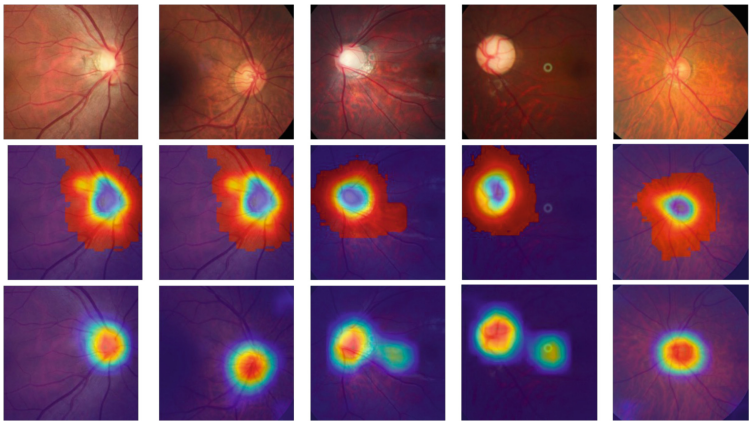


Fig. 5. Five samples of the original image (row 1), ground truth saliency map (row 2), and heatmap from our method (row 3) are shown.

with 3,144 negative and 1,712 positive glaucoma samples[1] Each sample contains a saliency map annotated by ophthalmologists by using mouse clicks to simulate the human visual attention process. Since our goal is to evaluate the explainability of our visualization algorithm, we decided to use all images to train the glaucoma classification model. We follow the work from Bylinskii et al. [1] to apply similarity (histogram intersection), cross-correlation, and Judd AUC to measure the performance of our algorithm. The 95% energy of the ROI heatmap was used

[1] the dataset is obtained from https://github.com/smilell/AG-CNN.

as a threshold to clean our heatmap. Our algorithm achieved 0.618 ± 0.0024 in similarity, 0.755 ± 0.0033 in cross-correlation, and 0.703 ± 0.0013 in Judd AUC. The complete evaluation is available in the supplementary material (Fig. 5).

5 Conclusion

In this study, we propose a novel category-independent deep learning visualization algorithm that does not rely on categorical labels to generate visualizations. Our evaluation demonstrates that our algorithm outperforms seven state-of-the-art algorithms by a significant margin on a multi-disease classification task using X-ray images. This indicates that our algorithm has the potential to enhance model explainability and facilitate its deployment in medical applications. Additionally, we demonstrate the flexibility of our algorithm by showing a clinical use case on retinal image glaucoma detection. Overall, our proposed Hessian-CIAM algorithm represents a promising tool for improving our understanding of deep learning models and enhancing their interpretability, particularly in medical applications.

Acknowledgements. This work is supported by the Agency for Science, Technology and Research (A*STAR) under its RIE2020 Health and Biomedical Sciences (HBMS) Industry Alignment Fund Pre-Positioning (IAF-PP) Grant No. H20c6a0031, the National Research Foundation, Singapore under its AI Singapore Programme (AISG Award No: AISG2-TC-2021-003), the Agency for Science, Technology and Research (A*STAR) through its AME Programmatic Funding Scheme Under Project A20H4b0141, A*STAR Central Research Fund "A Secure and Privacy Preserving AI Platform for Digital Health".

References

1. Bylinskii, Z., Judd, T., Oliva, A., Torralba, A., Durand, F.: What do different evaluation metrics tell us about saliency models? arXiv preprint arXiv:1604.03605 (2016)
2. Cao, C., et al.: Look and think twice: capturing top-down visual attention with feedback convolutional neural networks. In: 2015 IEEE International Conference on Computer Vision (ICCV), pp. 2956–2964 (2015)
3. Chattopadhay, A., Sarkar, A., Howlader, P., Balasubramanian, V.N.: Grad-CAM++: generalized gradient-based visual explanations for deep convolutional networks. In: IEEE Winter Conference on Applications of Computer Vision (WACV), pp. 839–847 (2018)
4. COMMISSION, E.: Proposal for a regulation of the European parliament and of the council (2021). https://artificialintelligenceact.eu/the-act/
5. Draelos, R.L., Carin, L.: HiResCAM: faithful location representation in visual attention for explainable 3D medical image classification. arXiv preprint arXiv:2011.08891 (2020)
6. Draelos, R.L., Carin, L.: Explainable multiple abnormality classification of chest CT volumes. Artif. Intell. Med. **132**(C), 102372 (2022)

7. Guo, Y., Liu, Y., Oerlemans, A., Lao, S., Wu, S., Lew, M.S.: Deep learning for visual understanding: A review. Neurocomputing **187**, 27–48 (2016). recent Developments on Deep Big Vision
8. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2016)
9. Hohman, F., Kahng, M., Pienta, R., Chau, D.H.: Visual analytics in deep learning: an interrogative survey for the next frontiers. IEEE Trans. Visual Comput. Graphics **25**(8), 2674–2693 (2019). https://doi.org/10.1109/TVCG.2018.2843369
10. Irvin, J., et al.: Chexpert: a large chest radiograph dataset with uncertainty labels and expert comparison. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 33, pp. 590–597 (2019)
11. Ishikawa, G., Xu, R., Ohya, J., Iwata, H.: Detecting a fetus in ultrasound images using Grad-CAM and locating the fetus in the uterus. In: ICPRAM, pp. 181–189 (2019)
12. Jiang, P.T., Zhang, C.B., Hou, Q., Cheng, M.M., Wei, Y.: Layercam: exploring hierarchical class activation maps for localization. IEEE Trans. Image Process. **30**, 5875–5888 (2021)
13. Li, L., Wang, B., Verma, M., Nakashima, Y., Kawasaki, R., Nagahara, H.: SCOUTER: slot attention-based classifier for explainable image recognition. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pp. 1046–1055 (2021)
14. Li, L., Xu, M., Wang, X., Jiang, L., Liu, H.: Attention based glaucoma detection: a large-scale database and CNN model. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2019)
15. Locatello, F., et al..: Object-centric learning with slot attention. In: Advances in Neural Information Processing Systems (NIPS), vol. 33, pp. 11525–11538 (2020)
16. Lundberg, S.M., Lee, S.I.: A unified approach to interpreting model predictions. In: Guyon, I., et al. (eds.) Advances in Neural Information Processing Systems (NIPS), pp. 4765–4774. Curran Associates, Inc. (2017)
17. Muhammad, M.B., Yeasin, M.: Eigen-CAM: class activation map using principal components. In: 2020 International Joint Conference on Neural Networks (IJCNN), pp. 1–7. IEEE (2020)
18. Oquab, M., Bottou, L., Laptev, I., Sivic, J.: Is object localization for free? weakly-supervised learning with convolutional neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 685–694 (2015)
19. Pearlmutter, B.A.: Fast exact multiplication by the hessian. Neural Comput. **6**(1), 147–160 (1994)
20. Petsiuk, V., Das, A., Saenko, K.: RISE: randomized input sampling for explanation of black-box models. arXiv preprint arXiv:1806.07421 (2018)
21. Poplin, R., et al.: Prediction of cardiovascular risk factors from retinal fundus photographs via deep learning. Nature Biomed. Eng. **2**(3), 158–164 (2018)
22. Ramaswamy, H.G., et al.: Ablation-CAM: visual explanations for deep convolutional network via gradient-free localization. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), pp. 983–991 (2020)
23. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-CAM: visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV), pp. 618–626 (2017)

24. Wang, H., et al.: Score-CAM: score-weighted visual explanations for convolutional neural networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 24–25 (2020)
25. Wang, X., Peng, Y., Lu, L., Lu, Z., Bagheri, M., Summers, R.M.: 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 (CVPR), pp. 2097–2106 (2017)
26. Yang, C., Rangarajan, A., Ranka, S.: Visual explanations from deep 3D convolutional neural networks for Alzheimer's disease classification. In: AMIA Annual Symposium Proceedings, vol. 2018, p. 1571. American Medical Informatics Association (2018)
27. Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., Torralba, A.: Learning deep features for discriminative localization. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2921–2929 (2016)