An Attention-Guided Deep Neural Network for Annotating Abnormalities in Chest X-ray Images: Visualization of Network Decision Basis*

Khadijeh Saednia, *Student Member, IEEE,* Ali Jalalifar, Shahin Ebrahimi, and Ali Sadeghi-Naini, *Senior Member, IEEE*

Abstract— **Despite the potential of deep convolutional neural networks for classification of thorax diseases from chest X-ray images, this task is still challenging as it is categorized as a weakly supervised learning problem, and deep neural networks in general suffer from a lack of interpretability. In this paper, a deep convolutional neural network framework with recurrent attention mechanism was investigated to annotate abnormalities in chest X-ray images. A modified MobileNet architecture was adapted in the framework for classification and the prediction difference analysis method was utilized to visualize the basis of network's decision on each image. A long short-term memory network was utilized as the attention model to focus on relevant regions of each image for classification. The framework was evaluated on NIH chest X-ray dataset. The attention-guided model versus the model with no attention mechanism could annotate the images in an independent test set with an F1-score of 0.58 versus 0.46, and an AUC of 0.94 versus 0.73. The obtained results implied that the proposed attention-guided model could outperform the other methods investigated previously for annotating the same dataset.**

Keywords— Deep convolutional neural network, Visualization method, Attention mechanism, Chest X-ray Annotation

I. INTRODUCTION

In recent years, machine learning in general and deep learning in particular has been the focus of intense research in the field of medical image analysis [1]. Deep Neural Networks (DNN), and especially convolutional neural networks (CNN), have turned into a methodology of choice in the field of medical image analysis [1]. Recently, these networks have been investigated in various applications such as image-guided medical event prediction, computer-aided detection, risk assessment, and automated image segmentation [1], [2].

One of the main drawbacks of DNNs is their lack of interpretability. Essentially, a DNN represents an extremely complex non-linear function, which makes it difficult to explain how a classification comes about. The lack of clarity and interpretability is a main obstacle for adaptation of DNN in applications related to the industry, government, and healthcare, in which the cost of a mistake is potentially so high [3]. As a result, clinicians usually hesitate to use these state-of-the-art intelligent tools in real-world clinical applications such as image classification, object segmentation, and disease characterization. Visualization of network decisions on medical images is one possible approach to address this challenge. Such visualization helps clinicians to understand better the basis of a network decision and potentially become more confident to adopt the technique. Computer scientists, on the other hand, can use this information to improve the accuracy of their networks. A number of methods have been proposed previously to visualize the feedback of a DNN to an input in terms of the areas or features that the network has more focus on. For example, the prediction difference analysis has been introduced to visualize areas in an image that contribute for or against a particular class [3].

Recently, attention-based models have demonstrated a remarkable ability for the problems of multiple-object localization and recognition, despite being trained only with labeled samples of each object during the training phase [4]. These models simulate the approach of human vision system in sequential object recognition, *e.g.* in reading, where the fovea moves continually from an object to the next relevant one, with the aim of recognizing a specific object, and adding it to the internal representation of the sequence. [5]. In [4], instead of a simple averaging over feature maps in a CNN architecture, weighted feature maps were learned in a recurrent network based on the class label probability vectors obtained in the previous time step, to suppress the irrelevant parts of an input. Development of methods for visualizing and explaining the basis of decision in DNNs, potentially facilitates a better understanding and evaluation of the network's attention mechanism and permits more interpretability in deep learning models [6].

Thoracic diseases are among most common causes of severe illness and death worldwide [7]. The chest X-ray radiography is the clinical routine for diagnosis of thoracic disorders, due to its accessibility and low cost [4]. Detecting many thoracic disorders from X-ray images is, however, challenging due to the projectional nature of X-ray radiography that increases the complexity of a diagnostic interpretation [8]. An automated end-to-end algorithm to

*This research was supported by Natural Sciences and Engineering Research Council (NSERC) of Canada, Terry Fox Foundation, and the Lotte and John Hecht Memorial Foundation.

K. Saednia, A. Jalalifar, and S. Ebrahimi are with the Department of Electrical Engineering and Computer Science, Lassonde School of Engineering, York University, Toronto, ON, Canada (e-mail: saednia@yorku.ca; alijfar@yorku.ca; eshahin@yorku.ca).

A. Sadeghi-Naini is with the Department of Electrical Engineering and Computer Science, Lassonde School of Engineering York University, Toronto, ON, Canada; also with the Department of Radiation Oncology and Physical Sciences Platform, Odette Cancer Centre and Sunnybrook Research Institute, Sunnybrook Health Sciences Centre, Toronto, ON, Canada; also with the Department of Medical Biophysics, University of Toronto, Toronto, ON, Canada (e-mail: asn@yorku.ca, phone: 416-736-2100 x20590).

978-1-7281-1990-8/20/$31.00 ©2020 IEEE

diagnose thoracic diseases reliably from chest X-rays is highly desirable in clinic [6].

In this paper, we investigated a deep CNN framework coupled with an attention mechanism to automatically annotate abnormalities in chest X-ray images accurately. A modified MobileNet architecture [9] was adapted for classification with a recurrent attention mechanism. The recurrent attention mechanism consisted of a recurrent neural network (RNN) to generate weighted feature maps based on the previous state of the network. The prediction difference analysis method was applied to visualize the basis of the network decision on each image. Obtained results demonstrated a considerably better performance of the attention-guided model compared to the model with no attention. The better performance of the attention-guided network was also evident in the visualization results of the network's decision basis.

II. Materials and Methods

A. Dataset

In this study, NIH chest X-ray dataset [10] was used to train and evaluate the proposed framework. The NIH Chest X-ray dataset consists of 112,120 anonymized frontal X-ray images with diagnosed disease labels from 30,805 patients. There are 15 classes (14 diseases, and one for "No findings") in this dataset. Images have been labeled with "No findings" or with one or more diagnosed diseases as detailed in Table 1. These labels were mined from the associated radiology reports, using natural language processing (NLP). The labels are expected to be >90% accurate and suitable for weakly-supervised learning [10]. Table 1 summarizes the statistics of the NIH X-ray Dataset. Out of the 15 classes in the dataset, 13 classes were used in this study. Images with a 'No finding' label were omitted because a main objective of the study was to visualize the basis of network decision for chest abnormalities. We also excluded images labeled with 'Hernia' as there were very few samples with this abnormality in the dataset. The dataset was randomly split into training (80%) and test (20%) sets at patient level. All images related to a patient was exclusively in the training or test set. 25% of the training data was used for model validation during the training process.

TABLE I - STATISTICS OF THE APPLIED DATASET

Disease	Number of Samples
Atelectasis	11559
Consolidation	4667
Infiltration	19894
Pneumothorax	5302
Edema	2303
Emphysema	2516
Fibrosis	1686
Effusion	13317
Pneumonia	1431
Pleural-thickening	3385
Cardiomegaly	2776
Nodule	6331
Mass	5782
Hernia	227
No finding	60361

B. Framework

The proposed framework (Fig. 1) consists of two branches. The first branch performs a crude classification and the second branch applies an attention mechanism to increase the contribution of relevant features in the final decision of the model. For training the model, the images of the training set and their corresponding label(s) were fed to the framework. The abnormality label(s) of each images was presented as a 13-bit vector demonstrating presence/absence of each of the 13 abnormalities. The network architectures have been explained in the following sections.

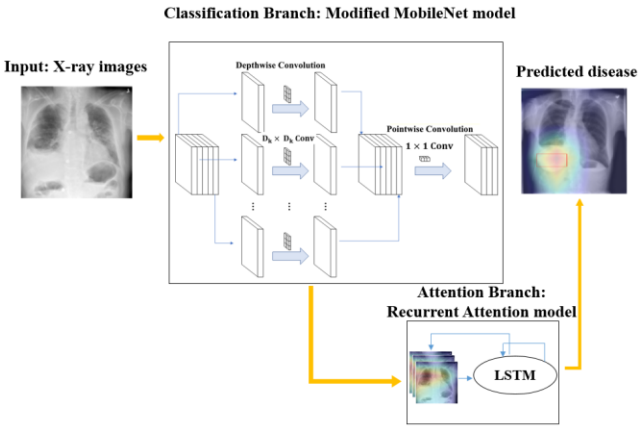

Fig. 1: Overview of the proposed attention-guided framework for annotation of abnormalities in chest X-ray images.

1) Classification Branch

A modified MobileNet architecture [7] was adapted for classification in the proposed method. This network is based on a streamlined architecture that uses depthwise separable convolutions. A depthwise separable convolution consists of two different operations: (1) depthwise convolution; (2) pointwise convolution. A standard convolution performs on the spatial dimension of the feature maps and on input/output channels. The depthwise convolution maps a single convolution filter on each input channel separately. The pointwise convolution with a 1×1 kernel size is utilized to merge the features created by the depthwise convolution [9].

The first layer of MobileNet is a full convolutional layer, and the others are composed of depthwise separable convolutional layers. All these layers are followed by a batchnorm and ReLU activation function. The last layer is a fully-connected layer with a softmax activation function for classification. In this work, the softmax function was replaced with a sigmoid activation function, and a global average pooling (GAP) layer was added before the fully-connected layer. For the attention-guided model (described in section D), the output (feature maps) of GAP layer along with the label vector were fed into the attention branch. The network in the classification branch was trained using the Adam optimizer with a categorical cross entropy loss function. A batch size of 1024 was used for the training process and the best model based on the lowest validation f1-score during the 10000 epochs of training was saved.

2) Attention Branch

Using the feature maps obtained from the GAP layer of the trained classification network, the correlation between the class (abnormality) labels and the abnormal regions in each image was explored by the attention branch.

Inspired by [11], a long short-term memory (LSTM) network was exploited to produce a probability vector of class labels at every time step from the network's previous hidden state, a class label probability vector generated in the previous step, and a context vector. At each time step, the context vector showed a dynamic representation of the relevant input part, in which higher value means higher relevance. The hidden activation of LSTM was a linear projection of the stochastic context vector followed by a tanh activation function. The feature maps generated in the classification network were fed into the attention model. Regions of the original image that provided more relevant information for classification with the feature maps, provided more contribution in the decision of attention branch by having higher weight values in the context vector. In simple term, the model paid more attention to the parts which were related to the correct class of abnormality. This mechanism was also applied to visualize the network's basis of decision, described in section C. The final prediction of the attention-guided framework was obtained by thresholding the final class label probability vector generated by the attention network. A categorical cross entropy loss function was applied with the Adam optimizer, a learning rate 0f 0.001, and an output unit size of 16 before the fully-connected layer for training the recurrent attention network. The validation f1-score was monitored as a criterion for saving the best model.

C. Visualization Methods

To demonstrate the effect of attention mechanisms and to permit a better interpretation of network decision, the prediction difference analysis (PDA) method was exploited. The PDA was used to find the pivotal pixels contributing to the decision of the network for each X-ray image [2]. The PDA method assigns a relevance value to each input feature in the feature map with respect to class c. The idea is to find the relevance of feature x_i assuming that it is unknown, and then to compare the network's prediction results, when x_i is known. In other words, the relevance is estimated by the difference between $p(c|x)$ and $p(c|x_{\setminus i})$, where $p(c|x_{\setminus i})$ is the prediction of the network without having feature x_i ($x_{\setminus i}$ shows all features except the i^{th} feature) [2]. In this study, the weights of GAP layer in modified MobileNet architecture were used as the input feature map of PDA method in the model without attention (described in section D). For the proposed attention-guided framework, these feature maps were passed to the recurrent attention network and the feature maps generated by trained attention network were processed with the PDA method for visualization.

D. Evaluation

Two different models were constructed and evaluated in this study for comparison: 1) a model without attention which used only the modified MobileNet architecture to annotate the images, 2) an attention-guided model that annotated the images using the the recurrent attention network trained on the feature maps of modified MobileNet. The performance of the models in annotating the X-ray images were evaluated, the F1-score, accuracy, binary accuracy, mean absolute error, and the area under receiver operating characteristic (ROC) curve (AUC) were calculated.

III. RESULTS

Fig. 2 shows the annotations made by the two networks for different patients along with the PDA heat maps showing the contribution of different regions to the decision made by each network. The results demonstrate that the attention mechanism guided the network to focus on the regions of image with more important information about the lung abnormalities.

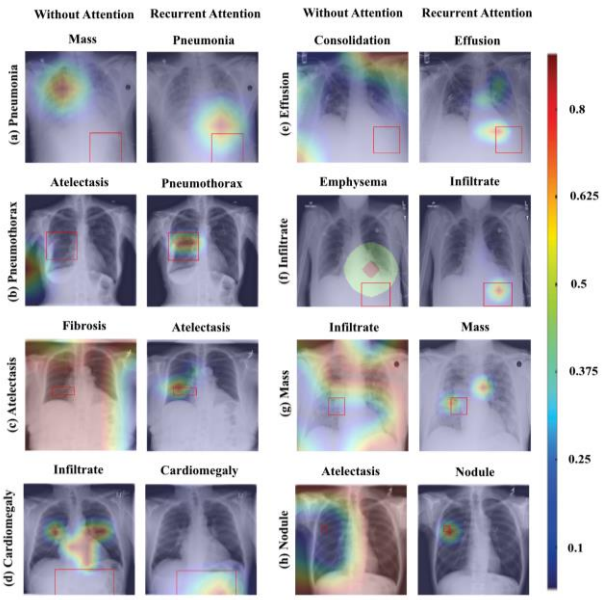

Fig. 2: The heat maps generated using the PDA for the models without and with recurrent attention. Eight pathologies in these figures are: (a) pneumonia, (b) pneumothorax, (c) atelectasis, (d) cardiomegaly, (e) effusion, (f) infiltration, (g) mass, and (h) nodule. The abnormality detected by each network was shown on top of the images. The red rectangles show the image areas used by the experts to diagnose the disorder . The color bar shows the level of contribution to the network decision (detected abnormality).

Fig. 3 depicts some of the predictions made by the attention-guided model. While the network detected the abnormality correctly in most of the cases, in some cases it was not able to pay attention to the correct parts of the image. Table 2 summarizes the classification results of the two models with different metrics including binary accuracy, F1-score, average area under the ROC curve (average AUC), and mean average error (MAE). The results demonstrate that the attention-guided model outperformed the model without attention considerably. Fig. 4 shows the confusion matrix for the attention-guided model. The performance of the network in detecting infiltration, effusion, and atelectasis was better compared to the other abnormalities. The better performance of the model in detecting these abnormalities can be due to the fact that there were a larger number of samples available for them in the dataset.

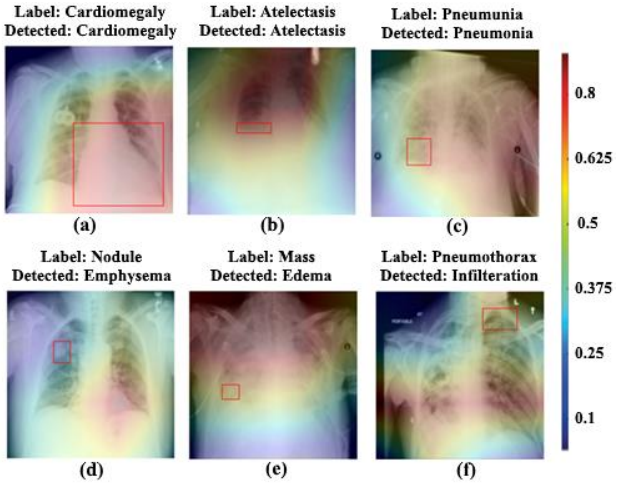

Fig. 3: The decision of the attention-guided network on representative images with different abnormalities. In (a) – (c) the network paid attention to the correct area of the images with abnormality and annotated them correctly, but in (d) – (f) the network attention was not on the regions with abnormality and the images were incorrectly annotated. The red rectangles show the image areas used by the experts to diagnose the disorder. The color bar shows the level of contribution to the network decision, *i.e.*, network's attention.

IV. DISCUSSION AND CONCLUSION

In this study, an attention-guided convolutional neural network framework was proposed for annotation of thoracic abnormalities in digital X-ray images. A modified MobileNet architecture was adapted as the base of the framework with an LSTM network to model the attention mechanism. The PDA method was applied to visualize the way the network makes decision on new images. The heat maps generated through the PDA showed which regions of the image had more contribution to the network decision. The recurrent attention mechanism used in this study guided the attention of the network step by step. This mechanism potentially improves the network at each step to make better decisions based on its previous performance on the training samples. The proposed framework was evaluated on the NIH chest X-ray dataset. The obtained results demonstrated that the attention-guided model considerably outperformed the model without an attention mechanism in annotating the x-ray images. The proposed attention-guided framework could annotate the images with an AUC of 0.94, showing a substantial improvement compared to the previous works. Previous studies reported an AUC of 0.78 [3], and 0.81 [11] for annotating abnormalities on the same dataset. The results obtained her is promising and encourage future studies to investigate the performance of other convolutional network architectures and attention mechanisms for detecting thoracic diseases using chest X-ray images.

TABLE II: SUMMARY OF THE CLASSIFICATION RESULTS.

Model	Binary Accuracy	F1-Score	Average AUC	MEA
Without Attention	0.85	0.46	0.73	0.23
Attention-guided	0.90	0.58	0.94	0.18

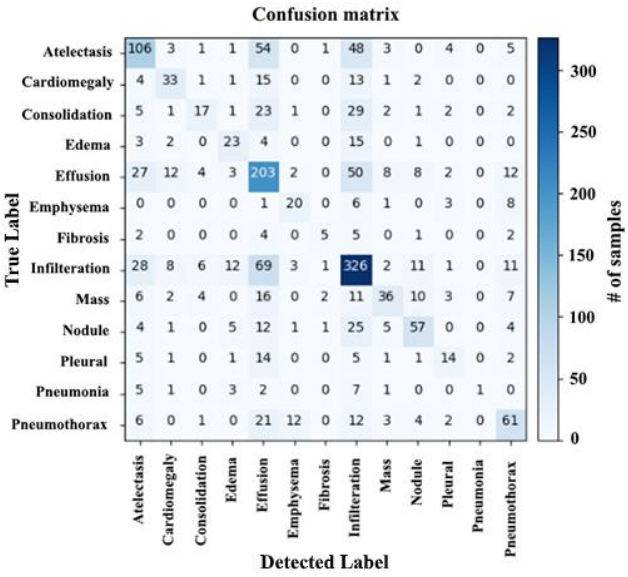

Fig. 4: Confusion Matrix summerizing the classifiction performance of the attention-guided model for the 13 classes.

REFERENCES

[1] Litjens, Geert, Thijs Kooi, Babak Ehteshami Bejnordi, Arnaud Arindra Adiyoso Setio, Francesco Ciompi, Mohsen Ghafoorian, Jeroen Awm Van Der Laak, Bram Van Ginneken, and Clara I. Sánchez. "A survey on deep learning in medical image analysis." Medical image analysis 42 (2017): 60-88.

[2] Zintgraf, Luisa M., Taco S. Cohen, Tameem Adel, and Max Welling. "Visualizing deep neural network decisions: Prediction difference analysis." arXiv preprint arXiv:1702.04595 (2017).

[3] Wang, Hongyu, and Yong Xia. "Chestnet: A deep neural network for classification of thoracic diseases on chest radiography." arXiv preprint arXiv:1807.03058 (2018).

[4] Ba, Jimmy, Volodymyr Mnih, and Koray Kavukcuoglu. "Multiple object recognition with visual attention." arXiv preprint arXiv:1412.7755 (2014).

[5] Ruuskanen, Olli, Elina Lahti, Lance C. Jennings, and David R. Murdoch. "Viral pneumonia." The Lancet 377, no. 9773 (2011): 1264-1275.

[6] Qin, Chunli, Demin Yao, Yonghong Shi, and Zhijian Song. "Computer-aided detection in chest radiography based on artificial intelligence: a survey." Biomedical engineering online 17, no. 1 (2018): 113.

[7] Howard, Andrew G., Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. "Mobilenets: Efficient convolutional neural networks for mobile vision applications." arXiv preprint arXiv:1704.04861 (2017).

[8] Zhou, Bolei, 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, pp. 2921-2929. 2016.

[9] Wang, Xiaosong, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M. Summers. "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, pp. 2097-2106. 2017.

[10] Xu, Kelvin, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. "Show, attend and tell: Neural image caption generation with visual attention." In International conference on machine learning, pp. 2048-2057. 2015.

[11] Guendel, S., et al., Learning to recognize abnormalities in chest x-rays with locationaware dense networks. arXiv preprint arXiv:1803.04565, 2018.