Comparing radiologists' gaze and saliency maps generated by interpretability methods for chest x-rays

Ricardo Bigolin Lanfredi[1] RICBL@SCI.UTAH.EDU
[1] *Scientific Computing and Imaging Institute, University of Utah, Salt Lake City, UT, USA*

Ambuj Arora[2]
[2] *School of Computing, University of Utah, Salt Lake City, UT, USA*

Trafton Drew[3]
[3] *Department of Psychology, University of Utah, Salt Lake City, UT, USA*

Joyce D. Schroeder[4]
[4] *Department of Radiology and Imaging Sciences, University of Utah, Salt Lake City, UT, USA*

Tolga Tasdizen[1]

Editors: Under Review for MIDL 2022

Abstract

The interpretability of medical image analysis models is considered a key research field. We use a dataset of eye-tracking data from five radiologists to compare the outputs of interpretability methods against the heatmaps representing where radiologists looked. We conduct a class-independent analysis of the saliency maps generated by two methods selected from the literature: Grad-CAM and attention maps from an attention-gated model. For the comparison, we use shuffled metrics, which avoid biases from fixation locations. We achieve scores comparable to an interobserver baseline in one shuffled metric, highlighting the potential of saliency maps from Grad-CAM to mimic a radiologist's attention over an image. We also divide the dataset into subsets to evaluate in which cases similarities are higher.

Keywords: Interpretability, XAI, Chest X-rays, Radiology, Eye Tracking, Gaze, Saliency Maps, Grad-CAM, Attention Gated Network

1. Introduction

The interpretability of deep learning models is an essential property for their adoption in the medical field (Kelly et al., 2019). One of the most used explanation methods in this field (Reyes et al., 2020) is Grad-CAM (Selvaraju et al., 2017). Grad-CAM uses the gradient of the network's outputs with respect to a spatial feature map to generate a posthoc coarse saliency map highlighting the areas of most importance for each of the network's outputs. Spatial attention maps are another method for producing saliency maps. They are included in the forward pass of networks and are self-explanatory masks that multiply feature maps of a network to select the most important regions for a network's decision (Schlemper et al., 2019). Mimicking humans is one of the motivations for their use (Xu et al., 2015).

© 2022 R.B. Lanfredi, A. Arora, T. Drew, J.D. Schroeder & T. Tasdizen.

We propose to quantify[1] the similarity between human attention and the saliency maps produced by these methods.

We use the REFLACX dataset (Bigolin Lanfredi et al., 2021; Goldberger et al., 2000; Lanfredi et al., 2021), which focuses on chest x-rays (CXRs), to build eye-tracking (ET) maps from radiologists' gazes and compare them with saliency maps from abnormality classification models. Figure 1 shows examples of ET maps and generated saliency maps. Differences are expected between them. Whereas a radiologist looks at multiple locations to inspect for abnormalities, interpretability methods are expected to highlight the areas where changes would cause a large impact on the output, i.e., abnormalities.

There are also reasons to believe humans and models might produce similar heatmaps. The low resolution of the Grad-CAM method should provide smooth saliency maps, similar to the ET maps. When comparing to human heatmaps, Ebrahimpour et al. (2019) showed the superiority of the similar class activation map (CAM) (Zhou et al., 2016) method. Since Grad-CAM provides one saliency map per class, we empirically test a few methods of combining the class saliency map into a single saliency map. Attention maps may be similar to ET maps since they are intrinsically class-independent and because of their human attention inspiration.

To better understand the expected range of similarity values, we calculate an upper bound by checking interobserver agreement and a lower bound by using a binary segmentation of the lung region. Because of the close proximity of the two bounds in traditional metrics, we use shuffled metrics to correct for center biases (Bylinskii et al., 2019). There is a tendency for fixations, i.e., image locations gazed by radiologists, to be in central regions of the images, and saliency maps that concentrate in these regions, independently of image content, achieve high scores. Shuffled metrics try to fix this problem and are formulated so that differences and similarities between heatmaps have different weights on the final scores depending on how commonly gazed their locations are. Given the structural similarity of CXRs, we calculate a specific center bias for this task, as shown in Figure 1(d). Finally, we evaluate the generated saliency map, reaching scores comparable to the interobserver agreement in one of the metrics. We further divide the dataset into subsets to understand which ones have a higher or lower similarity. We show that similarities are higher for abnormal CXRs and, in a more detailed subdivision, for parenchymal and pleural abnormalities, depending on the saliency map being evaluated.

1.1. Related work

In the field of interpretability, a few works have used ET maps to evaluate an explanatory saliency map (Ebrahimpour et al., 2019; Muddamsetty et al., 2020; Trokielewicz et al., 2019). Ebrahimpour et al. (2019) collected ET maps from participants listing objects present in natural images. They compared the data against the interpretability saliency maps for object-detection models, using the saliency map for the class with the highest score. In the field of medical images, Trokielewicz et al. (2019) compared the Grad-CAM (Selvaraju et al., 2017) saliency maps against humans in the task of iris recognition. Muddamsetty et al. (2020) did similar work for classification tasks of retinal images. These works analyzed only binary medical tasks and did not evaluate a strong baseline related to the biases present

1. Code at https://github.com/ricbl/etsaliencymaps

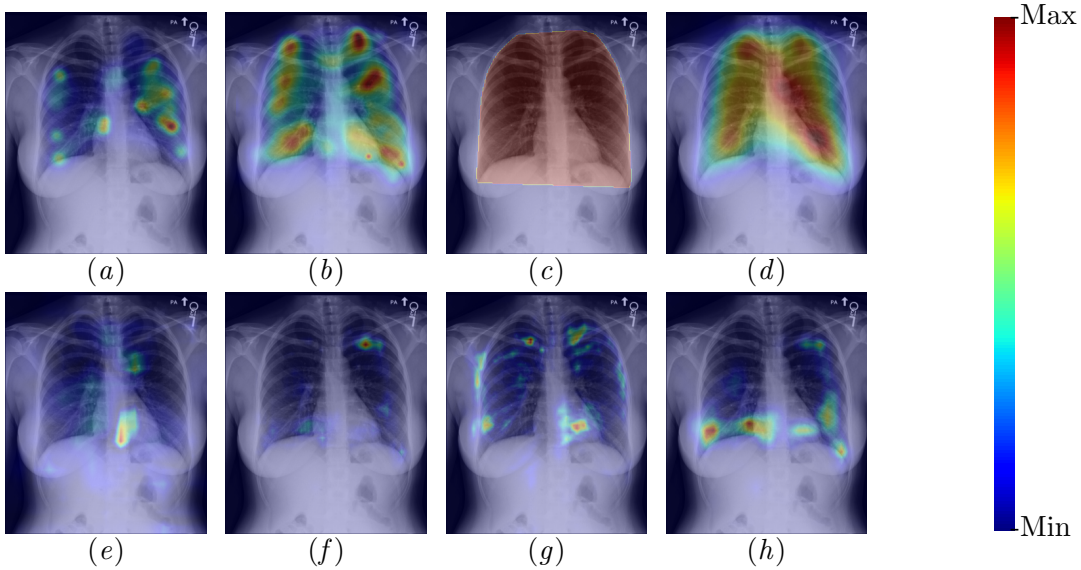

Figure 1: Examples of heatmaps over the respective CXR, and the heatmaps' color bar. a) ET map of a radiologist; b) average ET map of the remaining four radiologists; c) segmentation of the lung region used as a baseline d) average ET map from all CXRs, center bias (CB), after registration to match the location of the lungs; e) saliency map for model without attention gates (woAG) generated by Grad-CAM with uniform weights; f) saliency map for model with attention gates (wAG) generated by Grad-CAM with uniform weights; g) Attention map 1 (AM1) from wAG; h) Attention map 2 (AM2) from wAG.

in eye-tracking data. Karargyris et al. (2021) qualitatively checked the Grad-CAM saliency maps against ET maps in CXRs, but no quantitative analysis was performed. To the best of our knowledge, our study is the first to perform this quantitative analysis on CXRs.

The field of automatic generation of ET maps uses eye-tracking data as ground-truth and training data (Bylinskii et al., 2019). We employ the same comparison metrics as this field, but we do not focus on generating a saliency map that best matches ET maps.

2. Methods

2.1. Grad-CAM

Grad-CAM (Selvaraju et al., 2017) generates a saliency map for each class through the combination of the last spatial feature maps (LSFMs) of a network and the gradient of the network outputs with respect to each element of the LSFMs. The saliency map for each class c is produced by

$$GC_c = ReLU\left(\sum_k \alpha_c^k LSFM^k\right), \alpha_c^k = GAP\left(\frac{\partial logit(x)_c}{\partial LSFM^k}\right), \tag{1}$$

where GC_c is the saliency map provided by Grad-CAM for class c, $ReLU$ is a rectified linear unit, α_c^k is a weight for channel k of the last spatial layer of a network, GAP is global average pooling, $logit(x)_c$ is the logit output for the model being evaluated for class c, and $LSFM^k$ are the activations for channel k of the last spatial feature maps of a network. To combine the GC_c from all classes, we use

$$\frac{1}{\sum_c \psi_c} \sum_c \psi_c \times GC_c, \tag{2}$$

where ψ_c is a weight for the saliency map of each class. We consider three ways of choosing the weights ψ_c to mix the GC_c for each class c:

- **Thresholded**: uniformly mix the classes that are considered present in the image based on a threshold on the model's output, according to

$$\psi_c = \begin{cases} 1, \text{if } logit(x)_c > 0 \\ 0, \text{if } logit(x)_c < 0 \end{cases} . \tag{3}$$

  If all ψ_c are 0 for an image, we assign $\psi_c = 1$ for the "No Finding" label.

- **Weighted**: weight the classes using the output of the model, according to $\psi_c = \sigma(y_c)$, where σ is the sigmoid function.

- **Uniform**: uniformly mix all the classes: $\psi_c = 1$.

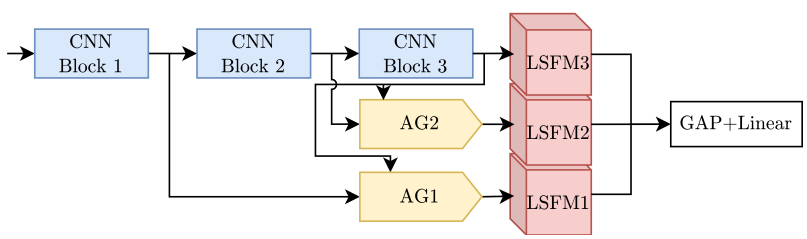

Figure 2: Overall structure of the network with attention gates (wAG). AG stands for attention gate. The network without attention gates (woAG) follows a similar architecture, with only LSFM3 input to the GAP operation. LSFM represents the activation maps used for calculating the Grad-CAM saliency map.

2.2. Attention gates

Figure 2 shows the architecture of the gated convolutional neural network (CNN), including the location of the attention gates. Figure 3 shows the employed attention gates. Each attention gate provides a saliency map through its attention map, and the attention maps can also be combined into a single saliency map through the use of Grad-CAM. Each channel from the LSFMs on Figure 2 is considered as one of the k channels from Equation (1). The output from CNN Block 3 in Figure 2 is directed to the attention gates before being called LSFM3, so that the gradient calculation from Equation (1) is influenced by LSFM3 in only one of the three CNN branches.

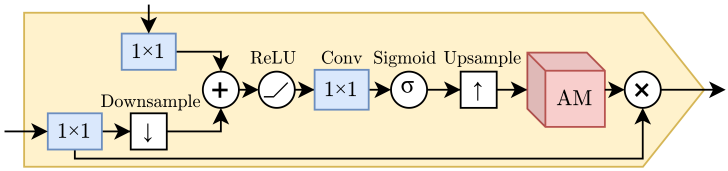

Figure 3: Operations inside the attention gate (AG). AM stands for attention map.

3. Experiments

Experiments employed CXRs from the MIMIC-CXR-JPG dataset (Goldberger et al., 2000; Johnson et al., 2019a,b) and eye-tracking data from the REFLACX dataset (Bigolin Lanfredi et al., 2021; Goldberger et al., 2000), which contains data collected while radiologists dictated reports. We used the parts of the REFLACX dataset where the same CXRs contained eye-tracking data for all five radiologists, totaling 91 CXRs. For each of the readings, one heatmap was generated from the parsed fixations.

3.1. Classification models

We trained two types of models, with (wAG) and without attention gates (woAG). They were trained and validated using the MIMIC-CXR-JPG dataset. For each type, we trained five models to calculate the variability of results, which are reported with their standard deviations. On the classification test set, woAG had an area under the receiver operating characteristic curve (AUC), averaged over the 14 labels, of 0.774 ± 0.003, whereas wAG had an average AUC of 0.769 ± 0.004. More details about the training process is presented in Appendix A.

3.2. Metrics and baselines

ET maps were generated by drawing Gaussians centered in each fixation and combining them through a sum weighted by the fixation duration. Following Le Meur and Baccino (2012), the Gaussians had a standard deviation of 1 degree of visual angle in each axis to represent location uncertainties for each fixation.

From the literature of automatic generation of human saliency maps (Bylinskii et al., 2019), we selected two metrics to compare saliency maps: normalized cross-correlation (NCC), i.e., Pearson's correlation coefficient, and the Borji formulation of AUC (Borji et al., 2013). The NCC was directly calculated between ET map and generated saliency map. We used a smooth formulation of AUC, sampling locations from smoothed heatmaps. For each sampled location, we picked the value of the generated saliency map at that location as one of the scores of the classifier evaluated by AUC. The locations for positive examples were sampled from a normalized ET map and for negative examples from the uniform distribution. We sampled 1000 positive and 1000 negative locations. To reduce the variability of the scores (Azam et al., 2016) and use the fact that CXRs had ET maps for five radiologists, we calculated the metrics against the average ET map of all combinations of four radiologists (1 vs. 4), which were considered the ground truth.

To have an upper bound for the metrics, we measured interobserver scores. To find a lower bound, we used an algorithm to segment the lungs and calculated the convex hull of the segmentations to include the mediastinum and the bilateral hemidiaphragms. Scores for the baselines are presented in Table 1. Upper and lower bounds were practically the same for NCC. This small range was probably caused by a strong bias toward having fixations around the lung area. To correct this bias, we used shuffled metrics (Bylinskii et al., 2019). For the NCC calculation, we drew from a closely related metric (Bylinskii et al., 2019), normalized scanpath saliency (NSS), and used the formulation from Gide and Karam (2016), resulting in

$$sNCC(GT, SM) = NCC(GT, SM) - NCC(CB, SM), \tag{4}$$

where sNCC is the shuffled NCC, GT is the ground truth saliency map, SM is the saliency map being evaluated, and CB is a heatmap representing the center bias in the dataset. For the shuffled AUC (sAUC), we sampled the locations for the negative examples from CB instead of uniformly.

Table 1: Scores for the tested methods of generating saliency maps (SM). Averages and standard deviations are calculated using 455 individual scores for the baselines and 2275 for the models. We highlight in bold the highest-scoring saliency map for each metric, excluding the interobserver upper bound.

SM	ψ_c	**NCC**	**AUC**	**sNCC**	**sAUC**
Interobserver	(Baseline)	0.632±0.126	0.790±0.042	0.028±0.128	0.558±0.051
Segmentation	(Baseline)	**0.637±0.107**	**0.735±0.046**	-0.187±0.086	0.505±0.026
Grad-CAM	Thresholded	0.252±0.253	0.596±0.109	-0.035±0.112	0.510±0.043
(woAG)	Weighted	0.408±0.191	0.683±0.079	-0.060±0.106	0.521±0.046
	Uniform	0.437±0.166	0.696±0.070	-0.067±0.104	0.522±0.044
Grad-CAM	Thresholded	0.194±0.174	0.583±0.074	-0.002±0.105	0.512±0.036
(wAG)	Weighted	0.299±0.173	0.672±0.064	0.027±0.123	0.528±0.044
	Uniform	0.343±0.154	0.678±0.066	**0.029±0.128**	**0.529±0.043**
AM1	-	0.254±0.141	0.672±0.075	-0.032±0.112	0.514±0.044
AM2	-	0.359±0.160	0.684±0.064	-0.007±0.132	0.522±0.052

We used bounding box annotations for lungs and heart in the calculation of the center bias in our dataset. We calculated the average bounding box, registered all bounding boxes to the average, applied the same transformation to the ET map, and combined them to get an average of the fixations. The resulting center bias is shown in Figure 1(d). For use in the metrics, the center bias is transformed to match the bounding box location of the respective CXR.

The results of the shuffled metrics for the baselines are presented in Table 1. The slightly positive value of the sNCC metric shows that the ET map from each radiologist is slightly more similar to the average of the other radiologists than to the center bias. The more extensive range between upper and lower bound baselines shows that considering the center bias is essential for calculating a meaningful score.

3.3. Results and discussion

Table 1 reports the metrics for all the tested methods. Not considering the baselines, the woAG model had the highest scores for the non-shuffled metrics and the wAG model for the shuffled metrics. In both cases, Grad-CAM with uniform ψ_c had the highest score. Uniform ψ_c might have achieved the best results because radiologists have to look for all abnormalities, including those not found in a particular image. Considering the sNCC metric, one of the models reached scores almost identical to the interobserver evaluation. For the AUC metrics, the interobserver evaluations had the highest scores with a good margin, highlighting that each metric measures different qualities of the heatmaps. Although the attention maps were not the highest scoring saliency maps, the Grad-CAM method had the highest shuffled scores when applied to the wAG model, showing a potential advantage of attention-gated models when compared to human attention.

We also analyzed scores after splitting normal and abnormal cases. Abnormal cases had a majority of radiologists selecting at least one abnormality for the image. As shown in Table 2, interobserver scores and the scores from a chosen interpretability method were higher for abnormal CXRs. This difference might have been caused by normal cases not having an evident area of interest and abnormality locations being areas of longer fixations by radiologists and stronger saliency for Grad-CAM. The scores for the segmentation baseline showed almost no change.

Table 2: Scores of baselines and of the Grad-CAM (wAG) with uniform ψ_c method when splitting the dataset into normal (N) and abnormal (Abn) CXRs. In parenthesis, we provide the number of 1 vs. 4 comparisons used to calculate each average.

Metric	**Label**	**Interobserver (IO)**	**Segmentation**	**Grad-CAM (wAG)**
sNCC	N	-0.056±0.113 (85)	-0.173±0.069 (85)	-0.047±0.103 (425)
sNCC	Abn	0.048±0.123 (370)	-0.191±0.090 (370)	0.045±0.128 (1850)
sAUC	N	0.532±0.044 (85)	0.504±0.016 (85)	0.497±0.031 (425)
sAUC	Abn	0.566±0.050 (370)	0.507±0.027 (370)	0.538±0.042 (1850)

To further understand our results, we separated the labels of the REFLACX dataset into three types: parenchymal, pleural and cardiomediastinal abnormalities. Parenchymal abnormalities involve lung tissue and can be located anywhere inside the lungs on frontal CXRs. Pleural abnormalities involve the membrane enclosing the lungs. On a frontal CXR, they are more commonly located near the lung apex or hemidiaphragms. Cardiomediastinal abnormalities are located between lungs, e.g., heart abnormalities. We grouped the labels to have a higher number of samples. Abnormality types were considered present when at least three radiologists selected at least one of the labels of the respective type.

Since the same CXR might be associated with more than one type of label, we calculated linear regression coefficients to evaluate how much each type of label was positively associated with the scores of each metric. We used the normal/abnormal label and the presence of each type of abnormality as independent variables, and the similarity scores as the dependent variable. Results are shown in Table 3. For the shuffled metrics, the intraob-

server agreement was stronger for pleural abnormalities, followed closely by parenchymal abnormalities. The scores for the model's saliency maps were positively correlated mainly with parenchymal abnormalities. Part of the differences was possibly caused by the penalty associated with shuffled metrics, because the area associated with some types of abnormalities might be fixated more often by radiologists in the average CXR. As seen in Table 3, when comparing the scores from shuffled and non-shuffled metrics, cardiomediastinal abnormalities were the most penalized for intraobserver agreement, and pleural abnormalities for the model's saliency maps. Part of the difference between types of labels might also be explained by the size of their visual evidence and how informative they are. Less ambiguous labels, such as "Enlarged cardiac silhouette" (the most common cardiomediastinal label), might have radiologists and models using fewer fixations to make their decision, leading to a higher variance in the location of the fixations. Location variance might also be increased for labels that occupy a large screen area, e.g., "Enlarged cardiac silhouette". The average areas and uncertainties for each label and type of labels are provided in Appendix B.

Table 3: Linear regression coefficients for each type of abnormality, for interobserver (IO) and Grad-CAM (wAG) with uniform ψ_c saliency maps (SM). Instead of standard deviations, we provide standard errors. Regression was performed with 455 points for IO scores and 2275 points for wAG.

Metric	**SM**	**Parenchymal**	**Pleural**	**Cardiomediastinal**
sNCC	IO	0.023±0.016	0.032±0.014	-0.021±0.013
sNCC	wAG	0.069±0.007	-0.017±0.006	-0.014±0.006
sAUC	IO	0.009±0.007	0.017±0.005	-0.013±0.005
sAUC	wAG	0.020±0.002	0.007±0.002	-0.006±0.002
NCC	IO	0.026±0.016	0.027±0.014	0.013±0.013
NCC	wAG	0.075±0.008	0.041±0.007	0.003±0.007
AUC	IO	0.006±0.005	0.016±0.004	0.004±0.004
AUC	wAG	0.030±0.004	0.034±0.003	0.003±0.003

4. Conclusion

Using a dataset of ET data from five radiologists, we showed that, when controlling for center bias, interpretability maps can be as similar to the ET maps from radiologists as ET maps from other radiologists. In other words, although the tested saliency maps are not good at highlighting areas fixated regularly in the average CXR, they excel at highlighting the specific areas in each CXR that radiologists fixate more than average. In our evaluation, the Grad-CAM method with uniformly weighted saliency maps of each class produced maps more similar to the ET maps. The attention-gated model produced saliency maps with the highest scores, and Grad-CAM outperformed the attention maps. Moreover, higher similarity scores were associated with the presence of abnormalities. Separating the dataset into types of labels showed that, for CXRs, saliency similarity potentially varies with abnormality size, ambiguity, and how much the location of an abnormality is commonly fixated.

Acknowledgments

Research reported in this publication was supported by the National Institute of Biomedical Imaging and Bioengineering of the National Institutes of Health under Award Number R21EB028367. The authors declare no competing interests. Christine Pickett provided copyediting support. Vivek Srikumar and Shireen Elhabian participated in discussions related to the project.

References

Shoaib Azam, Syed Omer Gilani, Moongu Jeon, Rehan Yousaf, and Jeong-Bae Kim. A benchmark of computational models of saliency to predict human fixations in videos. In Nadia Magnenat-Thalmann, Paul Richard, Lars Linsen, Alexandru C. Telea, Sebastiano Battiato, Francisco H. Imai, and José Braz, editors, *Proceedings of the 11th Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2016) - Volume 4: VISAPP, Rome, Italy, February 27-29, 2016*, pages 134–142. SciTePress, 2016. doi: 10.5220/0005678701340142. URL https://doi.org/10.5220/0005678701340142.

Christian F. Baumgartner, Konstantinos Kamnitsas, Jacqueline Matthew, Tara P. Fletcher, Sandra Smith, Lisa M. Koch, Bernhard Kainz, and Daniel Rueckert. Sononet: Real-time detection and localisation of fetal standard scan planes in freehand ultrasound. *IEEE Trans. Medical Imaging*, 36(11):2204–2215, 2017. doi: 10.1109/TMI.2017.2712367. URL https://doi.org/10.1109/TMI.2017.2712367.

Ricardo Bigolin Lanfredi, Mingyuan Zhang, William Auffermann, Jessica Chan, Phuong-Anh Duong, Vivek Srikumar, Trafton Drew, Joyce Schroeder, and Tolga Tasdizen. REFLACX: Reports and eye-tracking data for localization of abnormalities in chest x-rays, 2021. URL https://physionet.org/content/reflacx-xray-localization/1.0.0/.

Ali Borji, Dicky N. Sihite, and Laurent Itti. Quantitative analysis of human-model agreement in visual saliency modeling: A comparative study. *IEEE Trans. Image Process.*, 22(1):55–69, 2013. doi: 10.1109/TIP.2012.2210727. URL https://doi.org/10.1109/TIP.2012.2210727.

Zoya Bylinskii, Tilke Judd, Aude Oliva, Antonio Torralba, and Frédo Durand. What do different evaluation metrics tell us about saliency models? *IEEE Trans. Pattern Anal. Mach. Intell.*, 41(3):740–757, 2019. doi: 10.1109/TPAMI.2018.2815601. URL https://doi.org/10.1109/TPAMI.2018.2815601.

Mohammad K. Ebrahimpour, J. Benjamin Falandays, Samuel Spevack, and David C. Noelle. Do humans look where deep convolutional neural networks "attend"? In George Bebis, Richard Boyle, Bahram Parvin, Darko Koracin, Daniela Ushizima, Sek Chai, Shinjiro Sueda, Xin Lin, Aidong Lu, Daniel Thalmann, Chaoli Wang, and Panpan Xu, editors, *Advances in Visual Computing - 14th International Symposium on Visual Computing, ISVC 2019, Lake Tahoe, NV, USA, October 7-9, 2019, Proceedings, Part II*, volume 11845 of *Lecture Notes in Computer Science*, pages 53–65. Springer, 2019. doi: 10.1007/978-3-030-33723-0_5. URL https://doi.org/10.1007/978-3-030-33723-0_5.

Milind S. Gide and Lina J. Karam. A locally weighted fixation density-based metric for assessing the quality of visual saliency predictions. *IEEE Trans. Image Process.*, 25 (8):3852–3861, 2016. doi: 10.1109/TIP.2016.2577498. URL https://doi.org/10.1109/TIP.2016.2577498.

A. L. Goldberger, L. A. N. Amaral, L. Glass, J. M. Hausdorff, P. Ch. Ivanov, R. G. Mark, J. E. Mietus, G. B. Moody, C.-K. Peng, and H. E. Stanley. PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals. *Circulation*, 101(23):e215–e220, 2000. doi: 10.1161/01.CIR.101.23.e215.

Alistair Johnson, Matt Lungren, Yifan Peng, Zhiyong Lu, Roger Mark, Seth Berkowitz, and Steven Horng. MIMIC-CXR-JPG - chest radiographs with structured labels (version 2.0.0), 2019a. URL https://physionet.org/content/mimic-cxr-jpg/2.0.0/.

Alistair E. W. Johnson, Tom J. Pollard, Seth J. Berkowitz, Nathaniel R. Greenbaum, Matthew P. Lungren, Chih-ying Deng, Roger G. Mark, and Steven Horng. MIMIC-CXR-JPG: A large publicly available database of labeled chest radiographs. Eprint arXiv:1901.07042, 2019b.

Alexandros Karargyris, Satyananda Kashyap, Ismini Lourentzou, Joy T. Wu, Arjun Sharma, Matthew Tong, Shafiq Abedin, David Beymer, Vandana Mukherjee, Elizabeth A. Krupinski, and Mehdi Moradi. Creation and validation of a chest x-ray dataset with eye-tracking and report dictation for AI development. *Scientific data*, 8(1), December 2021. ISSN 2052-4463. doi: 10.1038/s41597-021-00863-5. Publisher Copyright: © 2021, The Author(s).

Christopher Kelly, Alan Karthikesalingam, Mustafa Suleyman, Greg Corrado, and Dominic King. Key challenges for delivering clinical impact with artificial intelligence. *BMC Medicine*, 2019. URL https://link.springer.com/article/10.1186/s12916-019-1426-2.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. URL http://arxiv.org/abs/1412.6980.

Ricardo Bigolin Lanfredi, Mingyuan Zhang, William F. Auffermann, Jessica Chan, Phuong-Anh T. Duong, Vivek Srikumar, Trafton Drew, Joyce D. Schroeder, and Tolga Tasdizen. Reflacx, a dataset of reports and eye-tracking data for localization of abnormalities in chest x-rays. Eprint arXiv:2109.14187, 2021.

Olivier Le Meur and Thierry Baccino. Methods for comparing scanpaths and saliency maps: strengths and weaknesses. *Behavior Research Methods*, pages 1–16, July 2012. doi: 10.3758/s13428-012-0226-9. URL https://hal.inria.fr/hal-00757615.

Satya M. Muddamsetty, Mohammad N. S. Jahromi, and Thomas B. Moeslund. Expert level evaluations for explainable AI (XAI) methods in the medical domain. In Alberto Del Bimbo, Rita Cucchiara, Stan Sclaroff, Giovanni Maria Farinella, Tao Mei, Marco Bertini,

Hugo Jair Escalante, and Roberto Vezzani, editors, *Pattern Recognition. ICPR International Workshops and Challenges - Virtual Event, January 10-15, 2021, Proceedings, Part III*, volume 12663 of *Lecture Notes in Computer Science*, pages 35–46. Springer, 2020. doi: 10.1007/978-3-030-68796-0_3. URL https://doi.org/10.1007/978-3-030-68796-0_3.

Mauricio Reyes, Raphael Meier, Sérgio Pereira, Carlos A Silva, Fried-Michael Dahlweid, Hendrik von Tengg-Kobligk, Ronald M Summers, and Roland Wiest. On the interpretability of artificial intelligence in radiology: Challenges and opportunities. *Radiology: Artificial Intelligence*, 2(3):e190043, 2020.

Jo Schlemper, Ozan Oktay, Michiel Schaap, Mattias P. Heinrich, Bernhard Kainz, Ben Glocker, and Daniel Rueckert. Attention gated networks: Learning to leverage salient regions in medical images. *Medical Image Anal.*, 53:197–207, 2019. doi: 10.1016/j.media.2019.01.012. URL https://doi.org/10.1016/j.media.2019.01.012.

Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-CAM: Visual explanations from deep networks via gradient-based localization. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pages 618–626. IEEE Computer Society, 2017. doi: 10.1109/ICCV.2017.74. URL https://doi.org/10.1109/ICCV.2017.74.

Mateusz Trokielewicz, Adam Czajka, and Piotr Maciejewicz. Perception of image features in post-mortem iris recognition: Humans vs machines. In *10th IEEE International Conference on Biometrics Theory, Applications and Systems, BTAS 2019, Tampa, FL, USA, September 23-26, 2019*, pages 1–8. IEEE, 2019. doi: 10.1109/BTAS46853.2019.9185980. URL https://doi.org/10.1109/BTAS46853.2019.9185980.

Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron C. Courville, Ruslan Salakhutdinov, Richard S. Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In Francis R. Bach and David M. Blei, editors, *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015*, volume 37 of *JMLR Workshop and Conference Proceedings*, pages 2048–2057. JMLR.org, 2015. URL http://proceedings.mlr.press/v37/xuc15.html.

Bolei Zhou, Aditya Khosla, Àgata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 2921–2929. IEEE Computer Society, 2016. doi: 10.1109/CVPR.2016.319. URL https://doi.org/10.1109/CVPR.2016.319.

Appendix A. Training details for the classification models

We used the Adam optimizer (Kingma and Ba, 2015) with a learning rate of 0.0001 and a weight decay of 0.00001. After three epochs without improvement on the validation average AUC, we multiplied the learning rate by 0.5. We trained with the binary cross-entropy loss for 75 epochs and with a batch size of 64. The 14 labels and the data split from the MIMIC-CXR-JPG dataset were used. All images from subjects displayed to radiologists

were moved to the test set. The training was limited to images filtered as follows: images without classification labels were discarded; only frontal CXRs were kept, i.e., images with "ViewPosition" metadata property equals to "AP" (anterior-posterior) or "PA" (posterior-anterior); and studies with more than one frontal image were excluded. The CNN blocks from Figure 2 were built following the Sononet-16 (Baumgartner et al., 2017) architecture, but with the modifications added by Schlemper et al. (2019) to include the attention gates.

For training, data were resized to have its shortest dimension equal to 224 pixels, rotated between -15 and +15 degrees, translated by up to 5% of its dimensions, scaled with a scale factor between 0.95 and 1.05, center cropped, randomly horizontally flipped, and normalized by the average intensities and standard deviation of the ImageNet dataset. For validation, images were scaled such that their longest dimension was a multiple of 16, and their shortest dimension the closest to 224 while keeping the aspect ratio. The image was then padded to a square. Saliency maps were generated with this padded version of the image and then cropped to the original aspect ratio.

Appendix B. Details of the grouping of labels

The REFLACX dataset provides ellipses locating most of the abnormalities labeled by radiologists. Each ellipse is also associated with a certainty label. We used the probabilities associated with each certainty label (10%, 25%, 50%, 75%, 90%) to calculate the Shannon's entropy of the associated binomial distribution (respectively, 0.47, 0.81, 1, 0.81, 0.47). We used the entropy as the level of uncertainty of the radiologist for that ellipse. The area of each ellipse was calculated in megapixels (MP). Table 4 shows the calculated statistics for the abnormality labels from the REFLACX dataset. As mentioned in Section 3.3, enlarged cardiac silhouette is one of the labels with the least uncertainty/ambiguity and highest area. Table 5 shows the statistics for the grouping of labels. The only label from the REFLACX dataset that was not considered an abnormality was "Quality issue". Cardiomediastinal abnormalities had the highest areas, while parenchymal abnormalities had the highest uncertainty and pleural abnormalities the smallest areas.

Table 4: List of abnormality labels from the REFLACX dataset, including their type, average area, uncertainty (entropy), and the number of ellipses used for the calculations.

Label	**Type**	**Area (MP)**	**Entropy**	**# ell.**
Abnormal mediastinal contour & Wide mediastinum	Cardiomediastinal	0.51	0.686	24
Airway wall thickening	Parenchymal	0.24	0.793	36
Atelectasis	Parenchymal	0.42	0.657	243
Consolidation	Parenchymal	0.49	0.620	194
Emphysema & High lung volume / emphysema	-	1.15	0.690	22
Enlarged cardiac silhouette	Cardiomediastinal	1.01	0.618	149
Enlarged hilum	Cardiomediastinal	0.22	0.621	7
Fracture & Acute Fracture	-	0.04	0.545	25
Groundglass opacity	Parenchymal	0.56	0.596	106
Hiatal hernia	-	0.18	0.811	1
Interstitial lung disease & Fibrosis	Parenchymal	0.38	0.752	20
Lung nodule or mass	Parenchymal	0.31	0.644	5
Mass	Parenchymal	0.14	0.469	1
Nodule	Parenchymal	0.02	0.598	23
Other	-	0.39	0.531	14
Pleural abnormality	Pleural	0.32	0.596	114
Pleural effusion	Pleural	0.40	0.633	112
Pleural thickening	Pleural	0.37	0.692	7
Pneumothorax	Pleural	0.22	0.636	26
Pulmonary edema	Parenchymal	0.50	0.708	138
Support devices	-	-	-	0

Table 5: Statistics of the location ellipses for groups of labels.

Type	**Area (MP)**	**Entropy**	**# ell.**
All Abnormalities	0.49±0.43	0.647±0.204	983
Cardiomediastinal	0.91±0.37	0.626±0.207	179
Parenchymal	0.42±0.35	0.667±0.201	536
Pleural	0.35±0.35	0.615±0.204	254