Presentation Video Link: https://www.youtube.com/watch?v=ZYfewy01shc

Detecting Lung Disease from Chest X Rays using Deep Learning

Ahmad Khan

Abstract

This paper replicates the current state of the art deep learning models used to detect 14 different pathologies in Lung Chest X rays and evaluates the model performances against the Stanford CheXpert dataset [1]. The model architectures considered include a pretrained DenseNet 121 architecture as utilized by Irvin et a [2], Category Residual Attention Learning (CRAL) network as proposed by Guan et al (2018) [3], Attention Guided Convolutional Networks (AG-CNN) as proposed by Guan et al (2018) [4], and a model that uses ideas from Residual Attention Network (RAN) for Image Classification (Wang et al) [5]. Each model is evaluated using AUC on validation set of 224 images with ground truths for pathologies based on seasoned radiologist consensus. Classification is done using both the binary classification ‘Ones’ and ‘3-Class Multi Classification’ approach used by Irvin et al. Results show that no single outperforms any model across all pathologies but the AG-CNN Ones approach tends to outperform a simpler pretrained DenseNet121 architecture suggesting that attention learning in CNNs can be effectively utilized.

Introduction

Detecting disease from Chest X Ray images requires specialized medical knowledge and clinical training which makes diagnosis of disease time consuming and costly. Using Deep Learning to instead detect X Ray disease however can make the process cheaper, quicker and perhaps even more accurate than doctors which makes this an important and critical problem to solve. Therefore, this paper aims to predict X ray disease and see if it performs as well or better than current state of the art methods and professionally trained radiologists.

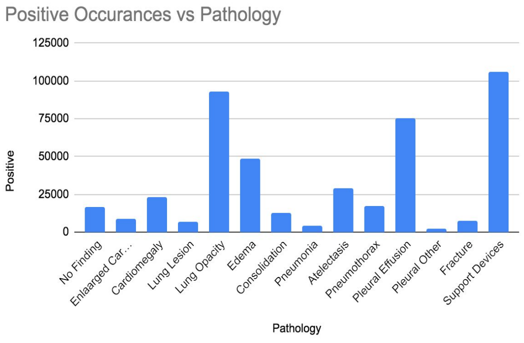
Literature Review

The existing literature on diagnosing disease from Chest X rays using deep learning utilizes Convolutional Neural Networks as the main algorithmic building block to achieve the X ray classification task.

1) Irvin et al and Rajpurkar et al [2] use a 121-layer DenseNet architecture to see how well it performs on the X ray disease classification task. DenseNe [6] by Huang et al improves performance by alleviating the vanishing gradient problem by connecting each layer to every other layer in a feed-forward fashion.
2) Liu et al[7] build a segmentation-based deep fusion network (SDFN) which trains a CNN to diagnose X ray disease using both the entire X ray image as well as one which first detects proposed local regions for the lungs using a lung region generator model with the outputs of both models combined during training.
3) Guan et al [4] utilize attention learning using ‘AG-CNN’ but this time to first find a local region of interest with disease using thresholding on a feature embedding activation map and then train a separate model on this smaller region of interest. Like Liu et all they then combine their outputs in final model.
4) Guan et al [3] also utilize a category-wise residual attention learning framework (CRAL) which helps predict disease by using attention learning to enhance good predictive features for each pathology..
5) Li et al [8] present a unified approach for both localization of disease and classification where they first utilize transfer learning with ResNet to develop convolutional features and then also slice the image into small patches to determine disease likelihood in each patch using local information coming from the patches.

Approach

5 different model architectures were built and evaluated namely pre-trained architectures like ResNet152 and DenseNet 121 as used by Irvin et al, AG-CNN as proposed by Guan et al, CRAL proposed by Guan et al, and modified CRAL using attention framework mentioned by Wang et al and each evaluated on a validation set of 224 images. The CheXpert data consists of 220K X ray images with 14 disease pathologies: No Finding, Enlarged Cardiomediastanum, Cardiomegaly, Lung Opacity,



Lung Lesion, Edema, Consolidation, Pneumonia, Atelectasis, Pneumothorax, Pleural Effusion, Pleural Other, Support Devices. It has three types of labels derived automatically from an automatic NLP text labeler. Positive to denote cases where the disease was present, Negative to denote absence of disease, and Uncertain where the labeler couldn't ascertain with confidence the presence/absence of a disease. The bar chart above shows the occurrence of the positive class for each.

For each model, 2 classification approaches were used to detect disease. The first approach is the binary classification **'Ones'** approach where Uncertain is treated as a Positive outcome. The second is the "**3 Class Classification"** approach. Here uncertain is treated as its own label so the output task becomes a 3*14 multi output multi classification task where each class is treated as its own category and within each class the goal is to predict the existence or not of each of the 14 diseases. During training, a softmax function is applied so that P(Disease = Positive) + P(Disease = Negative) + P(Disease = Uncertain) = 1. For evaluation, only the P(Disease=Positive) predictions are considered after a softmax is applied to the Positive and Negative class scores. This is to keep the approach consistent with Irvin et al.

The first model architecture used was the same as Irvin et al. Both a (ImageNet) pretrained DenseNet 121 and ResNet 152 without any fixed weights were retrained using 200K CheXpert X rays. The final output layer was replaced by 14 outputs for each class for the 'Ones' Classification case and with 3*14 output layer for the 3-Class Multi Classification approach with each row corresponding to one specific outcome (either Positive, Negative or Uncertain). The same was done for the pretrained ResNet 152. DenseNet-121 outperformed ResNet-152 slightly as shown below and so only DenseNet-121 was retained as the **baseline model** against which more complex model architectures next were tried.

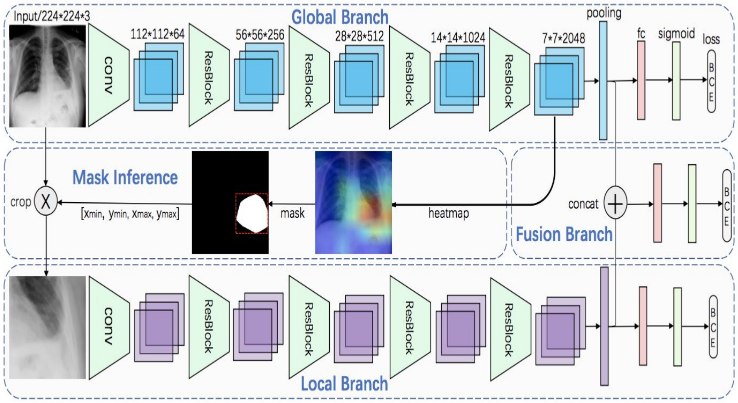
Mean Class AUC Results for Each Pretrained Architecture Model (DenseNet121 retained as Baseline since it performed better)

Model	**Mean Validation AUC**	**Mean Validation Accuracy**	**Mean Validation Loss**
Densenet121 with no fixed weights (Baseline)	***0.776***	***85.13%***	0.357
ResNet152 with no fixed weights	0.769	84.76%	0.347

Pretrained architectures like DenseNet 121 are very powerful at extracting feature embeddings that describe the images very well by utilizing feature maps of previous layers as inputs into the next layer. This encourages feature reuse and feature propagation leading to richer feature representation especially very deep down the network as the vanishing gradient problem is alleviated. ResNet 152 [9] also improves CNN performance by learning residual functions with reference to the inputs of the layers instead of just the output of the previous layer. However, while these approaches do a good job with prediction they theoretically have some limitations for specific prediction tasks like X Ray disease detection. First, X ray disease areas can be highly localized so global architectures might extract the broad image level features ((like shape of lungs etc) but miss the much finer and harder to detect small, highly localized disease patches. For this reason attention learning has been utilized more recently. Attention Learning first proposed for NLP machine translation by Vaswani et al [10] is being increasingly utilized in image detection to help models direct their attention to specific regions or features.

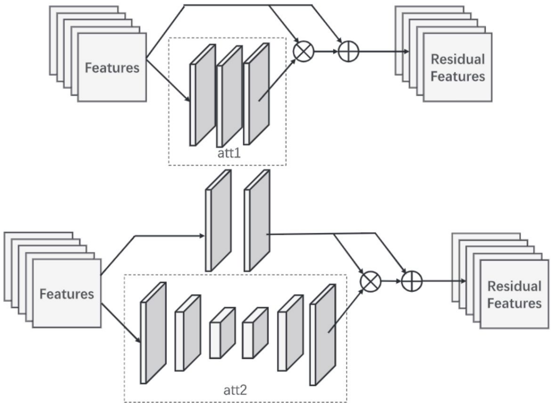
Liu et al [7] utilize Segmentation Deep Fusion Network (SDFN) to detect local patches where disease is present. This can be considered a type of 'hard attention' where the region is explicitly specified and cropped and then another model is trained on the detected sub region. However the drawback with this approach is that it relies on at least a few Chest X ray images to have bounding boxes labeled by radiologists so a lung region generator can learn to detect the lung region first. But the CheXpert data does not come with any bounding boxes for lung regions and so an alternative approach like Attention Guided Convolutional Networks [4] (AG-CNN) is utilized in this paper as the second model. AG-CNN differs from SDFN in that it automatically detects the lung disease region by a masking operation on the feature activation heatmap values generated from the last layer of DenseNet-121 as shown in Figure 1 below. First the global model is trained using a DenseNet-121 architecture on the entire input image and then the regions of the final convolutional feature map with values greater than 70% of the feature map values are cropped and resized into an image with only the diseased region. This smaller disease local region is then trained again using a DenseNet121 and the outputs of the global and local branches are combined in a final Fusion branch after applying max pooling and Relu activation on the feature map outputs. By combining both the global and local model's pooled outputs the model can learn not only the global representation of an image but also the local features on the probabilistic disease region after attention is guided to this region.

Figure 1: Attention Guided Convolutional Network (AG-CNN) Architecture (Image Credit: AG-CNN paper [4])



However, another way to utilize attention is to emphasize the class wise correlations for each of the disease pathologies. Therefore, the third model built is similar to what Guan et al utilize in CRAL [3]. A feature embedding module learns the high level features using a DenseNet-121 CNN and then re-weights this feature map using an attention module on the feature map. The attention module highlights for each class which region of the feature map are likely to contain a specific feature that can be predictive for disease identification. It achieves this by taking the Hadamard product of the attention scores with the feature map for reweighting and then concatenates with the global high level features so that global information is not lost along with the local attention. Two types of attention mechanism are considered and two models are built each using one of the specific attention scheme shown in Figure 2.

Figure 2: Cral Attention Modules. Both Att1 and Att2 are implemented as separate models (Image Credit: CRAL paper [3])



1) The first attention mechanism ATT1 based on the CRAL paper utilizes two 3x3 convolutional layers with 1024 inputs each followed by a Relu activation and then a 1x1 convolutional layer and Sigmoid Activation. This model is referenced as CRAL-ATT1 in the paper and shown in Figure 2 and is identity mapping.

2) The second ‘hourglass’ attention mechanism ATT2 is originally based on the attention module proposed by Wang et al [5] . Three consecutive pooling layers each of 3x3 size are applied followed by 3 bilinear interpolation upsampling operations to make the output the same size as the input to the attention module again. These are then followed by two 1x1 convolutions and a Sigmoid activation. By applying pooling consecutively the model is able to serve as a feature selector that enhances the good feature maps. This model is referenced as CRAL-ATT2 in the paper and shown in Fig 2.

Experimental Results and Discussion

All 4 models described in the preceding section were trained on the 224K image training set provided by CheXpert and validated on the 224 image validation set provided with ground truth provided by the consensus of radiologists. Each model was trained on 320x320 resized input images, with random horizontal flipping, and with ImageNet mean values subtracted for normalization. A learning rate of 0.00001 was used for each model, with Batch Size of 32 and maximum number of epochs equal to 5 except for the AG-CNN models which were trained with learning rate of 0.0001 and batch size 16 and epochs equal to 3 because of memory limitation. Models were all trained on a p2.xlarge GPU instance hosted by AWS using Pytorch. Default values for momentum and beta using the Adam optimizer were used as well.

Binary Cross Entropy with Logit Loss was used as the loss function and monitored to determine the best model to use after training over multiple epochs. For evaluation purposes AUC for each class was noted specifically “**Average AUC**” which is the mean AUC of each of the 14 pathologies as well as “**5 Class AUC**” which is the AUC of the 5 pathologies

(Cardiomegaly, Atelectasis, Pleural Effusion, Edema, and Consolidation) reported by Irvin et al. Accuracy was also monitored during training and validation. Each model's performance in terms of its Average AUC and 5 Class AUC is noted below in Figure 3 using the 'Ones' Classification Approach for each of the 14 pathologies with the best performing model highlighted in bold for each pathology.

Figure 3: Class Evaluation Results for Each Model using Ones Classification Approach

ClassName	**DenseNet-121 Ones (Baseline)**	**CRAL-ATT1 Ones**	**CRAL-ATT2 Ones**	**AG-CNN Ones**	**Irvin et al Comparison**
No_Finding	0.863	0.861	0.868	**0.879**	
Enlarged Cardio	0.528	**0.548**	0.431	0.443	
Cardiomegaly	0.817	0.797	0.744	**0.822**	*0.832*
Lung_Opacity	**0.900**	0.896	0.883	0.898	
Lung_Lesion	0.262	0.270	0.296	**0.562**	
Edema	0.931	**0.934**	0.926	**0.933**	*0.941*
Consolidation	**0.899**	0.896	0.823	0.879	*0.899*
Pneumonia	**0.864**	0.767	0.700	0.800	
Atelectasis	0.810	0.810	0.762	**0.848**	*0.858*
Pneumothorax	0.843	0.851	**0.920**	0.742	
Pleural_Effusion	0.921	0.919	0.918	**0.926**	*0.934*
Pleural_Other	0.790	0.824	0.742	**0.888**	
Support_Devices	0.936	0.924	0.932	**0.939**	
Average AUC	0.797	0.792	0.765	**0.812**	
5 Class AUC	0.875	0.871	0.834	**0.882**	**0.893**

From the above table we can notice the following. The AG-CNN model performs best using the Ones Classification approach for most of the diseases. Note that since AG-CNN has three branches global, fusion, and local the AUC referenced above for each pathology using the AG-CNN model is the fusion branch AUC which combines the output of both the global and local branches. The 5 Class AUC of AG-CNN comes out to be 0.882 while the 5 Class AUC of the baseline Densenet-121 for comparison is 0.875. Across all 14 classes the mean AUC is 0.812 for AG-CNN while it is 0.797 for baseline Densenet-121. For some pathologies like Pneumothorax however CRAL model using attention scheme 2 performs best with 0.920 AUC while Edema is best using CRAL with attention scheme 1 with AUC of 0.934. Both CRAL ATT1 and AG-CNN models either perform approximately as well or better for all diseases (except Pneumonia) when compared to the baseline DenseNet 121.

The improved individual results for each disease classification show that attention can benefit learning for certain classes. For example diseases like Lung Lesion (Nodule) performs best classification with the AG-CNN model with 0.562 in AUC with the improvement coming from the cropped local model (0.64 AUC) that is trained on only the attention masked crop of the image. In comparison the global branch on the entire image for AG-CNN has only 0.26 AUC. Lung Lesion is much more localized and occuring in a smaller area so models that only extract the global broad features like DenseNet-121 do badly with only 0.27 AUC. This can also be seen with Atelectasis where the lung tissue can collapse however this can be both a large area of the lung or just a small area and so attention learning can benefit in this case if the area of Atelectasis is small. However for other pathologies, it is not surprising that the simple Densenet-121 performs best like in the case for Pneumonia. As noted in Guan et al, the inflamed disease area for Pneumonia is relatively large so the generated feature heat map would be more scattered over. A generic model that extracts global features like pretrained DenseNet-121 would be better suited to a disease like Pneumonia than attention focused models that direct attention towards localized regions.

Next, we replicate the results above but this time for the 3-Class Multi-classification approach using uncertain labeled images as their own class. Results are shown below in Figure 4. From the table we can note the following. For the 3 Class classification case the AG-CNN framework is slightly modified as well to see if combining AG-CNN and CRAL leads to improved performance. Instead of using just the DenseNet-121 feature map it utilizes the weighted CRAL-ATT1 Feature map so that the input to the local branch is a feature map which is weighted by the attention scores from the CRAL attention module. This model is called AG-CNN-CRAL (Hybrid).

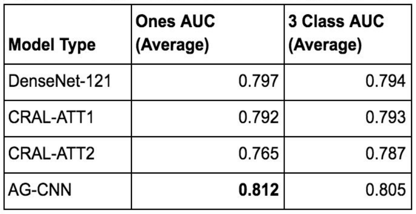
First like the 'Ones' Classification case above the AG-CNN model beats the DenseNet 121 on Average AUC but not by as much. On 5 Class AUC however CRAL-ATT1 performs best. And looking at individual diseases we can see that CRAL attention models tend to perform better for certain diseases (CRAL ATT2 for Edema, Consolidation and Pneumothorax while CRAL ATT1 for Cardiomegaly and Pleural Effusion). So once again no model performs best for all diseases but both CRAL models and AG-CNN outperform DenseNet 121 on many of the diseases (except Atelectasis/Pleural Other).

Figure 4: Class Evaluation Results for Each Model using 3 Class Multi approach

ClassName	DenseNet 121 Multi	CRAL-ATT1 Multi	CRAL-ATT2 Multi	AG CNN (Hybrid) Multi	Irvin et al Comparison
No_Finding	0.8471	0.900	0.867	**0.909**	
Enlarged Cardio	0.4868	0.522	**0.569**	0.492	
Cardiomegaly	0.8013	**0.817**	0.760	0.810	*0.854*
Lung_Opacity	0.9079	**0.913**	0.911	0.903	
Lung_Lesion	0.3605	0.202	0.202	**0.446**	
Edema	0.9213	0.918	**0.935**	0.897	*0.928*
Consolidation	0.8731	0.893	**0.903**	0.891	*0.937*
Pneumonia	0.6831	0.748	0.752	**0.775**	
Atelectasis	**0.8116**	0.800	0.800	0.810	*0.821*
Pneumothorax	0.8451	0.858	**0.888**	0.799	
Pleural_Effusion	0.9248	**0.931**	0.913	**0.929**	*0.936*
Pleural_Other	**0.9485**	0.893	0.811	0.871	
Support_Devices	0.9073	0.919	0.919	**0.928**	
Average AUC	0.794	0.793	0.787	*0.805*	
5 Class AUC	0.866	**0.872**	0.862	0.867	**0.895**

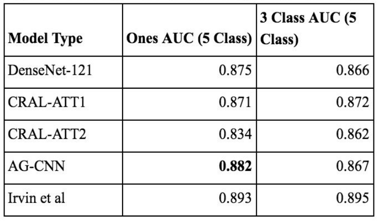
However, if we compare the 3 Class approach with the Ones Classification approach we see that most models perform very similarly in both Ones and 3 Class Multi Classification cases with not a big difference observed in Average AUC and 5 class AUC when comparing the same model with either classification approach except AG-CNN which performs slightly better using the Ones binary classification approach and CRAL-ATT2 which performs better on the 3 class approach.

Figure 5: Average 14 class AUC for both target classification approaches (Ones vs 3 Class)



Model Type	Ones AUC (Average)	3 Class AUC (Average)
DenseNet-121	0.797	0.794
CRAL-ATT1	0.792	0.793
CRAL-ATT2	0.765	0.787
AG-CNN	**0.812**	0.805

Figure 6: 5 average AUC for both target classification approaches (Ones vs 3 Class)

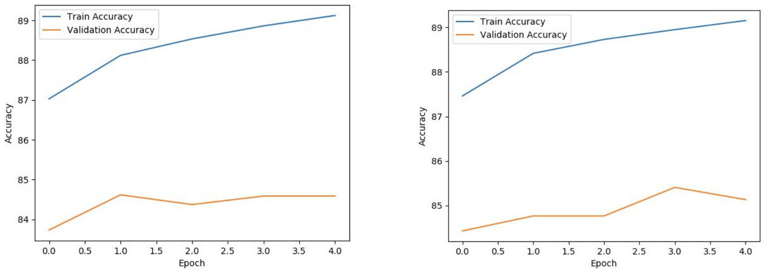


Model Type	Ones AUC (5 Class)	3 Class AUC (5 Class)
DenseNet-121	0.875	0.866
CRAL-ATT1	0.871	0.872
CRAL-ATT2	0.834	0.862
AG-CNN	**0.882**	0.867
Irvin et al	0.893	0.895

Figure 5 and 6 below summarize this observation. This is counter intuitive as according to Irvin et all using the less noisy 3 class labels should lead to improved performance for some diseases but this is not observed during the experiments with significant difference for some diseases like Atelectasis where the best One classification model is 0.857 AUC while the best 3 Class Classification model is AUC 0.811. One hypothesis for this could be that while the 3 Class approach factors in uncertainty in making predictions it nonetheless leads to overfitting of the data. The models become really good at identifying all 3 classes but in the real world there is also uncertainty in radiologist predictions so that uncertainty is not captured in the ground truth labels as the labels are just Positive or Negative with no borderline uncertainty captured. The accuracy curves below in Figure 7 show overfitting for the 3 Class Classification approach where positive label probability is used as prediction as there is greater divergence between the training and validation accuracy curves of each model. This was also observed for the Loss curves.

Figure 7: Three Class Classification Accuracy by Epoch (shows much greater divergence between training and validation accuracy)

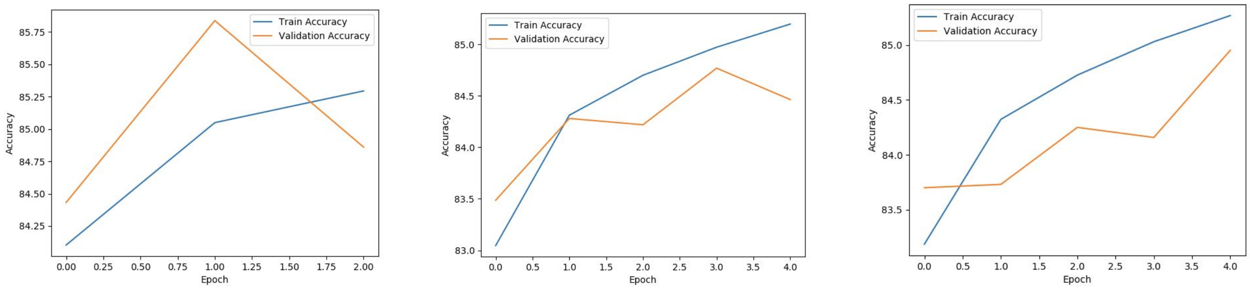
DenseNet 121 **Cral ATT1 Model**



However, for the Ones classification approach models this overfitting was not as severe as the difference between training and accuracy was much smaller as shown in the learning curves below in Figure 8 for each model and each epoch.

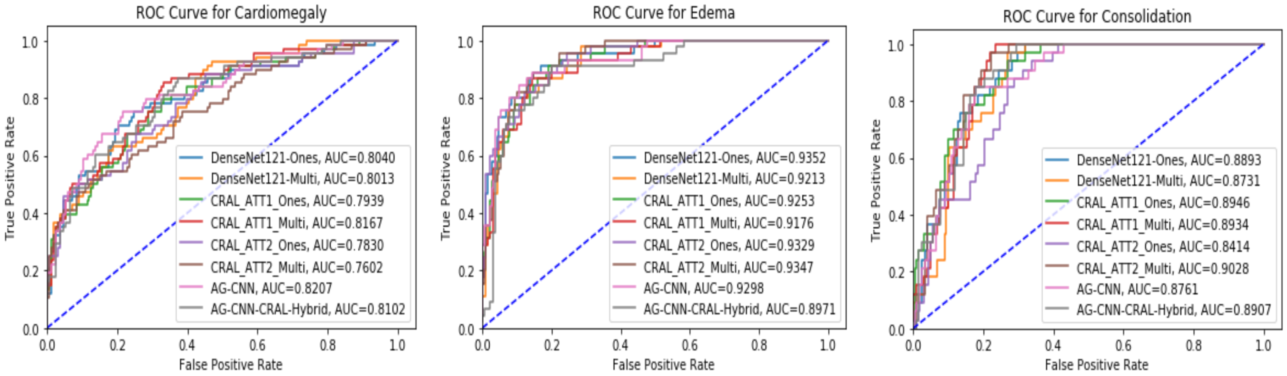
Figure 8: Ones Classification Accuracy by Epoch

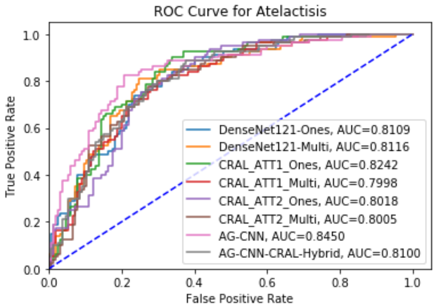
AGCNN Model **DenseNet 121 Model** **CRAL ATT1 Model**

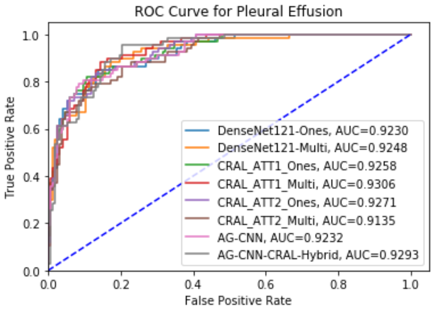


Finally, we plot the AUC curves for 5 diseases referenced by Irvin et al to show the difference in performance of each pathology with respect to each of the 8 models is when comparing their True Positive and False Positive Rates. We can see that AG-CNN and CRAL models tend to outperform a simpler DenseNet121 architecture as the ROC curve is shifted right.

AUC curve for the 5 major disease curves with each model results shown (Cardiomegaly, Edema, Consolidation, Atelectasis, Pleural Effusion)







Challenges/Areas of Improvement

It should be noted that the performance of the models including DenseNet-121 is still worse than what Irvin et al reported. One of the challenges with the modeling was the slow training time and memory bottleneck. This meant that training had to be stopped early after only 5 epochs or so and a parameter grid search for each model was not done as it would have taken too much time. Averaging model performance over multiple runs could also lead to less variance in predictions and more stable results. Therefore, this could potentially be a reason for why results did not match Irvin et al exactly and also why some models could have performed better than others. Secondly, both the AG-CNN and CRAL papers did not explicitly mention some network architecture parameters that could be critical in model performance and thus values for these were guessed. In attention module of the CRAL paper it was not specified how many filters would be in the convolutional layers of the attention module. In AG-CNN the cutoff threshold was mentioned to be 0.7 but it was not clear if this implied 70% percentile value or absolute value (because of Relu activation the median value was 10 and not close to 0.7 so 0.7 did not make much sense in an absolute context). These values had to be assumed and could significantly impact performance.

Additionally, the models could have been improved by doing sequential task prediction where instead of predicting all 14 diseases at the same time we assume disease occurrence is hierarchical in nature where some diseases occur first before others. That is we predict the parent disease and if only we feel confident that the parent disease is there that we predict a child disease. The 14 diseases are thus modeled as a parent-child dependency hierarchy as opposed to being independent classes of each other. Motivation for this learning approach comes from medical observation where certain lung diseases are organized in a hierarchy and the existence of the parent disease suggests the existence of a child disease. Utilizing this hierarchical information in the learning task has shown promising results for Pham et al [11] and could be incorporated to further improve model performance. Finally, other promising attention mechanisms like Squeeze and Excitation [12] framework could have been utilized as part of the models to see if they would lead to improved performance.

Conclusion

Based on the results we can thus conclude the following that using pre-trained general architectures like DenseNet-121 can achieve very good performance when it comes to X ray disease classification matching radiologist performance. However, new methods that utilize attention mechanisms in CNN architectures can improve performance even further than just using pre-trained architectures. Specifically we observed that using Attention Guided Convolutional Neural Network (AG-CNN) architecture that focuses attention and retrains model on localized high probability disease areas tends to outperform a simpler pretrained DenseNet-121 baseline. Similarly, the CRAL attention based architecture was also promising as it showed superior performance for certain disease classes like Pneumothorax. The results of this paper thus show promise in utilizing attention learning for X Ray disease classification using Convolutional Networks.

However, a few points should be noted. First, no single architecture outperforms the other across all 14 disease types. Even a simpler DenseNet-121 can beat AG-CNN for some diseases which are more global like Pneumonia so there is no clear winner across all diseases. Second, the type of classification itself (Binary vs MultiLabel) can make some difference in model performance as ground truth uncertainty can lead to some variation in model performance. However, this variation was not as severe for this project compared to what was observed in Irvin et al. Finally, the performance of many diseases using DenseNet-121 is still less than what was quoted by Irvin et al in their paper.

References

1) https://stanfordmlgroup.github.io/competitions/chexpert/
2) J. Irvin, P. Rajpurkar, M. Ko, Y. Yu, S. Ciurea-Ilcus, C. Chute, H. Marklund, B. Haghgoo, R. Ball, K. Shpanskaya, J. Seekins, D. A. Mong, S. S. Halabi, J. K. Sandberg, R. Jones, D. B. Larson, C. P. Langlotz, B. N. Patel, M. P. Lungren, and A. Y. Ng. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. arXiv preprint arXiv:1901.07031, 2019.
3) Q. Guan and Y. Huang. Multi-label chest x-ray image classification via category-wise residual attention learning. Pattern Recognition Letters, 2018.
4) Guan Q, Huang Y, Zhong Z, Zheng Z, Zheng L, Yang Y. Diagnose like a Radiologist: Attention Guided Convolutional Neural Network for Thorax Disease Classification. ArXiv180109927 Cs, 2018.
5) F. Wang, M. Jiang, C. Qian, S. Yang, C. Li, H. Zhang, X. Wang, and X. Tang. Residual attention network for image classification. In CVPR, 2017. 2, 6
6) Gao Huang, Zhuang Liu, Kilian Q. Weinberger, and Laurens van der Maaten. Densely connected convolutional networks. arXiv preprint arXiv:1608.06993, 2016b.
7) H. Liu, L. Wang, Y. Nan, F. Jin, and J. Pu. Sdfn: Segmentation-based deep fusion network for thoracic disease classification in chest x-ray images. arXiv preprint arXiv:1810.12959, 2018
8) Z. Li, C. Wang, M. Han, Y. Xue, W. Wei, L.-J. Li, and L. Fei-Fei. Thoracic disease identification and localization with limited supervision. CVPR, 2018.
9) K. He, X. Zhang, S. Ren, J. Sun, "Deep residual learning for image recognition", 2015.
10) Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In Advances in Neural Information Processing Systems, pages 6000–6010.
11) Hieu H. Pham, Tung T. Le, Dat Q. Tran, Dat T. Ngo, Ha Q. Nguyen, Interpreting chest X-rays via CNNs that exploit disease dependencies and uncertainty labels
12) J. Hu, L. Shen, and G. Sun. Squeeze-and-excitation networks. arXiv preprint arXiv:1709.01507, 2017