Do You See What I See? A Comparison of Radiologist Eye Gaze to Computer Vision Saliency Maps for Chest X-ray Classification

Jesse Kim, Helen Zhou*, Zachary Lipton*

Machine Learning Department

Carnegie Mellon University
School of Computer Science

MOTIVATION

Deep learning for automated diagnosis has shown promise in several medical domains.

Large public **chest X-ray datasets** have catalyzed the use of deep learning for chest X-ray classification.

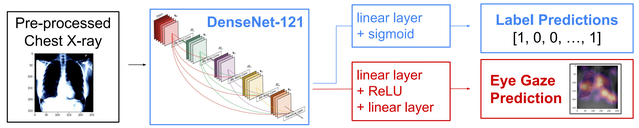
To interpret these models, studies often turn to **saliency maps**.

However, (to our knowledge) no studies have **systematically compared radiologist eye gaze to saliency maps** produced from deep learning models for chest X-ray classification.

METHODS

Datasets: (1) **MIMIC-CXR** Database v2.0.0[1,2]
(2) Eye Gaze Data for Chest X-Rays (**Eye-Gaze-CXR**)[3]

Models: DenseNet[4] and DenseNet-Aug



loss functions:

$$L_{\text{DN}} = BCE(y, \hat{y})$$

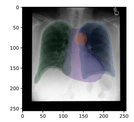
$$L_{\text{DN-Aug}} = \begin{cases} BCE(y, \hat{y}) + \lambda \cdot MSE(\hat{E}, E), & \text{if } E \text{ exists} \\ BCE(y, \hat{y}), & \text{otherwise} \end{cases}$$

Saliency Methods: gradient wrt input (SL), GradCAM (GC), DeepLift (DL), Layer Conductance (LC), Gradient SHAP (GS), SmoothGrad (NT)

Qualitative comparison of all six saliency methods vs. ground truth

Quantitative region-level comparison:
- determine a "most salient region" for each sample: take highest z-score (relative to other images) for the proportion of total saliency within each region
- Plot each class's distribution over "most salient regions."

Regions: aortic knob (A), right lung (R), left lung (L), mediastanum (M)



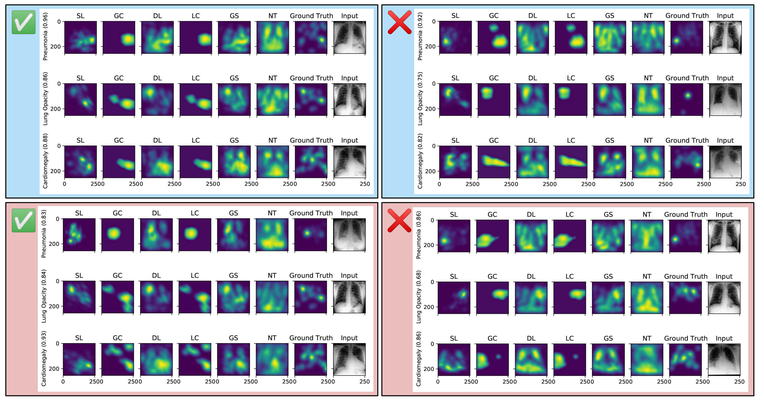
Quantitative image-level comparison: compute avg. structural similarity index (SSIM) score between model saliency maps & radiologist eye gaze among correct & incorrect preds.

RESULTS

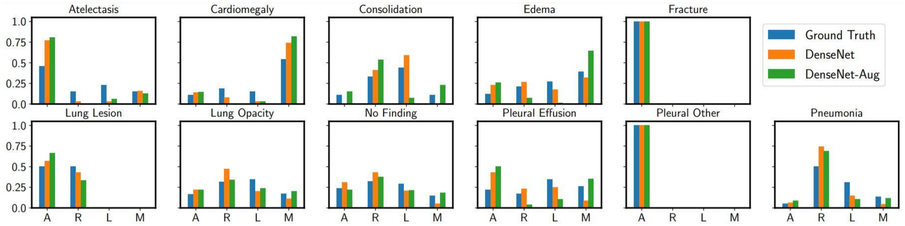
AUROCs

Label	DenseNet [4]	DenseNet-Aug
Atelectasis	0.759 (0.742 – 0.773)	0.751 (0.735 – 0.765)
Cardiomegaly	0.788 (0.774 – 0.801)	0.778 (0.766 – 0.792)
Consolidation	0.745 (0.719 – 0.770)	0.749 (0.722 – 0.776)
Edema	0.835 (0.821 – 0.848)	0.833 (0.821 – 0.846)
Enlarged CM	0.719 (0.682 – 0.753)	0.722 (0.685 – 0.758)
Fracture	0.676 (0.633 – 0.717)	0.680 (0.637 – 0.724)
Lung Lesion	0.737 (0.702 – 0.768)	0.727 (0.692 – 0.764)
Lung Opacity	0.694 (0.678 – 0.709)	0.697 (0.681 – 0.712)
No Finding	0.793 (0.776 – 0.808)	0.803 (0.788 – 0.817)
Pleural Eff.	0.888 (0.879 – 0.898)	0.884 (0.874 – 0.893)
Pleural Other	0.843 (0.810 – 0.874)	0.851 (0.824 – 0.877)
Pneumonia	0.711 (0.687 – 0.734)	0.713 (0.689 – 0.735)
Pneumothorax	0.832 (0.798 – 0.865)	0.816 (0.781 – 0.850)
Support Dev.	0.885 (0.875 – 0.895)	0.885 (0.875 – 0.894)

Saliency maps for DenseNet and DenseNet-Aug



Distribution over "most salient regions" for each class:



Average similarity (SSIM scores) of DenseNet and DenseNet-Aug saliency maps vs. radiologist eye gaze data, among "correct" (✅) and "incorrect" (❌) predictions:

Saliency method	Lung Opacity (n=98)		Pneumonia (n=80)		Cardiomegaly (n=53)		Edema (n=33)		Pleural Effusion (n=23)		Atelectasis (n=13)	
	✅	❌	✅	❌	✅	❌	✅	❌	✅	❌	✅	❌
SL	**0.450**	0.442	**0.465**	0.435	**0.398**	0.373	0.467	**0.490**	**0.378**	0.359	**0.427**	0.386
GC	**0.362**	0.332	**0.402**	0.381	0.351	**0.353**	0.383	**0.397**	**0.417**	0.394	**0.363**	0.285
DL	**0.402**	0.396	0.399	**0.403**	**0.407**	0.401	0.417	**0.457**	0.372	**0.373**	**0.388**	0.370
LC	**0.362**	0.332	**0.402**	0.381	0.351	**0.353**	0.383	**0.397**	**0.417**	0.394	**0.363**	0.285
GS	0.450	**0.451**	**0.423**	0.419	**0.441**	0.420	0.458	**0.486**	0.413	**0.436**	0.391	**0.410**
NT	**0.406**	**0.406**	0.358	**0.371**	**0.443**	0.428	0.433	**0.459**	0.456	**0.464**	0.397	**0.435**
SL	**0.465**	0.422	**0.485**	0.415	**0.373**	0.341	0.423	**0.462**	**0.395**	0.361	**0.489**	0.361
GC	**0.366**	0.336	**0.428**	0.381	**0.281**	0.266	0.362	**0.381**	0.341	**0.348**	**0.352**	0.322
DL	0.415	**0.419**	**0.421**	0.412	**0.400**	0.375	**0.417**	0.439	**0.420**	0.401	**0.433**	0.372
LC	**0.366**	0.336	**0.428**	0.381	**0.281**	0.266	0.362	**0.381**	0.341	**0.348**	**0.352**	0.322
GS	0.432	**0.443**	**0.434**	0.432	**0.441**	0.414	0.423	**0.450**	**0.446**	0.410	**0.427**	0.372
NT	0.391	**0.412**	0.335	**0.383**	**0.453**	0.450	0.421	**0.435**	0.435	**0.447**	0.406	**0.441**

DISCUSSION

Qualitative Saliency Map Analysis:
- Saliency methods can produce significantly different maps
  - NT is the most spread out. GC and LC are similar, possibly due to upsampling the last layer.
- Saliency maps for correct predictions focus on similar areas as ground truth
- Confidently incorrect saliency maps still bear some resemblance to ground truth eye gaze data (especially DenseNet-Aug). Two current hypotheses:
  - Certain regions are important for multiple labels
  - Model invariance (saliency sanity checks required)

Quantitative Saliency Region Level Analysis:
- General agreement between ground truth and model attribution distributions.
  - Region based attribution may lessen reliance on pixel-based explanations

Quantitative Saliency Map Analysis:
- SSIM score higher for DenseNet & DenseNet-Aug when prediction is correct (excluding No Finding, NT, & Edema)
  - Suggests shared spatial attribution information between saliency maps and eye gaze data.

Overall Model Performance:
- DenseNet & DenseNet-Aug have similar test set AUCs.
- Inconclusive if incorporating radiologist eye gaze data will improve predictive performance.

FUTURE WORK

Next Steps:
- Collect more eye gaze data
- Use attention mechanisms to improve DenseNet-Aug performance

Considerations:
- Is it desirable for saliency maps to mimic human gaze?
  - Eye gaze can locate important, robust features, but also suffers from central bias[5]
  - Models may pick up on important characteristics unintuitive to the human eye.
- Our work is purely comparative and does not necessarily promote saliency maps for interpretive purposes.

[1] Johnson, A. E., Pollard, T. J., Berkowitz, S. J., Greenbaum, N. R., Lungren, M. P., Deng, C.-y., Mark, R. G., and Horng, S. Mimic-CXR, a de-identified publicly available database of chest radiographs with free-text reports. Scientific Data, 6(1):1–8, 2019a.
[2] Goldberger, A., Amaral, L., Glass, L., Hausdorff, J., Ivanov, P., Mark, R., and Stanley, H. Physiobank, physiotoolkit, and physionet: Components of a new research resource for complex physiologic signals. pp. e215–e220, 2000. Circulation [Online] 101 (23).
[3] Karargyris, A., Kashyap, S., Lourentzou, I., Wu, J. T., Sharma, A., Tong, M., Abedin, S., Beymer, D., Mukherjee, V., Krupinski, E. A., et al. Creation and validation of a chest x-ray dataset with eye-tracking and report dictation for AI development. Scientific Data, 8(1): 1–18, 2021.
[4] Seyyed-Kalantari, L., Liu, G., McDermott, M., Chen, I. Y., and Ghassemi, M. Chexclusion: Fairness gaps in deep chest x-ray classifiers. In BIOCOMPUTING 2021: Proceedings of the Pacific Symposium, pp. 232–243. World Scientific, 2020
[5] Le Meur, O. and Baccino, T. Methods for comparing scan-paths and saliency maps: strengths and weaknesses. Behavior Research Methods, 45(1):251–266, 2013.

**equal contribution*