COMPUTER ASSISTED READING OF CHEST RADIOGRAPHS

Nandinee Fariah Haq, Z. Jane Wang

The University of British Columbia
Vancouver, Canada.

ABSTRACT

Chest radiographs or X-ray images are a common diagnostic tool to identify different thoracic diseases and other abnormal cardiopulmonary conditions. The advancements of artificial intelligence paves the way to machine learning based computer-assisted systems that can support the radiologists in disease diagnosis and report generation from chest radiographs. In this work we report an implementation of a deep-learning based framework to interpret the disease signature from chest X-rays. The model was trained on a large dataset consisting of both frontal and lateral X-ray images of the chest with multiple thoracic disease labels. We report a mean area under ROC curve (AUC) of 0.86, with the AUC of individual diseases in the range of 0.76 to 0.93. We also generated disease-level colormaps to visually present the X-ray image region most indicative of the disease.

Index Terms— X-rays, convolutional neural networks, densenet, activation maps

1. INTRODUCTION

Chest radiographs are the most frequently performed radiological examinations in clinical routines [1]. Radiographs or X-ray images are a common diagnostic tool to identify different thoracic diseases and other abnormal cardiopulmonary conditions [2]. With the advances in medical imaging technology the amount of radiology examinations ordered is increasing. However the number of radiologists to visually inspect and generate reports from the increasing amount of radiographs is inadequate [3, 4]. As a result there is a dramatic increase in radiologists' workloads [5] resulting in an increase in the total radiology turnaround time and reducing the overall quality of patient care owing to the delays in medical imaging interpretation [6, 7]. A computer-assisted system to analyse radiographs for primary screening have the potential to accelerate the radiologists' workflow and thereby improving the overall quality of healthcare.

With the advancements of artificial intelligence, computer-assisted systems were reported in literature to better aid the radiologists in disease diagnosis and report generation. Traditional feature extraction and machine learning based approaches were proposed for the diagnosis of thoracic diseases from chest X-rays. These approaches include textural or geometrical feature extraction and a statistical or machine learning based approach to classify the samples with the targeted disease from other samples [8, 9, 10]. However most of the studies target one disease and the features used in detection framework varies vastly across different diseases.

In recent years deep learning based frameworks have shown to generate promising results in various natural image recognition and detection tasks. The application of deep learning eliminates the necessity of handcrafted feature extraction thereby making it more applicable for scenarios where the existence of multiple diseases need to be detected simultaneously. However deep learning based automatic interpretation of chest X-ray images remain a challenging job due to the shortage of publicly available large scale medical datasets [11]. Earlier works in this research direction therefore mostly used networks pre-trained on other image databases to extract features [12, 13]. However medical images can be very different from other images and hence pre-trained networks might not be the proper feature extractor for them.

A very basic requirement in applying advanced machine learning frameworks to successfully interpret medical images is access to sufficient data. Recently two large scale dataset on chest X-ray images have been released publicly to facilitate further research on medical image interpretation [14, 15]. A few studies have been reported since the publication of the dataset in [14] that tackle the detection and diagnosis of thoracic diseases with a deep-learning based framework. Most of these studies focussed on detecting one disease from others [16], or implemented one deep learning based network per disease class to detect each diseases separately [14, 17], with a few taking into account the dependencies between disease labels [18, 19]. However the automatic labeller used for this dataset to extract disease labels from radiology reports later shown to perform weakly when applied to a more recent dataset in [15], where a more reliable labeller was proposed and the dataset and the labeller is since applied to more accurate detection framework in [20], which only focussed on detecting diseases from frontal chest X-rays and reported the best result with a framework that converts the multi-label detection problem using several single-label classifier. How-

Corresponding author: N.Haq (nandinee@ece.ubc.ca).

978-1-7281-2723-1/19/$31.00 ©2019 IEEE

ever, using single-label detectors in a multi-label settings assumes that the disease labels are independent, whereas in medical settings there are significant dependencies between the disease labels which should be taken into account when classifying multiple diseases simultaneously [18].

In this work, we implemented a deep learning based framework to facilitate the reading of chest radiographs. The model was trained on large chest X-ray dataset and was designed to simultaneously generate multi-label predictions of common thoracic diseases from both frontal and lateral chest radiographs. A visualization colormap is then generated to highlight the image region that is most indicative of the predicted diseases.

2. MATERIALS AND METHODS

2.1. Dataset

For this work we used the recently published CheXpert dataset [15], which is a large publicly available dataset of chest radiographs. The dataset contains 223,648 chest X-ray images among which 191,229 images are frontal chest X-rays and the rest 32,419 are lateral X-rays. The images are obtained between 2002 and 2017 from 64,740 unique patients, with 35,917 male and 28,822 female subjects. The disease labels were generated using an automatic rule-based labeler from their associated radiology reports. The labeler classified the labels as positive, negative and uncertain values, where uncertain label is assigned when it has no positive mentions and atleast one uncertain mention in the associated radiologist reports. For this work we used nine disease labels based on their prevalence in the dataset. The disease labels and their number of samples is reported in Table. 1. Table. 2 reports the number of frontal and lateral images in the dataset for each disease positive samples. Fig. 1 shows a few frontal posterior-anterior (PA) images from the dataset with different disease labels.

Table 1: Number of cases from different disease types with positive, negative and uncertain labels.

Disease label	Positive	Negative	Uncertain
Atelectasis	33,456	156,453	33,739
Cardiomegaly	27,068	188,493	8,087
Consolidation	14,816	181,090	27,742
Edema	52,291	158,373	12,984
Lung Opacity	105,707	112,343	5,598
Pneumonia	6,047	198,831	18,770
Pneumothorax	19,456	201,047	3,145
Pleural Effusion	86,254	125,766	11,628

Table 2: Number of frontal/lateral images per disease-positive samples.

Disease label	Frontal	Lateral
Atelectasis	29,795	3,661
Cardiomegaly	23,451	3,617
Consolidation	13,015	1,801
Edema	49,717	2,574
Lung Opacity	94,328	11,379
Pneumonia	4,683	1,364
Pneumothorax	17,700	1,756
Pleural Effusion	76,963	9,291

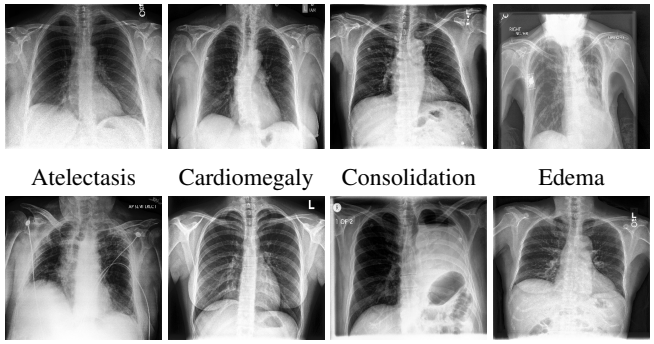


Fig. 1: Example of frontal posterior-anterior (PA) images from the dataset for different diseases.

2.2. Method

Our framework to classify the disease labels consists of three parts: data preprocessing, disease likelihood generation and disease localization. At first, the images were downsampled to a manageable size of 320×320 pixels and normalized to a range of [0,1]. The dataset was then passed through a data augmentation block that horizontally flips the X-ray images to generate more data. A deep convolutional neural network is then trained on the augmented data to generate disease likelihood from the X-ray images.

In this work we considered a model proposed in [21] which consists of densely connected convolutional layer blocks, known as DenseNet. Within the dense blocks each convolutional layer has a feed-forward connection to every other layers. We have used a DenseNet model with four dense blocks and four convolutional blocks. The structure of the network is shown in Fig.2. The output layer consists of eight fully-connected dense layer. To allow multi-label classification sigmoid function is used as the activation function in the output layer. The initial weights of the last fully connected layer was generated by the scheme proposed in [22]. The initial weights for other layers are assigned from a model pretrained on ImageNet dataset [23].

The dataset is multi-label and has uncertain labels for one

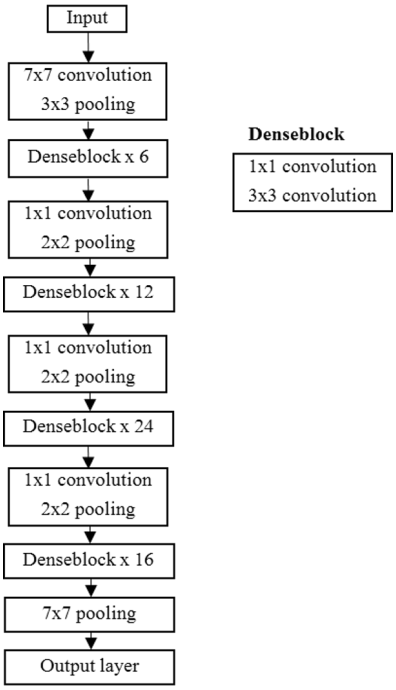


Fig. 2: Deep network architecture.

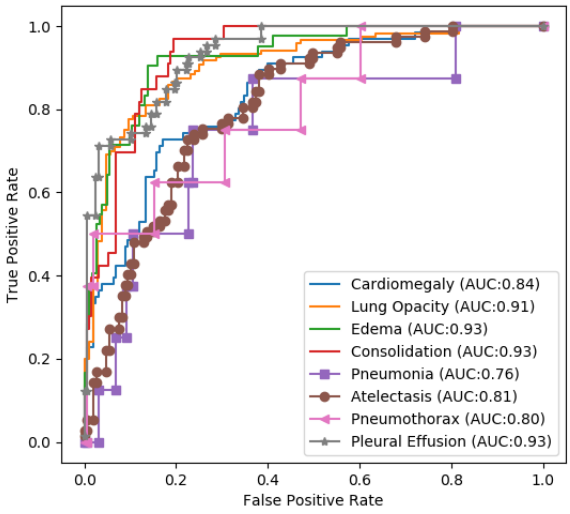


Fig. 3: Receiver operating characteristic curve for the .

or more diseases to a lot of samples, and if all the samples with one or more uncertain labels are removed, the dataset size decreases as well which is not ideal for training a deep network. Hence here we have included all the samples to train the network and ignored the uncertain labels while updating the gradient during training. It is also a highly imbalanced dataset, with different number of samples for each diseases, and almost 2 to 14 times more negative samples that the positive ones, as can be seen from Table. 1. To handle the class imbalance, we optimized the following weighted binary cross-entropy function as a loss function during training:

$$\mathcal{L}_{X|y} = \sum_{y\in\{0,1\}}^{i} -w_{i+} \times y_i log(\bar{y}_i) - w_{i-} \times (1-y_i) log(1-\bar{y}_i) \tag{1}$$

$$where, \quad w_{i+} = \frac{|n_{i-}|}{N}; \quad w_{i-} = \frac{|n_{i+}|}{N}$$

where $\mathcal{L}_{X|y}$ is the loss term for sample X with the label y, y_i is the ground truth label for disease class-i and $\bar{y}_i$ is the predicted likelihoo. $|n_{i+}|$ and $|n_{i-}|$ are the total number of positive and negative samples respectively for class-i and N is the total number of samples.

Finally, we generated disease-level colormaps to visualize the location of the diseases predicted by the network. To generate the colormaps, we incorporated a technique proposed in [24]. Let for a given image $I(x, y)$, $\mathcal{X}_m(x, y)$ is the m-th feature map of the last convolutional layer and ω_i^m is the weight term corresponding to class-i for the m-th feature map at the last classification layer. Then the i-th disease-level colormap for image $I(x, y)$ is generated as:

$$\mathcal{C}_{i|I(x,y)} = \sum_m \omega_i^m \times \mathcal{X}_m(x, y) \tag{2}$$

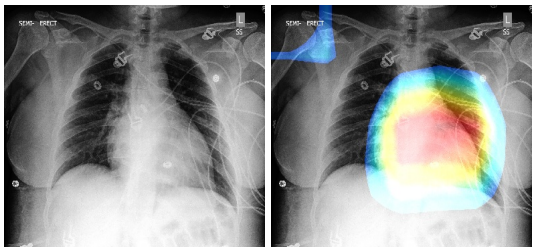
The generated disease-level colormap is then upsampled to the dimension of the original image and overlayed on the image for visualization.

3. RESULTS

We trained the deep network model with the with a batch size of 32 for 20 epochs. The initial learning rate was 10^{-4} and the learning rate was reduced by a factor of 10 each time a plateau was reached in the training loss. The network was trained with the Adam optimizer with $\beta_1 = 0.9$ and $\beta_2 = 0.999$. The network was trained on a Titan X GPU.

The receiver operating characteristic (ROC) curves for the different disease labels on the test dataset is shown in Fig. 3. The trained model achieved a mean area under ROC (AUC) curve of 0.86. The best performing labels were Consolidation (AUC 0.93), Pleural Effusion (AUC 0.93), Edema (0.93) and Opacity (AUC 0.91). The worst performing label was Pneumonia, with an AUROC of 0.76. This is partly because pneumonia is not entirely a radiologic diagnosis, rather it is a clinical diagnosis where other clinical information are incorporated to make a decision. However in this work only radiographs are used to generate the predictions.

We also compared the performance of the implemented framework with results reported in [20]. Table. 3 shows the comparison between the AUCs of the two approaches. As



(a) (b)

Fig. 4: Disease level colormap for a patient with ground truth label of Cardiomegaly. (a) The original X-ray image. (b) Generated Cardiomegaly colormap showing area indicative of the disease.

can be seen, our implementation outperforms the literature-based approach for most of the diseases. The mean AUC for the literature based approach on these eight diseases is 0.82, whereas we report an AUC of 0.86 on detecting these eight diseases.

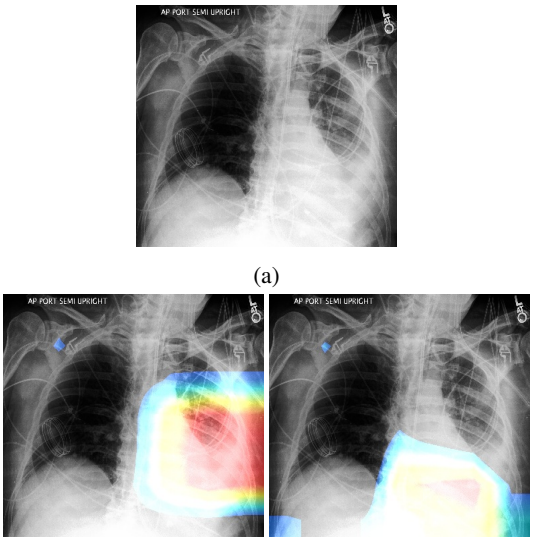
Table 3: Comparison with the state-of-art method on CheXpert dataset.

Disease label	Allaouzi *et al.* [20]	Our Implementation
Atelectasis	0.72	**0.81**
Cardiomegaly	0.88	**0.84**
Consolidation	0.77	**0.93**
Edema	0.87	**0.93**
Lung Opacity	0.76	**0.91**
Pneumonia	0.79	**0.76**
Pneumothorax	0.86	**0.80**
Pleural Effusion	0.90	**0.93**

We also generated disease-level colormaps to visualize the areas of the image that generated the class predictions. Fig. 4 shows an example of the disease-level colormaps where the image sample has a single positive disease label-cardiomegaly. The generated cardiomegaly-colormap shows the image region that influenced the labelling decision of the deep network. Fig. 5 shows another test image where the ground truth positive labels were pleural effusion and lung opacity. Visual inspection reveals opacities in the left lung. The generated pleural effusion colromap and lung opacity colormap both highlights the area around the left lung.

4. CONCLUSION

In this work we implemented a framework to extract disease-level colormaps from chest radiographs. We trained a deep convolutional neural network on a recently published dataset that consists of lateral and frontal chest X-ray images from



(a)

(b) (c)

Fig. 5: Disease level colormap for a patient with ground truth diagnosis of pleural effusion and lung opacity. (a) The original X-ray image. (b) Generated Pleural Effusion colormap. (c) Generated Lung Opecity colormap

a large number of subjects. The trained model utilizes a multi-label classification setting thereby incorporating the dependencies between disease labels. By training a DenseNet model on eight most prevalent diseases on the dataset we have achieved a mean AUC of 0.86, with individual disease AUCs ranging from 0.76-0.93. We then generated disease-level colormaps to visualize the area of the image that mostly influenced the prediction of the diseases. The colormaps highlight the area that should be paid more attention to while generating radiology reports, thus it can help towards better radiologic reading of chest X-rays. One possible limitation of the work is the absence of radiologists' input on the validation of the colormaps. In future we plan to implement a validation framework to include radiologists' annotation to verify the colormaps and improve the localization accuracy of the model.

5. REFERENCES

[1] Les R Folio, *Chest imaging: an algorithmic approach to learning*, Springer Science & Business Media, 2012.

[2] Alistair EW Johnson, Tom J Pollard, Seth Berkowitz, Nathaniel R Greenbaum, et al., "Mimic-cxr: A large publicly available database of labeled chest radiographs," *arXiv preprint arXiv:1901.07042*, 2019.

[3] Andrew B Rosenkrantz, Wenyi Wang, Danny R Hughes, and Richard Duszak Jr, "A county-level analysis of the us radiologist workforce: physician supply and subspe-

cialty characteristics," *Journal of the American College of Radiology*, vol. 15, no. 4, pp. 601–606, 2018.

[4] Farah S Ali, Samantha G Harrington, Stephen B Kennedy, and Sarwat Hussain, "Diagnostic radiology in liberia: a country report," *Journal of Global Radiology*, vol. 1, no. 2, pp. 6, 2015.

[5] Ahmed Hosny, Chintan Parmar, John Quackenbush, Lawrence H Schwartz, and Hugo JWL Aerts, "Artificial intelligence in radiology," *Nature Reviews Cancer*, vol. 18, no. 8, pp. 500, 2018.

[6] Sarah Bastawrous and Benjamin Carney, "Improving patient safety: avoiding unread imaging exams in the national va enterprise electronic health record," *Journal of digital imaging*, vol. 30, no. 3, pp. 309–313, 2017.

[7] Abi Rimmer, "Radiologist shortage leaves patient care at risk, warns royal college," *Bmj*, vol. 359, pp. j4683, 2017.

[8] KC Santosh and Sameer Antani, "Automated chest x-ray screening: Can lung region symmetry help detect pulmonary abnormalities?," *IEEE transactions on medical imaging*, vol. 37, no. 5, pp. 1168–1177, 2017.

[9] Norliza Mohd Noor, OM Rijal, and Chang Yun Fah, "Wavelet as features for tuberculosis (mtb) using standard x-ray film images," in *6th International Conference on Signal Processing, 2002.* IEEE, 2002, vol. 2, pp. 1138–1141.

[10] A Asuntha, A Brindha, S Indirani, and Andy Srinivasan, "Lung cancer detection using svm algorithm and optimization techniques," *J. Chem. Pharm. Sci*, 2016.

[11] Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, and Ronald M Summers, "Tienet: Text-image embedding network for common thorax disease classification and reporting in chest x-rays," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 9049–9058.

[12] Yaron Anavi et al., "Visualizing and enhancing a deep learning framework using patients age and gender for chest X-ray image retrieval," in *Medical Imaging*. SPIE, 2016, vol. 9785, p. 978510.

[13] Amit Shah et al., "Deeply learnt hashing forests for content based image retrieval in prostate MR images," in *Medical Imaging*. SPIE, 2016, vol. 9784, p. 978414.

[14] Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, et al., "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*, 2017, pp. 2097–2106.

[15] Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, et al., "Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison," *arXiv preprint arXiv:1901.07031*, 2019.

[16] Pranav Rajpurkar, Jeremy Irvin, Kaylie Zhu, Brandon Yang, et al., "Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning," *arXiv preprint arXiv:1711.05225*, 2017.

[17] Zhe Li, Chong Wang, Mei Han, Yuan Xue, et al., "Thoracic disease identification and localization with limited supervision," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 8290–8299.

[18] Li Yao, Eric Poblenz, Dmitry Dagunts, Ben Covington, et al., "Learning to diagnose from scratch by exploiting dependencies among labels," *arXiv preprint arXiv:1710.10501*, 2017.

[19] Qingji Guan, Yaping Huang, Zhun Zhong, Zhedong Zheng, et al., "Diagnose like a radiologist: Attention guided convolutional neural network for thorax disease classification," *arXiv preprint arXiv:1801.09927*, 2018.

[20] Imane Allaouzi and Mohamed Ben Ahmed, "A novel approach for multi-label chest x-ray classification of common thorax diseases," *IEEE Access*, vol. 7, pp. 64279–64288, 2019.

[21] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger, "Densely connected convolutional networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4700–4708.

[22] Xavier Glorot and Yoshua Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, 2010, pp. 249–256.

[23] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, et al., "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 2009, pp. 248–255.

[24] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba, "Learning deep features for discriminative localization," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2921–2929.