Detecting COVID-19 with Chest X-ray using CNN with Attention

Ruiming(Ray) Li,* Jinlu(Krystal) Ma,† and Songyu Yan‡
University of Pennsylvania
(Dated: July 20, 2021)

I. INTRODUCTION

The COVID-19 pandemic has disrupted public health systems worldwide, causing 3.17 millions of loss of human life as of April, 2021[1], and poses devastating economic and social disruption to the world. As the number of COVID-19 cases remains high, there is an ongoing need to screen infected patients as quickly and accurately as possible to give them appropriate treatment. Chest X-ray, among other diagnosis tools, has proven to be a cost-effective way to achieve that. In fact, pooled results showed that chest X-ray correctly diagnosed COVID-19 in 80.6% of people who had COVID-19 [4].

In this project, we experimented with 6 different models aiming to detect COVID-19 using chest X-rays. These 6 models include a machine learning baseline of Random Forest, two deep learning baselines of an AlexNet trained by us and a pretrained ResNet50 on ImageNet, and 3 advanced deep learning models that apply attention mechanism to AlexNet. Although there were previous works on COVID-19 detection with X-ray using CNN, there was not an architecture that combines AlexNet with attention mechanism. After experimenting with 3 variations of AlexNet with attention, we achieved a highest accuracy of 96.8%, a 7.7% improvement from AlexNet without attention, showing effectiveness of attention in COVID-19 detection task.

II. RELATED WORK

As large scale X-ray datasets become available in recent years, there has been a surge in applying deep learning methods to chest X-ray image classification and detection tasks. For example, Rajpurkar et al.[11] developed an algorithm, CheXNet, which is a 121-layer convolutional neural network that can detect pneumonia from chest X-rays at a level exceeding practicing radiologists. Besides traditional CNN, many recent works also applied attention mechanism to chest X-ray analysis. Note that the attention mechanism mentioned here mainly refers to the "trainable attention", which is used as a mechanism to help improve the training of the models by paying attention attention to more important features. Some examples of the recent works that use attention to various tasks associated with imaging include Rubin et al.[12], which designed a Dual-Network architecture to extract the image information of both frontal and lateral views for the chest radiography abnormality assessment task. Ypsilantis et al.[14] proposed a recurrent attention model to identify cardiomegaly regions. Pesce et al.[8] on the other hand introduced a soft attention mechanism, which locates lesions with highlighting part of the saliency map generated by CNN. Guan et al.[2] used attention to generate masks, which help to amplify lesion regions. Although not applied on classifying X-rays, Jetley et al.[5] guides our experiments in this project, where they applied attention to the popular CNN architectures of VGGNet (Simonyan & Zisserman [13]) and ResNet (He et al.[3]), and captures coarse-to-fine attention maps at multiple levels to classify images in the CIFAR-10 and CIFAR-100 datasets.

III. DATASET

We are using chest X-ray images of COVID patients from the COVID-19 Radiography Database on Kaggle[6]. The Kaggle database has chest X-ray images of 3616 COVID-19 positive cases along with

* rayli@seas.upenn.edu
† jinluma@seas.upenn.edu
‡ yveyan@seas.upenn.edu

Figure 1. Example COVID-19 Positive Chest X-Ray Image

Figure 2. Example Normal Chest X-Ray Image

10,192 normal, 6012 lung opacity (non-COVID lung infection), and 1345 viral pneumonia. Each image has a dimension of $299 \times 299 \times 1$, meaning that the image is grayscaled and there are in total 89,401 features to train on.

IV. METHODS

A. Image Pre-processing and Augmentation

In order to achieve a balanced number of COVID-19 positive and negative images, we are using all of the 3616 COVID-19 positive images from the dataset and another 3616 normal lung X-ray images. We then split the data into training, validation and testing set, each consists of 4648, 1162 and 1423 number of images. For all the deep learning models, we chose Binary Cross Entropy Loss with Logits(BCEWithLogitLoss [10]) as the loss function, and Adam[9] with 0.0001 learning rate as the optimizer. We loaded images into batches of size $64 \times 3 \times 299 \times 299$, and augmented training samples via random rotation, random horizontal flip and normalization to avoid overfitting.

B. Architecture Prior

All of our models are based on the assumption that a COVID-19 patient's chest X-ray displays abnormal patterns, distinguishable from healthy chest X-rays.

The intuition for introducing attention into traditional CNN is that when identifying if a patient is COVID-19 positive, doctors will pay attention to anomaly patterns in the X-ray image and then make an decision. Using this intuition, we incorporated attention mechanism into our convolution model AlexNet, where the model selectively processes data by focusing on segments that are more informative.

The implementation of attention layers serves two purposes - interpreting the model by visualizing the attention scores, and improving model performance.

C. Evaluation Metrics

As the problem has been formulated into a binary classification task applied on a balanced data set, we simply choose *Accuracy* as our evaluation metrics across all models, defined as

$$\frac{\text{Total Number of Correct Predictions}}{\text{Total Number of Images}} \tag{1}$$

D. Models

1. Random Forest

We choose a Random Forest classifier as the machine learning baseline method as it is non-parametric and has few hyperparameters, leading to low biases in general and simple to implement as a baseline. However, Random Forest's complexity grows with the number of training samples. To prevent exceeding the available memory, we started with 5 estimators and incremented 2 estimators per batch by setting the "warm start" parameter in sklearn to be true:

```
{    'max_depth': None,
     'max_features': 'auto',
     'max_leaf_nodes': None,
```

```
'max_samples': None,
'min_impurity_decrease': 0.0,
'min_impurity_split': None,
'min_samples_leaf': 1,
'min_samples_split': 2,
'min_weight_fraction_leaf': 0.0,
'n_estimators': 5,
'warm_start': True }
```

We flattened the augmented training images from dimension $3 \times 299 \times 299$ to 1×268203, and trained the model accordingly. The Random Forest model is not expected to produce high predictability power as it does not take into account local semantics, disregarding relationships across neighboring pixels.

2. *Deep Learning Baseline1 - AlexNet*

We adopted AlexNet's[7] architecture for our first deep learning baseline model. To further improve model performance and prevent both exploding and vanishing gradient, we further added batch-normalization layers and trained the model with Xavier Initialization. We also changed the final output dimension to 2 to cater to binary classification problem.

Table I. Improved AlexNet Architecture

Layer	In-Dim	Out-Dim	Filter
Convolution	3	64	11*11
Maxpool(3*3) + ReLu + 2D Batch Normalization			
Convolution	64	192	5*5
Maxpool(3*3) + ReLu + 2D Batch Normalization			
Convolution	192	384	5*5
Maxpool(3*3) + ReLu + 2D Batch Normalization			
Convolution	384	256	3*3
Maxpool(3*3) + ReLu + 2D Batch Normalization			
Dropout(0.5)			
Convolution	256	256	3*3
Average Pool(6*6)			
Linear + ReLu	9216	4096	
Linear + ReLu	4096	4096	
Linear	4096	1	

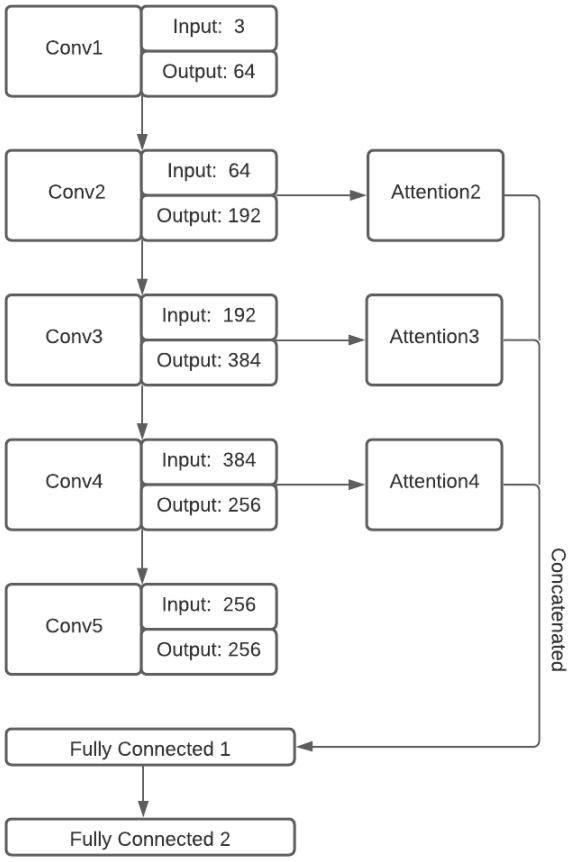

Figure 3. AlexNet + Attention Architecture

3. *Learning Baseline2 - Residual Neural Network*

We then performed transfer learning on ResNet50[3], fine-tuning only the last fully connected Linear layer and adding 1 Linear layer that maps to the output space. The transferred network was trained on more than a million images from the ImageNet database. In preprocessing stage, we normalized data with the required mean=[0.485, 0.456, 0.406], and std=[0.229, 0.224, 0.225]. As our project analyzes gray-scaled X-ray images, we expect a shift in training data distribution.

4. *Advanced Deep Learning - Attention*

Our approach is to insert attention layers after certain convolution layers in the AlexNet. We experimented with 3 variations of architecture, which

we will discusse later. Figure 3 shows a preview of the third variation, "AlexNet + Attn 3". Referencing Learn to Pay Attention [5], we defined the output of the last fully connected convolution layer as the 'global image' g, also known as the attention Query. We used the dot product between a local filters f_i, also known as the attention Key and Value, and the 'global image' as the attention score s_i, defined as:

$$s_i =< f_i, g > \text{ for } i \in \{1, ...n\} \tag{2}$$

where i indicates the spatial location within a convolution layer. A softmax layer is implemented to normalize attention scores to a probability distribution,

$$\alpha_i = \frac{exp(s_i)}{\sum_j^n exp(s_j)} \text{ for } i \in \{1, ...n\} \tag{3}$$

The final output of the attention layer O is then a weighted average of α_i and its corresponding f_i,

$$O = \sum_{i=1}^{n} \alpha_i f_i \tag{4}$$

The final output is produced on the attention output O, followed by 2 fully connected linear layers.

If multiple attention layers are present, the output vectors from each attention mechanism are concatenated $O = [O_1^1, O_2^1, ..., O_1^2, O_2^2, ...]$ before sent to fully connected layers. We used concatenation, instead of the last attention output vector, because different levels of convolution layers capture different semantics relations within training images. The final output is expected to interpret both local and global information.

In order to accelerate training, we used matrix multiplication to compute attention scores. This raises a problem as filters from different layers have different dimensions. We thus projected all convolution output filters to be of the same dimension as the 'global image'. Below we describe the 3 variations of the attention models:

1. **AlexNet+Attn 1**

   Attention block is added after the 4th convolution layer. The attention block takes filters generated from the 4th convolution layer and the 5th convolution layer, 'global image' g as inputs, and returns the attention output O^4.

Table II. AlexNet+Attn 1 architecture

Layer	In-Dim	Out-Dim	Filter
Convolution1	3	64	11*11
Maxpool(3*3) + ReLu + 2D Batch Normalization			
Convolution2	64	192	5*5
Maxpool(3*3) + ReLu + 2D Batch Normalization			
Convolution3	192	384	5*5
Maxpool(3*3) + ReLu + 2D Batch Normalization			
Convolution4	384	256	3*3
Attention 4	256	256	
Maxpool(3*3) + ReLu + 2D Batch Normalization			
Dropout(0.5)			
Convolution5	256	256	3*3
Linear + ReLu	256	128	
Linear	128	1	

2. **AlexNet+Attn 2**

   On top of AlexNet+Attn 1, AlexNet+Attn 2 adds an additional attention block after the 3rd convolution layer, where it takes the filters f^3 from the 3rd convolution layer and 'global image' g as inputs, and returns the attention output O^3. f^3 are projected to be the same size as g. O^3 and O^4 are concatenated horizontally, and then sent to the fully connected layers.

Table III. AlexNet+Attn 2 Architecture

Layer	In-Dim	Out-Dim	Filter
Convolution1	3	64	11*11
Maxpool(3*3) + ReLu + 2D Batch Normalization			
Convolution2	64	192	5*5
Maxpool(3*3) + ReLu + 2D Batch Normalization			
Convolution3	192	384	5*5
Attention 3	256	256	
Maxpool(3*3) + ReLu + 2D Batch Normalization			
Convolution4	384	256	3*3
Attention 4	256	256	
Maxpool(3*3) + ReLu + 2D Batch Normalization			
Dropout(0.5)			
Convolution5	256	256	3*3
Linear + ReLu	256*2	128	
Linear	128	1	

3. **Alexnet+Attn 3**

   On top of AlexNet+Attn 2, AlexNet+Attn 3 adds an additional attention block after the 2nd convolution layer using the same approach.

Table IV. AlexNet+Attn 3 Architecture

Layer	In-Dim	Out-Dim	Filter
Convolution1	3	64	11*11
Maxpool(3*3) + ReLu + 2D Batch Normalization			
Convolution2	64	192	5*5
Attention 2	256	256	
Maxpool(3*3) + ReLu + 2D Batch Normalization			
Convolution3	192	384	5*5
Attention 3	256	256	
Maxpool(3*3) + ReLu + 2D Batch Normalization			
Convolution4	384	256	3*3
Attention 4	256	256	
Maxpool(3*3) + ReLu + 2D Batch Normalization			
Dropout(0.5)			
Convolution5	256	256	3*3
Linear + ReLu	256*3	128	
Linear	128	1	

V. EXPERIMENTS AND RESULTS

Our machine learning baseline, which uses random forest model, reached a 78.0% accuracy on the test dataset. Our self-trained and tuned AlexNet, which we chose as our deep learning baseline performed fairly well, reaching 89.1% test accuracy. We also experimented with transfer learning using ResNet50. Although we thought ResNet50 would perform just as well as AlexNet if not better, it turned out to only reached a test accuracy of 77.6%. This tells us that even though taking the weights directly from ResNet50 can reduce training time, the fact that it is trained on ImageNet had little positive impact on the real task to predict COVID-19, indicating ImageNet has very different features than our COVID-19 X-ray dataset. Lastly, we have 3 models with added attention mechanism. The first setup, "AlexNet + Attn 1" reached a test accuracy of 92.4%. The second setup, "AlexNet + Attn 2" has the highest test accuracy among all 3 attention-induced models, reaching as high as 96.8%, a much better improvement from our deep learning baseline - the self-trained AlexNet. Our last attention setup, "AlexNet + Attn 3" has a test accuracy of 94.4%.

Table V. Validation and Test Accuracy for 6 Models

Models	**Val Acc**	**Test Acc**
Random forest	77.9%	78.0%
AlexNet	94.3%	89.1%
ResNet50	83.8%	77.6%
AlexNet+Attn 1	95.0%	92.4%
AlexNet+Attn 2	97.2%	96.8%
AlexNet+Attn 3	95.3%	94.4%

Table V summarizes the validation and test accuracies for all 6 models. Training and validation loss and accuracy graphs for the 5 models (no random forest) are shown in Figure 4 to Figure 13.

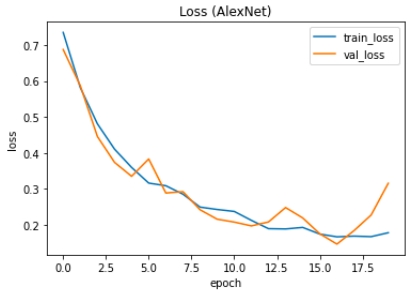

Figure 4. Training and Validation Loss (AlexNet)

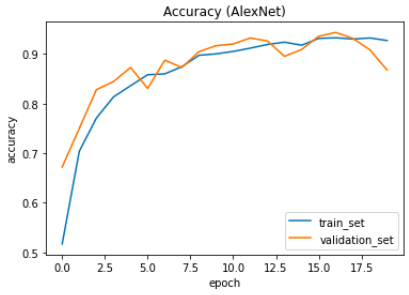

Figure 5. Training and Validation Accuracy (AlexNet)

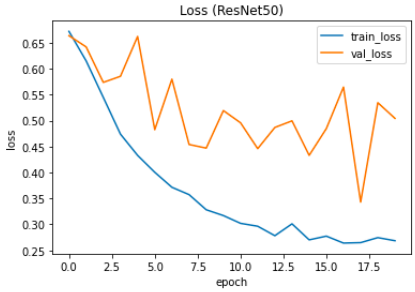

Figure 6. Training and Validation Loss (ResNet50)

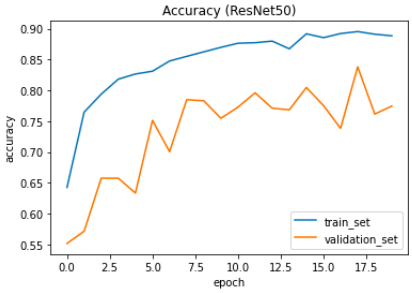

Figure 7. Training and Validation Accuracy (ResNet50)

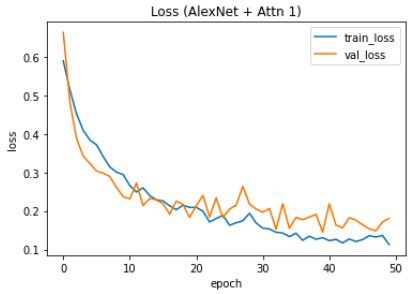

Figure 8. Training and Validation Loss (AlexNet + Attn 1)

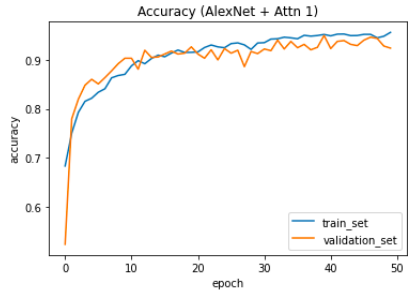

Figure 9. Training and Validation Accuracy (AlexNet + Attn 1)

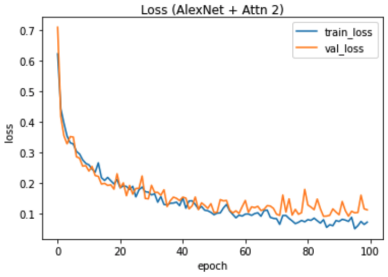

Figure 10. Training and Validation Loss (AlexNet + Attn 2)

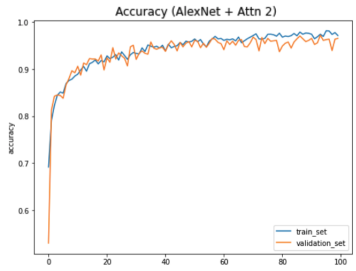

Figure 11. Training and Validation Accuracy (AlexNet + Attn 2)

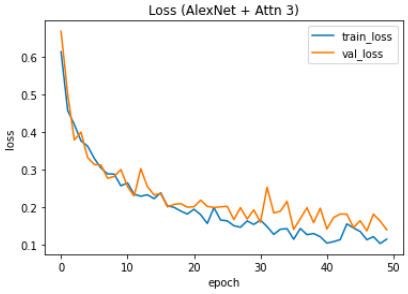

Figure 12. Training and Validation Loss (AlexNet + Attn 3)

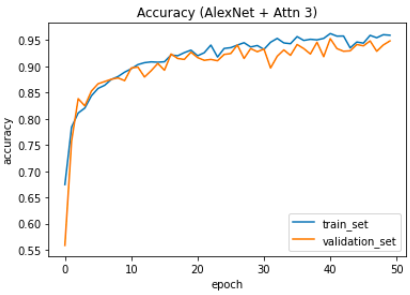

Figure 13. Training and Validation Accuracy (AlexNet + Attn 3)

In order to make sure that our model is learning the right things, rather than some trivial features on the X-ray such as watermarks, we also visualized our attention weights. Figure 14 shows the heat map of three attention layers in AlexNet + Attn 3 model for 4 sample X-rays. Since the attention weights are calculated from feature maps in various locations in AlexNet, you can see that the further the attention block is inserted, the harder it is to interpret whether the attention is paid to the correct places on the X-ray. However, from the heat map for the first attention layer, it is easily seen that the model pays more attention to where the viral pneumonia hotspots are present around the mid regions and the lower zones in the lungs, rather than areas like neck and shoulders. The second and third layer heat map also appears to be a more "zoomed in" version of the first layer heat map. Therefore, the visualization of attention weights helped confirm that our model is indeed paying attention to the correct region of interest in the X-rays. Figure 15 shows the first attention layer weights for 15 X-ray samples. The figure consists of 3 columns and 5 rows, representing the heat map for the first attention layer weights for 15 samples.

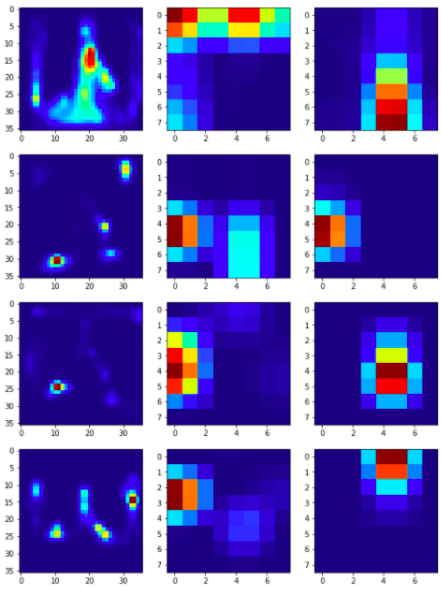

Figure 14. Visualization of 3 Layers of Attention Weights for 4 X-ray Samples (AlexNet + Attn 3)

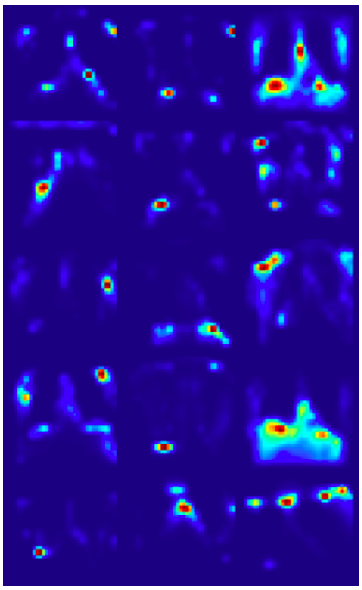

Figure 15. Visualization of First Layer Attention Weights for 15 X-ray Samples (AlexNet + Attn 3)

VI. CONCLUSION

In this project, we successfully classified COVID chest X-ray images with a maximum of 96.8% test accuracy. We found that the architecture with the highest accuracy is AlexNet with attention layers after the 3rd and 4th convolution layer. We also found that using attention mechanism gives a 7% improvement in accuracy compared to AlexNet baseline. By visualizing the attention weights in different layers, we verified that our deep learning model focuses on meaningful regions in the image and does not classify based on irrelevant features. The attention weights are also helpful in informing doctors of the potential regions of interest for classifying COVID patients.

A. Social Impact Analysis

The primary contribution of our project is to expedite the diagnosis and treatment of COVID patients. By automating the process of identifying COVID-19 from X-ray chest images, doctors can pay attention to patients identified by our algorithm to have COVID-19 quickly and make more informed decisions on their treatments. With our advanced deep learning method that utilizes the attention mech-

anism, we can also highlight regions in which the model pays attention to and inform doctors of the potential areas of concern in any X-ray chest image. We do not expect this to be the only means for doctors to diagnose a patient, but such additional information that can be helpful in diagnosis and speed up the process.

With the idea of using a pre-trained AlexNet with attention mechanism, we also expect to lower the cost of training for other diseases in the future. For example, in the case of the emergence of a new virus that is not COVID-19, we can quickly perform diagnosis using the proposed architecture. This is critical for stopping a disease early and preventing a pandemic.

However, our project may suffer from biases in the data potentially, depending on the population on which the X-ray images are collected. We do not have any demographic data of the patients and our algorithm may oversample a certain population and perform less accurately on less-seen populationâs X-ray chest images. However, this bias should be limited since X-ray images are not as affected by demographic features as a normal image.

VII. FUTURE WORK

As we achieved a maximum test accuracy of 96.8% and visualized attention on the image, there are two main areas of future work we are interested in:

1) Further explore possible architectures and fine-tune hyperparameters for greater accuracy. Due to the limit of time, we were only able to fine tune the model based on AlexNet architecture and insert attentions at a few layers. Given more time and computing power, we can further fine-tune our hyper-parameters and explore different variations of ResNet and other architectures such as visual transformers to increase test accuracy. This will help doctors better identify potential patients with COVID-19 and reduce false negative rate.

2) Better interpretation of results by layering attention activation map with original X-ray images, and comparing attention to feature-importance measures such as LIME and SHAP. If we can impose the activation map on original images, we can better visualize and find patterns on the original X-ray that lead to a positive COVID-19 classification. Furthermore, we can compare attention weights with feature importance measure such as LIME and SHAP to give a fuller picture for explanation, so we can inform doctors of potential patterns for identifying COVID-19 patients and enrich human's understanding on the disease.

ACKNOWLEDGMENTS

We appreciate the guidance from Prof. Koerding, Prof. Lyle Ungar and teaching assistants throughout the semester.

BIBLIOGRAPHY

[1] *Coronavirus Update(Live)*. 2021. URL: https://www.worldometers.info/coronavirus// (visited on 04/08/2021).

[2] Qingji Guan et al. "Diagnose like a Radiologist: Attention Guided Convolutional Neural Network for Thorax Disease Classification". In: *CoRR* abs/1801.09927 (2018). arXiv: 1801.09927. URL: http://arxiv.org/abs/1801.09927.

[3] Kaiming He et al. *Deep Residual Learning for Image Recognition*. 2015. arXiv: 1512.03385 [cs.CV].

[4] *How accurate is chest imaging for diagnosing COVID-19?* 2020. URL: https://www.cochrane.org/CD013639/INFECTN_how-accurate-chest-imaging-diagnosing-covid-19.

[5] Saumya Jetley et al. "Learn To Pay Attention". In: *CoRR* abs/1804.02391 (2018). arXiv: 1804.02391. URL: http://arxiv.org/abs/1804.02391.

[6] *Kaggle COVID-19 Radiography Database*. 2021. URL: https://www.kaggle.com/tawsifurrahman/covid19-radiography-database.

[7] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. "ImageNet Classification with Deep Convolutional Neural Networks". In: *Advances in Neural Information Processing Systems*. Ed. by F. Pereira et al. Vol. 25. Curran Associates, Inc., 2012. URL: https://proceedings.neurips.cc/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf.

[8] Emanuele Pesce et al. "Learning to detect chest radiographs containing pulmonary lesions using visual attention networks". In: *Medical Image Analysis* 53 (Apr. 2019), 26â38. ISSN: 1361-8415. DOI: 10.1016/j.media.2018.12.007. URL: http://dx.doi.org/10.1016/j.media.2018.12.007.

[9] *Pytorch Adam Optimizer*. https://pytorch.org/docs/stable/optim.html. Accessed: 2021-04-26.

[10] *Pytorch BCELoss With Logit*. https://pytorch.org/docs/stable/generated/torch.nn.BCEWithLogitsLoss.html. Accessed: 2021-04-26.

[11] Pranav Rajpurkar et al. "CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning". In: *CoRR* abs/1711.05225 (2017). arXiv: 1711.05225. URL: http://arxiv.org/abs/1711.05225.

[12] Jonathan Rubin et al. "Large Scale Automated Reading of Frontal and Lateral Chest X-Rays using Dual Convolutional Neural Networks". In: *CoRR* abs/1804.07839 (2018). arXiv: 1804.07839. URL: http://arxiv.org/abs/1804.07839.

[13] Karen Simonyan and Andrew Zisserman. *Very Deep Convolutional Networks for Large-Scale Image Recognition*. 2015. arXiv: 1409.1556 [cs.CV].

[14] Petros-Pavlos Ypsilantis and Giovanni Montana. *Learning what to look in chest X-rays with a recurrent visual attention model*. 2017. arXiv: 1701.06452 [stat.ML].