Quantitative evaluation of activation maps for weakly-supervised lung nodule segmentation

Finn Behrendt*[a], Suyash Sonawane*[a], Debayan Bhattacharya[a], Lennart Maack[a], Julia Krüger[b], Roland Opfer[b], and Alexander Schlaefer[a]

[a]Institute of Medical Technology and Intelligent Systems, Hamburg University of Technology, 21073 Hamburg, Germany
[b]Jung Diagnostics GmbH, 22335 Hamburg, Germany

ABSTRACT

The manual assessment of chest radiographs by radiologists is a time-consuming and error-prone process that relies on the availability of trained professionals. Deep learning methods have the potential to alleviate the workload of radiologists in pathology detection and diagnosis. However, one major drawback of deep learning methods is their lack of explainable decision-making, which is crucial in computer-aided diagnosis. To address this issue, activation maps of the underlying convolutional neural networks (CNN) are frequently used to indicate the regions of focus for the network during predictions. However, often, an evaluation of these activation maps concerning the actual predicted pathology is missing. In this study, we quantitatively evaluate the usage of activation maps for segmenting pulmonary nodules in chest radiographs. We compare transformer-based, CNN-based, and hybrid architectures using different visualization methods. Our results show that although high performance can be achieved in the classification task across all models, the activation masks show little correlation with the actual position of the nodules.

Keywords: Nodule Detection; Chest Radiographs; Weakly Supervised Segmentation; Explainable AI; Deep Learning

1. INTRODUCTION

Manually assessing chest X-rays (CXR) requires expert knowledge, is a time-consuming process and is prone to errors.[1] As a result, there is a growing interest in exploring the application of deep learning methods to assist in detecting common pathologies found in CXRs. These methods hold the potential to reduce the workload of radiologists and improve diagnostic accuracy in the interpretation of CXRs.[2]

In recent years, deep learning-based models have been developed as highly effective tools for detecting and classifying multiple pathologies in chest x-rays.[2] While these models have proven beneficial in assisting radiologists with CXR assessments, there is a greater need for localization of the identified conditions to aid radiologists further. Additionally, an essential requirement for a support tool is its ability to provide explanations for its decision-making process.[2] However, deep learning models, such as convolutional neural networks (CNN), lack explanations for why a particular class is predicted. To gain insights into the black-box behavior of CNNs, various approaches have been employed to visualize regions and features sensitive to prediction changes. These visualization techniques aim to achieve a sense of explainability by generating heatmaps highlighting regions important for prediction. By employing these methods, some level of interpretability can be attained, enabling a suggestion of how the CNNs arrive at their decisions.[3]

While CNNs have dominated state-of-the-art in classification tasks, the recent development of attention-based vision transformers[4] (ViT) challenge the classification performance of CNNs.[5,6] In addition, the self-attention mechanism in ViTs allows directly visualizing the regions of interest for the classification task at hand.[6,7] As the activation maps provide a coarse localization of the classified pathology, they could be provided to radiologists for model verification and to guide their attention to a specific region in the CXR. While recently the

*Authors contributed equally.

Further author information: Finn Behrendt: E-mail: finn.behrendt@tuhh.de

Medical Imaging 2024: Computer-Aided Diagnosis, edited by Weijie Chen,
Susan M. Astley, Proc. of SPIE Vol. 12927, 129272P · © 2024
SPIE · 1605-7422 · doi: 10.1117/12.3006416

evaluation of activation maps from CNNs gained interest,[8–10] in many research publications, activation maps are provided alongside classification results solely facilitating qualitative analysis.[2] Furthermore, ViTs are not evaluated for this task despite their easily accessible attention maps and promising features. Therefore, this study aims to evaluate the attention maps generated by various model architectures, including Convolutional Neural Networks (CNNs), ViTs and hybrid models. In addition to the classification task, we derive segmentation predictions through post-hoc analysis techniques like GradCAM[3] or by directly utilizing the attention patterns of ViTs. We train the classification task based on a data set of CXRs that contains pulmonary lung nodules. The segmentation predictions are then compared to pixel-wise ground truth annotations from radiologists. This evaluation aims to gain insights into the effectiveness and accuracy of the attention maps generated by different model architectures in the context of chest x-ray analysis.

Our research reveals notable difficulties in attaining pixel-wise segmentation using weakly-supervised activation maps from classification networks. Despite achieving high classification performance, all models demonstrate unsatisfactory pixel-wise segmentation results. We conclude that the network's activation maps follow complex image patterns and cannot be straightforwardly interpreted as segmentation maps without further supervision.

2. METHODS

This study aims to assess the activation maps of deep learning classification networks for the pixel-wise segmentation of pulmonary lung nodules. First, we train the models to classify whether a CXR contains nodules. Subsequently, we utilize the activation maps to evaluate the pixel-wise segmentation performance compared to the ground-truth bounding boxes.

2.1 Data

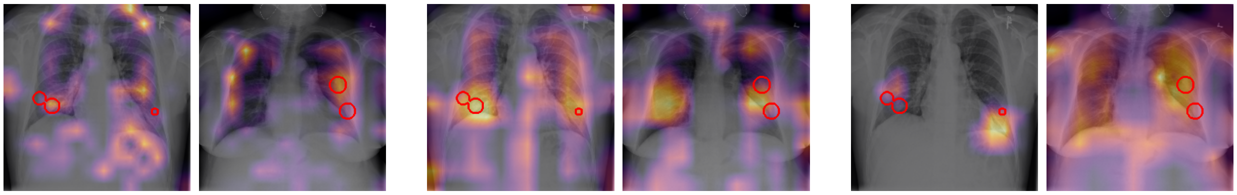
We conduct experiments using the Node21 competition data set, which comprises 4882 frontal CXRs sourced from four distinct public data sets: JRST (N=242), PadChest (N=1680), Chestx-ray14 (N=1804), and Open-I (N=1218). These x-rays are carefully reviewed and annotated by radiologists. Of the CXRs, 3748 are without nodules, while 1134 exhibit at least one nodule, totaling 1476. The Node21 challenge's annotation protocol involves identifying solitary, solid, or subsolid nodules while filtering out clusters of more than three nodules. The data set exclusively includes nodules with a diameter ranging from 6 mm to 30 mm. The annotation is given as bounding box information. To render a segmentation mask close to the nodule shape, we fit a circle inside the bounding boxes. All CXRs are resized to a resolution of 224×224 pixels using bilinear interpolation to save computational resources. We split the data into a balanced held-out test set of 400 CXRs such that there is an equal amount of x-rays with and without nodules. Furthermore, we split the remaining training data into 5 training and validation sets via cross-validation.

2.2 Models and Training

For transformer-based architectures, DeiT[5] and PatchConvNet[7] are selected to analyze activation maps. DeiT is a variation of ViT, leveraging knowledge distillation, while PatchConvNet combines CNN architecture with the attention mechanism of ViTs. These models are compared to a state-of-the-art DenseNet121 CNN architecture. We initialize the models with ImageNet weights and train for 50 epochs. The checkpoint with the lowest validation loss is chosen for evaluation. The learning rate is reduced by a factor of 0.8 every epoch. For DenseNet121, we use an initial learning rate of 5e-6 and a batch size of 32. For DeiT we use an initial learning rate of 1e-5 and a batch size of 32 and for PatchConvNet an initial learning rate of 1e-6 and a batch size of 16 is used.

2.3 Visualization of Activation Maps

After training the models with image-level labels, we extract saliency maps. Thereby, we utilize GradCAMs[3] concerning the last layer for all model architectures. Additionally, we extract the attention maps from transformer-based architectures and visualize the attention scores of the last layer. Hereby, we use the maximum values of the different attention heads.



Deit Dense Net Patch Conv Net

2.4 Evaluation Metrics

To assess classification performance, two metrics are used: the Area under Precision-Recall Curve (AUPRC) and the Area under Receiver Operator Curve (AUROC). Furthermore, we utilize the pixel-wise AUPRC for the images containing nodules to evaluate the segmentation performance. The generated attention maps are normalized to the range of 0 to 1 and compared to the binary masks (ground truth) to calculate the AUPRC.

3. RESULTS

In Table 1 (left), we observe strong classification performance across all models regarding both the AUROC and AUPRC. While the Dense Net and Patch Conv Net perform similarly, the best performance is achieved by the Deit.

In contrast to the strong classification performance, we observe poor segmentation performance across all models in Table 1 (right). Notably, the attention-based model architectures show improved segmentation metrics. The highest performance is achieved from the attention maps of Patch Conv Net.

Considering Figure **??**, we observe that while the activation maps significantly differ across the network architectures, they follow specific patterns in the CXRs like lung boundaries or the contour of the hearth, rather than pointing to the given nodules.

Model	AUPRC	AUROC
Deit	80.93 ± 0.90	81.42 ± 1.08
Dense Net	75.50 ± 0.64	76.89 ± 0.89
Patch Conv Net	74.67 ± 3.14	77.56 ± 2.64

	AUPRC	
Model	GradCAM	Attention
Deit	3.45 ± 2.38	3.42 ± 1.36
Dense Net	1.36 ± 0.34	N/A
Patch Conv Net	2.51 ± 1.57	5.9 ± 0.92

Table 1: Nodule classification (left) and segmentation (right) Performance. All metrics are provided in percent, model predictions across different folds are provided as (mean ± standard deviation).

4. DISCUSSION AND CONCLUSION

In this study, we aimed to explore the potential of activation maps for explainable AI and investigate their usability for localization without any guidance during training. Additionally, we compared two different visualization mechanisms: post-hoc visualization using CNN-based activation maps and direct visualization using the attention maps of ViTs. Our experiments indicate that all models can classify whether nodules are present in CXRs. However, despite the high classification performance, segmentation performance remains unsatisfactory, although slightly higher scores can be observed for attention-based visualizations.

We attribute the subpar segmentation performance to the appearance of nodules in CXRs, characterized by subtle intensity shifts that require the context of the whole image to be detected. Each extracted activation map holds a specific significance for the model's prediction. Some activation maps are utilized to identify specific edges or intensities that correlate with the presence of the pathology, facilitating detection. On the other hand, some activation maps capture the overall image context and intensity distributions at specific locations, which may serve for calibration purposes. However, due to the black-box nature of deep learning networks, we find that directly extracting a specific feature map that consistently highlights the pathology is not feasible. The considered deep learning networks rely on aggregates of multiple feature maps for the classification, which hinders the direct applicability of activation maps in the nodule segmentation task. To address this challenge, we highlight the need for additional guidance during training.[11,12] Furthermore, we note that another reason for the poor segmentation performance might be that the evaluated networks rely on features not directly related to the nodules but other image features that correlate with the nodules.

In conclusion, while Attention Maps can offer valuable insights for explainable AI, particularly in generic images, achieving precise localization of nodules in chest radiographs without guidance during training remains challenging. Our study provides a quantitative comparison and highlights the importance of additional strategies to improve localization using attention maps. Future research should explore advanced guidance techniques, particularly for attention-based models, to unlock the full potential of attention maps for accurate and reliable localization in medical image analysis.

REFERENCES

[1] Brady, A. P., "Error and discrepancy in radiology: inevitable or avoidable?," *Insights into Imaging* **8**, 171–182 (Feb. 2017).

[2] Çallı, E., Sogancioglu, E., van Ginneken, B., van Leeuwen, K. G., and Murphy, K., "Deep learning for chest x-ray analysis: A survey," *Med. Image Anal.* **72**, 102125 (2021).

[3] Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., and Batra, D., "Grad-cam: Visual explanations from deep networks via gradient-based localization," in [*Proceedings of the IEEE international conference on computer vision*], 618–626 (2017).

[4] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., and Houlsby, N., "An image is worth 16x16 words: Transformers for image recognition at scale," in [*ICLR*], (2021).

[5] Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., and Jégou, H., "Training data-efficient image transformers & distillation through attention," in [*International conference on machine learning*], 10347–10357, PMLR (2021).

[6] Behrendt, F., Bhattacharya, D., Krüger, J., Opfer, R., and Schlaefer, A., "Data-efficient vision transformers for multi-label disease classification on chest radiographs," *Current Directions in Biomedical Engineering* **8**(1), 34–37 (2022).

[7] Touvron, H., Cord, M., El-Nouby, A., Bojanowski, P., Joulin, A., Synnaeve, G., and Jégou, H., "Augmenting convolutional networks with attention-based aggregation," *arXiv preprint arXiv:2112.13692* (2021).

[8] Ozer, C. and Oksuz, I., "Explainable image quality analysis of chest x-rays," in [*Medical Imaging with Deep Learning*], (2021).

[9] Zhang, J., Chao, H., Dasegowda, G., Wang, G., Kalra, M. K., and Yan, P., "Overlooked trustworthiness of saliency maps," in [*International Conference on Medical Image Computing and Computer-Assisted Intervention*], 451–461, Springer (2022).

[10] Kang, H., Park, H.-m., Ahn, Y., Van Messem, A., and De Neve, W., "Towards a quantitative analysis of class activation mapping for deep learning-based computer-aided diagnosis," in [*Medical Imaging 2021: Image Perception, Observer Performance, and Technology Assessment*], **11599**, 119–131, SPIE (2021).

[11] Joshi, A., Mishra, G., and Sivaswamy, J., "Explainable disease classification via weakly-supervised segmentation," in [*Interpretable and Annotation-Efficient Learning for Medical Image Computing: Third International Workshop, Held in Conjunction with MICCAI 2020, Lima, Peru, October 4–8, 2020, Proceedings 3*], 54–62, Springer (2020).

[12] Pesce, E., Joseph Withey, S., Ypsilantis, P.-P., Bakewell, R., Goh, V., and Montana, G., "Learning to detect chest radiographs containing pulmonary lesions using visual attention networks," *Medical Image Analysis* **53**, 26–38 (2019).