Automated Diagnosis and Visualization from X-Ray Imaging

Santoshmurti Daptardar[1], Sanyam Rajpal[1]
[1]Luddy School of Informatics, Computing and Engineering, Indiana University Bloomington, IN 47408

***Abstract*—Chest radiograph interpretation is critical for the detection of thoracic diseases, including tuberculosis and lung cancer, which affects millions of people globally each year. This time-consuming task typically requires expert radiologists to interpret the images, leading to fatigue-based diagnostic error and lack of diagnostic expertise in areas of the world where radiologists are not available. Recently, deep learning approaches have been able to achieve expert-level performance in medical image interpretation tasks, powered by large network architectures and fueled by the emergence of large labeled datasets. The purpose of this study is to investigate the performance of various deep learning techniques on the detection of pathologies in chest radiographs and to visualize these pathologies. Moreover, the visualization capabilities of CNNs have not been fully investigated. We test Grad-CAMs for visualizing pathologies in X-Rays and discuss them from a radiological perspective.**

I. INTRODUCTION

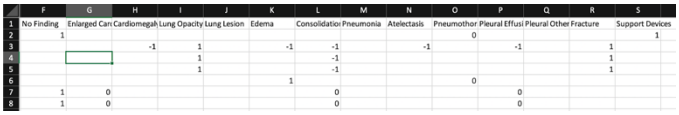
Chest radiography is the most common imaging examination globally, with over 2 billion procedures performed each year [4]. It is critical for screening, diagnosis, and management of many life-threatening diseases. Automated chest radiograph interpretation at the level of practicing radiologists could thus, provide substantial benefit in many clinical settings, from improved workflow prioritization and clinical decision support to large-scale screening and global population health initiatives.

Recent advancements in deep learning and large datasets have enabled algorithms to match the performance of medical professionals in a wide variety of other medical imaging tasks, including diabetic retinopathy detection [5], skin cancer classification [6], and lymph node metastases detection [7]. Automated diagnosis from chest imaging has received increasing attention [8],[9], with specialized algorithms developed for pulmonary tuberculosis classification [10],[11] and lung nodule detection [12], but the use of chest radiographs to discover other pathologies such as pneumonia and pneumothorax motivates an approach that can detect multiple pathologies simultaneously.

In this work, we aimed to assess the performance of various deep learning techniques to automatically detect the presence of 14 different disease classes in chest radiographs and visualize the areas indicating the presence of these diseases in the X-Ray.

II. DATA

We used CheXpert dataset [1], which is a large public dataset for chest radiograph interpretation, consisting of 224,316 chest radiographs of 65,240 patients. The data was collected from Stanford Hospital between 2002-2017. As shown in fig. 1, each X-Ray is labeled for the presence of 14 observations as positive (1), negative (0) or uncertain (-1).



	F	G	H	I	J	K	L	M	N	O	P	Q	R	S
1	No Finding	Enlarged Car	Cardiomegal	Lung Opacity	Lung Lesion	Edema	Consolidatio	Pneumonia	Atelectasis	Pneumothor	Pleural Effusi	Pleural Other	Fracture	Support Devices
2	1									0				1
3			-1	1		-1	-1		-1		-1		1	
4				1			-1						1	
5				1			-1						1	
6						1				0				
7	1	0					0				0			
8	1	0					0				0			

Fig. 1: Labeled image spreadsheet provided with data

These 14 labels and their distribution is shown in fig. 2. Additionally, an image can have multiple positive labels which implies that the problem is a multi-label classification. Resolution of images in the data are 390 x 320 pixels.

Pathology	Positive (%)	Uncertain (%)	Negative (%)
No Finding	16627 (8.86)	0 (0.0)	171014 (91.14)
Enlarged Cardiom.	9020 (4.81)	10148 (5.41)	168473 (89.78)
Cardiomegaly	23002 (12.26)	6597 (3.52)	158042 (84.23)
Lung Lesion	6856 (3.65)	1071 (0.57)	179714 (95.78)
Lung Opacity	92669 (49.39)	4341 (2.31)	90631 (48.3)
Edema	48905 (26.06)	11571 (6.17)	127165 (67.77)
Consolidation	12730 (6.78)	23976 (12.78)	150935 (80.44)
Pneumonia	4576 (2.44)	15658 (8.34)	167407 (89.22)
Atelectasis	29333 (15.63)	29377 (15.66)	128931 (68.71)
Pneumothorax	17313 (9.23)	2663 (1.42)	167665 (89.35)
Pleural Effusion	75696 (40.34)	9419 (5.02)	102526 (54.64)
Pleural Other	2441 (1.3)	1771 (0.94)	183429 (97.76)
Fracture	7270 (3.87)	484 (0.26)	179887 (95.87)
Support Devices	105831 (56.4)	898 (0.48)	80912 (43.12)

Fig. 2: Distribution of pathology labels by classes in the dataset

III. METHODS

A. Data Processing

The dataset has 3 classes for each of the 14 labels. The competition of this dataset suggested many ways of dealing with the uncertain labels (-1). Since the dataset is collected from a credible source, we prefer replacing the (-1) values with (0) for simplicity. Moreover, as this dataset belongs to a competition, we sampled validation and testing data from the training data and downscaled the images to 128 x 128 pixels.

B. Classification

We used various deep learning models and compared them to concurrently detect the presence of 14 different pathologies in frontal-view chest radiographs. Various current state-of-the art models such as DenseNet-121, MobileNet-V2, Inception-V3, VGG-16 were trained on the CheXpert dataset. Common parameters used on all these models for training are learning rate is 0.0002, batch size is 64, optimizer used is Adam and since it is a multi-label classification, binary cross entropy is used as the loss function. Moreover, these models were trained on 30,000 training samples and 5000 validation samples.

Only the last layer in these models are modified according the classification task. After the basic architecture of these models, we have flattened the image, then used a fully connected layer with 1024 hidden units and ReLU activation function. For regularization, we have added a dropout of value 0.5 and the final output layer is a fully connected layer with 14 units (one for each label) and Sigmoid activation function.

We also created an attention model on top of VGG-16 as VGG-16 has highest test accuracy, to see if it can outperform all other models mentioned earlier. Fig. 3 represents the architecture of the attention model. After the basic architecture of VGG-16, we performed batch normalization and added two 2-D convolution layers with a kernel size of (1,1), ReLU activation function, one with 64 filters and the other with 16 filters, to reduce the depth of the image. Then a Locally Connected 2-D layer (sigmoid activation, kernel size is [1,1], one filter) followed by another 2-D Convolution layer (kernel size is [1,1], linear activation, number of filters same as depth of image). Thereafter, we multiply the 2-D convolution layer with batch normalization layer. The 2-D Global Average Pooling has been applied on the resultant multiplied output as well as the previous convoluted output. The ratio of the two global average pooling output is calculated. This process is called as rescaling. Dropout of 0.5 is applied to rescaling and then a fully connected layer with 256 units and ReLU activation function is used followed by another dropout with value 0.25. Finally, we have output layer with 14 units and sigmoid activation function.

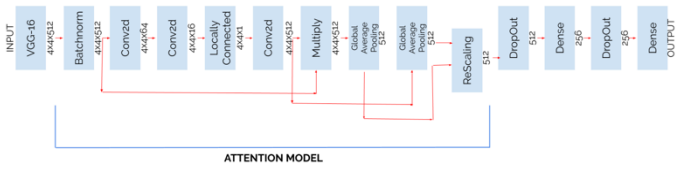


Fig. 3: Attention Model Architecture

C. Visualization

Once the network was trained for multi-label classification, we also generated gradient weighted class activation maps (Grad-CAM) [13],[14]. This visualization technique helps us understand the network and is also useful as an approximate visual diagnosis for presentation to radiologists.

Grad-CAMs generate a heatmap that shows which region of the image weights more for the classification and it can be applied to any CNN architecture. Since VGG-16 model attained the highest test accuracy (86.42%) on the dataset, we used it as reference model for Grad-CAM through which heatmaps will be generated for predicted classes given a test image.

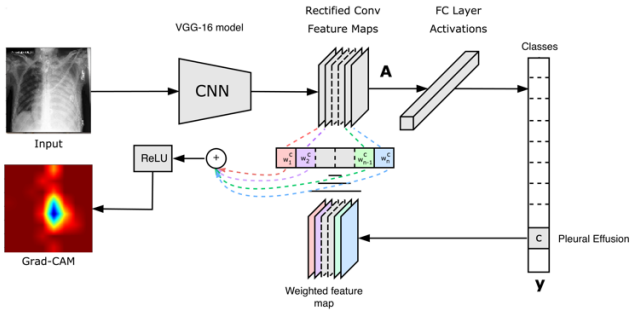


Fig. 4: Grad-CAM Architecture

As shown in fig. 4, we feed a test image to the trained VGG-16 model which predicts the classes for that image. We then compute the gradient of class output with respect to feature map of last convolution layer and pool the gradient over all axes leaving out channel dimension.

After this, we weigh (multiply) the output feature map with computed pooled gradient value to get weighted feature map. This weighted feature map is averaged along channel dimension and ReLU activation is applied on it. Resulting heatmap is then normalized between 0 and 1, resized to input image size, denormalized by multiplying by 255 and colormap is applied on it.

Resulting heatmap is color image whereas input is grayscale. So, the input image is converted to color and then the heatmap is superimposed on it by selecting appropriate alpha or transparency value.

IV. RESULT

A. Classification performance

The following fig. 5 shows comparison of test accuracy of various deep learning models on 5000 test images for multi-label classification task. From the figure, it can be inferred that VGG-16 model outperforms all other models by at least 1%. It is followed by attention model which achieves accuracy of 85.39%. DenseNet-121 model has an accuracy of 84.96% followed by Inception-V3 (83.76%) and then MobileNet-V2 having lowest accuracy of 83.24%.

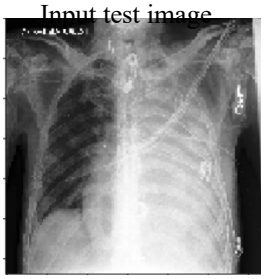
Model	MobileNet-V2	Inception-V3	DenseNet-121	Attention Model	VGG-16
Test Accuracy	83.24%	83.78%	84.96%	85.39%	**86.42%**

Fig. 5: Test accuracy comparison across various models

B. Visualization

For a given test image, we predict all the possible classes out of the 14 classes based on the trained VGG-16 model and visualize the areas most indicative of each of the predicted classes in the form of a heatmap using Grad-CAM technique. We expected attention model to have highest accuracy but since VGG-16 attained the highest accuracy we used it for creating heatmaps. Following table (fig. 6) shows the predicted probabilities for the given test image

Input test image



Pathology	Predicted Probability (%)
No Finding	1.0
Enlarged Cardiomediastinum	5.3
Cardiomegaly	31.4
Lung Opacity	**80.0**
Lung Lesion	2.5
Edema	10.4
Consolidation	12.0
Pneumonia	2.6
Atelectasis	15.3
Pneumothorax	7.0
Pleural Effusion	**98.7**
Pleural Other	1.6
Fracture	2.3
Support Devices	**71.8**

Fig. 6: Table showing predicted probabilities for the given test image

The Grad-CAMs are generated only for the positive classes, because we want to visualize the regions where pathologies are present and not the other way around. Thus, for the input test image shown in fig. 6, three heatmaps are generated for the three positive classes – Lung opacity shown in fig. 8, Pleural Effusion shown in fig. 7, and Support devices shown in fig. 9.

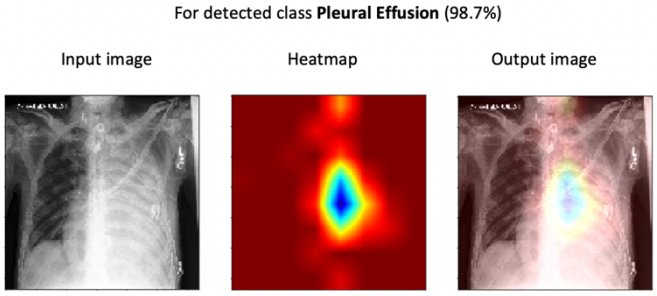


Fig. 7: Heatmap for class Pleural Effusion

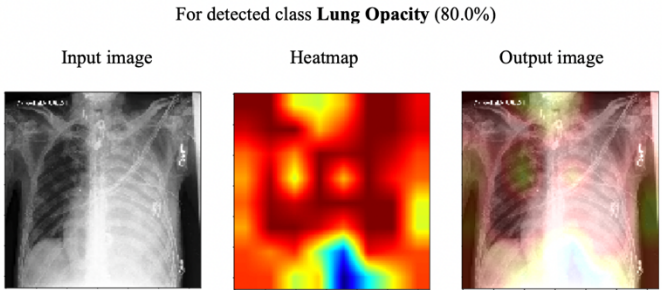


Fig. 8: Heatmap for class Lung Opacity

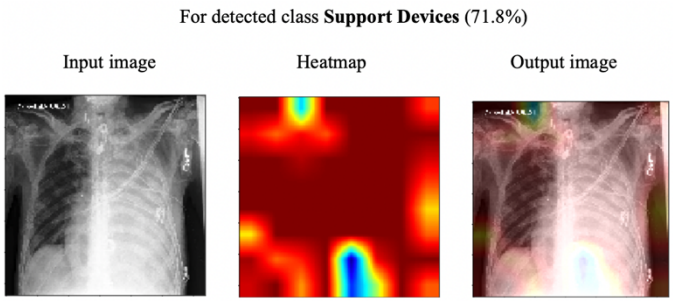


Fig. 9: Heatmap for class Support Devices

V. CONCLUSION

We tried multiple architectures and to integrate the best of them with an attention model. We generated the heatmap along with probability of the positive classes for a given test image which can be used for medical applications. We try to build an easily interpretable heatmap so that making inferences from the diseases becomes easy for the doctor.

Grad-CAM use deeper feature maps, which typically results in better localization due to the higher-level nature of the features in deeper layers but are available only at reduced resolution due to pooling. This is a trade-off which may or may not lead to better results.

The accuracy achieved for now doesn't beat the state of the art in CheXpert competition but if we could somehow label the uncertain images with some very strong and logical explanation and increase the resolution of images, then the model could outperform the state of the art for the current dataset and can become extremely handy in such cases where the certainty of a disease can't be quantified.

ACKNOWLEDGMENT

We would like to thank Prof. Minje Kim for giving us the opportunity to develop and present our own project as a part of his deep learning course. Moreover, his constant support and guidance throughout the duration of the project helped us complete the project duly.

REFERENCES

[1] Irvin, Jeremy et al. "CheXpert: A Large Chest Radiograph Dataset with Uncertainty Labels and Expert Comparison." AAAI (2019).

[2] Pasa, F., Golkov, V., Pfeiffer, F. et al. Efficient Deep Network Architectures for Fast Chest X-Ray Tuberculosis Screening and Visualization. Sci Rep 9, 6268 (2019) doi:10.1038/s41598-019-42557-4

[3] Rajpurkar P, Irvin J, Ball RL, Zhu K, Yang B, et al. (2018) Deep learning for chest radiograph diagnosis: A retrospective comparison of the CheXNeXt algorithm to practicing radiologists. PLOS Medicine 15(11): e1002686. https://doi.org/10.1371/journal.pmed.1002686

[4] Raoof S, Feigin D, Sung A, Raoof S, Irugulpati L, Rosenow EC. Interpretation of plain chest roentgenogram. Chest. 2012 Feb;141(2):545–58. pmid:22315122

[5] Gulshan V, Peng L, Coram M, C. Stumpe M, Wu D, Narayanaswamy A, et al. Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs. JAMA. 2016 Nov 29;316.

[6] Esteva A, Kuprel B, Novoa RA, Ko J, Swetter SM, Blau HM, et al. Dermatologist-level classification of skin cancer with deep neural networks. Nature. 2017 Feb;542(7639):115–8. pmid:28117445

[7] Ehteshami Bejnordi B, Veta M, Johannes van Diest P, van Ginneken B, Karssemeijer N, Litjens G, et al. Diagnostic Assessment of Deep Learning

Algorithms for Detection of Lymph Node Metastases in Women With Breast Cancer. JAMA. 2017 12;318(22):2199–210. pmid:29234806

[8] Cicero M, Bilbily A, Colak E, Dowdell T, Gray B, Perampaladas K, et al. Training and Validating a Deep Convolutional Neural Network for Computer-Aided Detection and Classification of Abnormalities on Frontal Chest Radiographs. Invest Radiol. 2017;52(5):281–7. pmid:27922974

[9] Bar Y, Diamant I, Wolf L, Lieberman S, Konen E, Greenspan H. Chest pathology detection using deep learning with non-medical training. In: 2015 IEEE 12th International Symposium on Biomedical Imaging (ISBI). 2015. p. 294–7.

[10] Maduskar P, Muyoyeta M, Ayles H, Hogeweg L, Peters-Bax L, van Ginneken B. Detection of tuberculosis using digital chest radiography: automated reading vs. interpretation by clinical officers. Int J Tuberc Lung Dis Off J Int Union Tuberc Lung Dis. 2013 Dec;17(12):1613–20.

[11] Lakhani P, Sundaram B. Deep Learning at Chest Radiography: Automated Classification of Pulmonary Tuberculosis by Using Convolutional Neural Networks. Radiology. 2017 Apr 24;284(2):574–82. pmid:28436741

[12] Setio AAA, Ciompi F, Litjens G, Gerke P, Jacobs C, Riel SJ van, et al. Pulmonary Nodule Detection in CT Images: False Positive Reduction Using Multi-View Convolutional Networks. IEEE Trans Med Imaging. 2016 May;35(5):1160–9. pmid:26955024

[13] Selvaraju, R. R. *et al*. Grad-CAM: Why did you say that? Visual Explanations from Deep Networks via Gradient-based Localization. 1610.02391V2 1–5 (2016)

[14] Zhou, B., Khosla, A., Lapedriza, A., Oliva, A. & Torralba, A. Learning deep features for discriminative localization. in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 2921–2929 (2016)