An Application of DenseNet for Medical X-ray Classification

Joanna Halpern
joanna.halpern@mail.mcgill.ca
260410826

Talia Wise
talia.wise@mail.mcgill.ca
260659717

Chenghao Liu
chenghao.liu@mail.mcgill.ca
260736137

***Abstract*—Pneumonia is a potentially deadly disease that affects millions of people worldwide every year. It is commonly diagnosed through chest X-ray images and can often appear similar to other thoracic diseases, making automatic X-ray classification of pneumonia valuable to populations with inadequate accessibility to medical imaging specialists. CheXnet is a 121 layer dense convolutional neural network (DenseNet-121) that classifies images of chest X-rays labeled with 14 different diseases from the ChestX-ray14 dataset. In this paper we reproduce the CheXnet algorithm, test DenseNet's performance on binary classification between pneumonia and other diseases, and we perform robustness testing on the model. Ultimately, we find that only some of the results of the CheXnet study are reproducible within the time and budget constraints that we have. Our study, however, affirms their hypothesis that DenseNet-121 classifies the diseases in their dataset with high accuracy. We also find that CheXnet is robust to considerable amounts of noise and changes in contrast and brightness.**

I. Introduction

Independent replication of scientific studies has had an important impact in verifying the quality of accepted scientific results. But in computational sciences, which require the collection of large datasets that are expensive and time consuming to collect, such studies are often virtually impossible (Peng et al, 2011). As such, studies which aim to reproduce the algorithms and measurements of a study on the original dataset are valuable in fields such as computational radiology.

In this paper, we aim to reproduce the results achieved by the authors of CheXNet (Rajpurkar et al. 2017). The authors perform two studies on the ChestX-ray14 dataset (Wang et al. 2017) . First they compare the diagnoses of four different radiologists on a 420 image test set with the classification results of their CheXnet algorithm and find that CheXnet outperforms the average radiologist. They then use CheXnet to classify the ChestX-ray14 dataset and achieve better accuracy than any previous results on this dataset. The authors of this study did not release their code, their train-test split, or the four radiologists' diagnoses. As such, reproducing this study is not possible in its entirety.

We reconstruct CheXnet by implementing DenseNet-121 and adjusting it to a multi-class classification for the ChestX-ray14 dataset. We then test this classifier on the test set given by the ChestX-ray14 dataset and find that our implementation resulted in majority-class classification, achieving an F1-score of 0.033. We also use a similarly implemented model to classify pneumonia-positive vs. pneumonia-negative images, as well as pneumonia-positive vs. edema-positive images and find that our models achieve majority-class classification on each of these experiments as well, with F1-scores of 0.00.

One of the common concerns in X-ray imaging is controlling the amount of radiation the patient is exposed to. Higher radiation amounts produce better X-rays but are more dangerous to the patient. We test the CheXnet algorithm's robustness by introducing various common X-ray faults to the image data and checking to see how they affect accuracy. We find that some small amounts of additive noise or increased contrast mildly improve classification, while larger amounts of noise and reduced brightness decrease classification scores. This shows that the model is robust to some common faults in X-ray images.

II. Reproducibility Study

III. Theory

CheXNet

CheXNet is a 121 layer dense convolutional neural network that detects pneumonia from chest X-rays at a level exceeding practicing radiologists. The authors use DenseNet (Huang et al., 2016) to classify frontal X-rays with pneumonia against those without, and then compare these classification results with those of four radiologists, ultimately achieving better accuracy than the average radiologist. They also train Densenet to classify fourteen thoracic diseases, which include pneumonia and 13 other diseases that are often mistaken for pneumonia or appear similar in X-rays. They achieve better accuracy than state-of-the-art results on all 14 diseases.

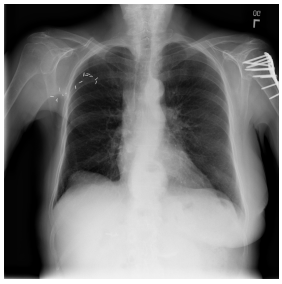
DenseNet

Densenet is a relatively new convolutional network in which each layer is connected to every other layer in a feed-forward fashion. For each layer, the feature maps of all previous layers are used as input instead of just the feature map from the one previous layer, as in regular neural networks. Densenet has been shown to achieve better than previous state-of-the-art results on common datasets like ImageNet. And even when the results are not better than state-of-the-art, Densenet still achieves similar levels of accuracy using many less parameters that other models. On top of substantially reducing the number of parameters, Densenet also encourages feature reuse since later layers do not have to relearn features learned by earlier layers since they access this information directly through concatenation (as opposed to resnet where the

features are added instead of concatenated), and it strengthens feature propagation since features learned by an early layer are directly accessible by later layers. The vanishing-gradient problem is alleviated since the greater connectivity between layers in Densenet allows each layer to have direct access to the original input signal as well as the gradients of the loss function.

IV. Methodology

Dataset



We used the ChestX-ray14 dataset from Wang et al., 2017. This dataset contains 112,120 frontal-view chest X-rays labeled with any combination of 14 different thoracic diseases. In the CheXnet paper, the authors randomly split the dataset into a train (98637 images), validation (1672 images) and test (420 images) with each patients images only belonging to a single category.

We found that the distribution of positive cases between the 14 different diseases varied widely. And surprisingly there where far fewer pneumonia-positive images than images of any other class, especially given that pneumonia classification is the main focus of the CheXnet experiment.

The following table shows the number of images positive for each class in the training set:

'Atelectasis'	7996
'Cardiomegaly'	1950
'Effusion'	9261
'Infiltration'	13914
'Mass'	3988
'Nodule'	4375
'Pneumonia'	978
'Pneumothorax'	3705
'Consolidation'	3263
'Edema'	1690
'Emphysema'	1799
'Fibrosis'	1158
'Pleural_Thickening'	2279
'Hernia'	144

Feature Extraction

The following feature extraction settings where used in the original CheXNet paper. The feature extraction in our models are similarly implemented.

1) downscale the images to 224 x 224
2) normalize based on the mean and standard deviation of the ImageNet dataset
3) augment the training set with random horizontal flipping

Training and Validation

Training and testing of the models was done on a Google Compute Cloud Instance which was provided by McGill University. This instance had 4 CPUs, 15 GB of RAM, 1 Nvidia K80 (16 GB) GPU, and 500 GB of storage. The code for training and testing the model was implemented using PyTorch. The implementation of a testing-only version of ChexNet in PyTorch (from Weng et al.) was used as a reference for our code as was an implementation of DenseNet in PyTorch (from https://github.com/andreasveit/densenet-pytorch).

First loaded the data and labels into a PyTorch Dataloader which we were able to do successfully. Next, we built the model in PyTorch using PyTorch's torchvision package which contains a 121 layer DenseNet model. We then changed the last layer of DenseNet-121 to a classifier with a linear layer followed by a softmax. This way the model would give predictions for each class of how likely it was that that image was in that class. To calculate the loss function we used cross entropy loss and we used a stochastic gradient descent optimizer with Nesterov momentum with a learning rate of 0.1, a momentum factor of 0.1, and weight decay (L2 penalty) of 1e-4. We used a mini-batch size of 16 because that is what they used in the original paper.

In the CheXnet paper, the model was pre-trained on the Imagenet dataset. The weights before training are initialized to Imagenet's weights rather than random ones and the X-ray images are normalized with the mean and standard deviation of the Imagenet dataset. This has been shown to help DenseNet converge on image datasets and is fairly standard practice so we use the pre-trained version of the model in all of our experiments.

We first trained DenseNet to classify pneumonia-positive and pneumonia-negative X-rays. We then trained a binary classification model on X-rays with Pneumonia and Edema. We conducted these experiments in order to compare binary classification of individual thoracic diseases to the multi-label classification used in the original CheXnet model.

We then trained a multi-class version of DenseNet to classify all 14 different thoracic diseases.

In each case, first we trained the model on a sample of 16 images to test for convergence, and then we trained the model on a larger training set. For each model that we trained, we printed the current precision and loss of both the training data and the validation data at each epoch. That way we were able

to save the model where the precision was the highest, and the loss the lowest for the validation. We were also able to stop training when we noticed that the precision was not increasing and the loss was not decreasing anymore.

V. Results

Pneumonia detection

The model classified every image as non-pneumonia. F1-score: 0.00

Pneumonia/Edema binary classification

The model classified every image as Edema. F1-score: 0.00

Multi-class classification

Test scores by class

	precision	recall	f1-score	support
Atelectasis	0.00	0.00	0.00	826
Cardiomegaly	0.00	0.00	0.00	243
Effusion	0.00	0.00	0.00	835
Infiltration	0.38	0.03	0.05	1893
Mass	0.07	0.09	0.08	413
Nodule	0.00	0.00	0.00	587
Pneumonia	0.00	0.00	0.00	59
Pneumothorax	0.07	0.88	0.13	439
Consolidation	0.00	0.00	0.00	241
Edema	0.00	0.00	0.00	110
Emphysema	0.00	0.00	0.00	191
Fibrosis	0.00	0.00	0.00	154
Pleural_Thickening	0.00	0.00	0.00	247
Hernia	0.00	0.00	0.00	21
avg / total	0.12	0.08	0.03	6259

Average (macro) F1 Score:0.01874694479676278

```
---------------- TRAIN DATA ------------------
---- Classification Report -----
                    precision    recall  f1-score   support

       Atelectasis       0.00      0.00      0.00      2942
      Cardiomegaly       0.00      0.00      0.00       754
         Effusion        0.00      0.00      0.00      2730
      Infiltration       0.41      0.03      0.06      6663
              Mass       0.09      0.14      0.11      1481
            Nodule       0.00      0.00      0.00      1883
         Pneumonia       0.00      0.00      0.00       210
      Pneumothorax       0.07      0.87      0.13      1552
     Consolidation       0.00      0.00      0.00       952
             Edema       0.00      0.00      0.00       479
         Emphysema       0.00      0.00      0.00       635
          Fibrosis       0.00      0.00      0.00       491
Pleural_Thickening       0.00      0.00      0.00       759
            Hernia       0.00      0.00      0.00        71

       avg / total       0.14      0.08      0.04     21602

---- F1 Scores -----
[0.         0.         0.         0.06260388 0.10518094 0.
 0.         0.1338143  0.         0.         0.         0.
 0.         0.        ]
---- Average (macro) F1 Score -----
0.021542794365561512


---------------- VALIDATION DATA ------------------
---- Classification Report -----
                    precision    recall  f1-score   support

       Atelectasis       0.00      0.00      0.00       444
      Cardiomegaly       0.00      0.00      0.00        97
         Effusion        0.00      0.00      0.00       394
      Infiltration       0.34      0.03      0.05       995
              Mass       0.11      0.14      0.13       244
            Nodule       0.00      0.00      0.00       236
         Pneumonia       0.00      0.00      0.00        38
      Pneumothorax       0.07      0.89      0.13       208
     Consolidation       0.00      0.00      0.00       121
             Edema       0.00      0.00      0.00        45
         Emphysema       0.00      0.00      0.00        69
          Fibrosis       0.00      0.00      0.00        82
Pleural_Thickening       0.00      0.00      0.00       121
            Hernia       0.00      0.00      0.00        18

       avg / total       0.12      0.08      0.03      3112
```

Average (macro) F1 Score: 0.021603016270971945

VI. Discussion

While at first it seemed like reproducing the experiments in the CheXnet paper would be fairly straightforward, we found that the paper did not go into sufficient detail to allow us to fully reproduce the results. Implementing DenseNet is not nearly as straightforward as many older and more well established models, and we had to build some parts of the model ourselves just in order to get it to run. Additionally, the CheXnet authors did not release the way they split their data. It is unclear in their paper whether they used the test set that is distributed with the ChestX-ray14 dataset or if they created their own test set. In addition, the CheXnet authors conducted an experiment comparing their model to diagnoses given by four different radiologists on their test set, and they did not provide these extra labels, making this part of the experiment impossible unless one has access to multiple physicians. We emailed the authors of the CheXnet paper in order to try to gain access to the code of their model (in order to perform an ablation study) and perhaps gain more information about their dataset, but have not received a response.

Particularly, modifying DenseNet from a binary classifier into a multi-label classifier was a fairly involved process and there are several different ways one might do so. The authors of CheXnet, however, do not explain their methodology in doing so at all. They only write that they made ”simple modifications” to the model. We did find an already-implemented model of CheXnet created by a different team than the original authors on github, but the creators of this model only released a pre-trained version that can be used for classifying chest X-rays, and did not release the code for building and training this model (Weng et al. - This is the model which we use for robustness testing).

To reporoduce the original paper we would need more time to run experiments as well as more expertise on multi-label classification. The original paper was difficult to reproduce from the information given.

VII. Model Robustness Study

VIII. Theory

X-ray images often include noise and varying brightness and contrast due to difference in equipment and radiation doses. In particular, for digital images, Gaussian noise originates from data acquisition such as sensor noise, while Poisson noise is generated from fluctuating electrical currents, and salt-and-pepper noise may arise from defective pixels. X-ray noise is usually modeled through Gaussian and Poisson distributions (Gravel et al. 2004). We also tested the effects of salt and pepper noise and speckle noise as controls.

Theoretically, robustness of an algorithm with respect to image input can be tested with images that contain varying levels of commonly observed noise as well as images that have

differing brightness, contrast, and sharpness. Moreover, should the algorithm perform better with additive noise on the images, fine tuning using these image pre-processing parameters can be used to achieve higher test accuracies in practice.

IX. Methodology

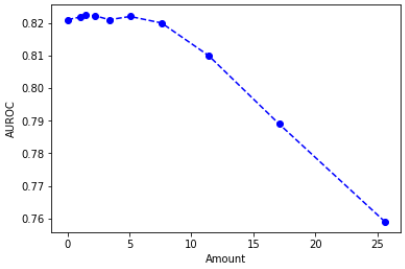
483 frontal chest X-ray images were randomly selected from the test dataset and then tested on the implemented CheXNet algorithm. Varying amounts of Gaussian noise, Salt-and-pepper noise, Poisson noise, and Speckle noise was added to these images separately. We should note that the first two noise distributions are more commonly observed in X-ray images and the results are tested against that of Speckle noise, a noise distributions that is not commonly appear in X-ray images as baselines, in which each pixel equals to the original pixel multiplied with a Gaussian distribution added towards the original pixel.

These noisy images were then classified into 14 different diseases using the CheXNet model, the weights of the network are obtained from the checkpoint file generated by the optimal validation performance conducted by Weng et al. (Github). The classification accuracies for each category is presented as the area under the receiving operating characteristic curve (AUROC), and the average AUROC over 14 disease categories is reported.

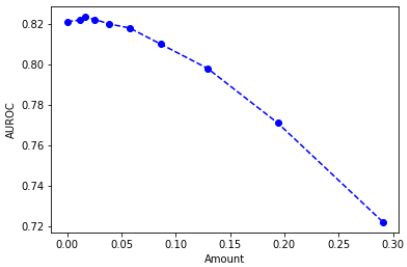
Following this, brightness, contrast, sharpness are altered for each of these 483 images, and are classified into disease categories using the same method as described above. We then compared the difference between the observed trends in each of these alterations.

X. Results

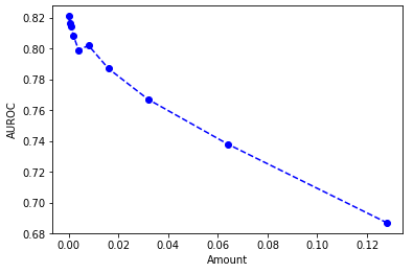
- Gaussian noise



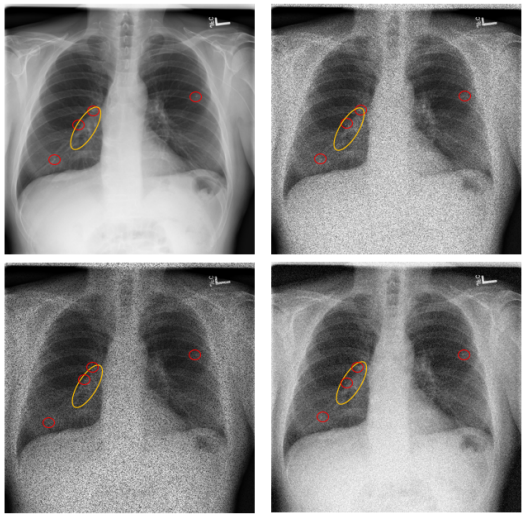
- Speckled noise



- Salt and pepper noise



- Noise comparisons with sample image



Upper left: normal image showing infiltration and nodules with nodules in red circle and infiltration in orange circle; upper right: speckled noise with amount 0.28, infiltration obvious, nodule not so obvious; Lower left: salt-and-pepper noise with 0.12 density, nodule not obvious, infiltration not so obvious; lower right: gaussian noise with sigma of 35, infiltration and nodule both obvious

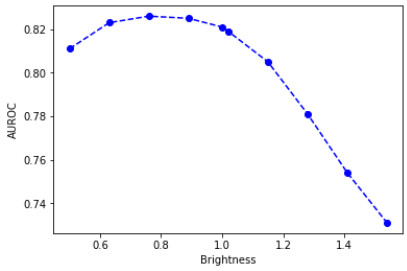
The plots above show varying amounts of Gaussian noise, speckled noise, and salt-and-pepper noise against average AUROC. The X-axis corresponds to accuracies achieved with 0 being the original images. The amount of Gaussian noise corresponds to sigma with mean and variance of 0 and 1, while amount of speckled noise corresponds to the intensity of additive Gaussian noise, and amount of salt-and-pepper corresponds to the density of combined white and black pixels. We should note that to achieve full image-dependency, no scale was added for Poisson noise, and thus it only returned one AUROC value of 0.823, slightly higher than that of original images' AUROC of 0.821.

As readily seen, the average AUROC does not seem to be much affected by Poisson noise. It also stays relatively constant or even increases slight when minor Gaussian noise and speckled noise is added, and only after considerable amount of noises is added, the AUROC starts to decrease. Even at very high noise levels such as sigma = 35 for Gaussian noise, the model still achieved a reasonable AUROC - on par with previous state-of-the-art X-ray classification results from Wang et. al 2017, with average AUROC of 0.738. However, the average AUROC is much less resistant to that of salt and pepper noise, showing large decreases immediately after the noise is added. Regardless of the decrease at the end, we can see that the algorithm is fairly resistant to Poisson and

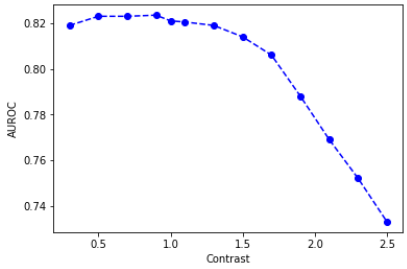
Gaussian noise which commonly arise in X-rays, while it is much less resistant to the more random speckles and salt and pepper noise.

Below we plot the effect on AUROC by varying brightness, contrast, and sharpness. In these plots original performance is at 1.0, while less than 1.0 implies darkening the image, decreasing the contrast, or decrease the sharpness, and more than 1.0 increases all of these respectively. Lower brightness led to a small improvement in performance while increasing the brightness results in much lower ability to classify correctly. Surprisingly, increasing contrast did not increase the performance but rather lowered it, while decreasing the contrast did not seem to damage the performance by a lot. At last, sharpness did not influence the average AUROC by much as fluctuations in performance were within 2%.

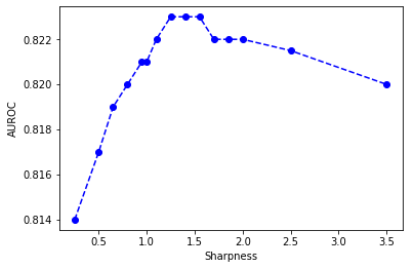
- Brightness



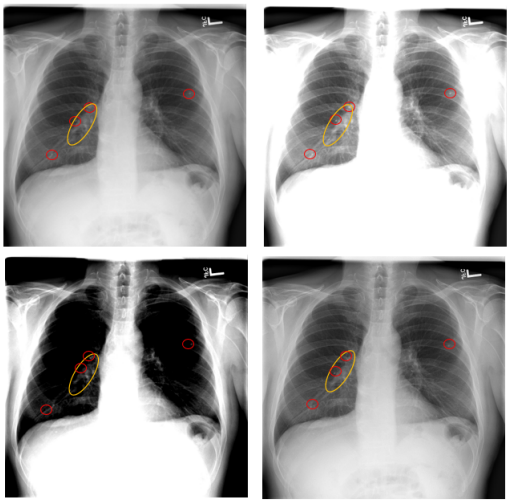
- Contrast



- Sharpness



- Brightness, contrast, and sharpness comparison with sample image



Upper left: normal image showing infiltration and nodules with nodules in red circle and infiltration in orange circle; upper right: brightened image with 1.5 scale, infiltration and nodule both obvious; Lower left: image with high contrast with 2.5 scale, nodule not obvious, infiltration not so obvious; lower right: sharpened image with 1.5 scale, infiltration and nodule both somewhat obvious

XI. Discussion

To begin with, we shall discuss the observed differences in the CheXNet resistance to different types of noises. Gaussian noise treated images transforms its pixels in such a way that it becomes the sum of itself and a Gaussian distribution with mean 0 and standard deviation sigma (amount). With low standard deviation, dark/bright regions of the X-ray will remain relatively dark/bright, and thus only minorly influences the X-ray classification performance. In particular, at low standard deviation, small fine structures are erased which might smooth out irregularities. The same is true for speckled noise. Salt-and-pepper noise, however, acts to randomly turn pixels completely black or white, which greatly disturbs the ability of CheXNet to classify the images. Healthy lungs look black in X-rays so adding a mix of black and white noise may generate white spots in healthy areas - potentially indicative of diseases, and black spots in unhealthy areas - potentially ignoring the illness, thus reducing classification accuracy. To further explain the difference in the apparent ability of CheXNet to resist Gaussian noise better than that of speckled noise, we will briefly look at the generation of speckled noise. An image treated with speckled noise will have its pixels become the sum of itself and a Gaussian distribution multiplied by itself, which implies that there would be a large fluctuation in the pixel intensity in the output image. In other words, dark pixels may become very dark or fairly bright, and vice versa, causing the pre-trained network to mis-classify on the images. Poisson noise results in small difference between the final and initial image and thus lead to negligible change in performance.

To examine the differences resulting from altering brightness, contrast, and sharpness, we can employ similar arguments. For brightness, severely darkening the image will lead to indistinguishability of bright and dark spots which CheXNet uses to classify, leading to lower performance. Increasing the exposure will lead to the revelation of many previously hard-to-identify fine structures that are unrelated to

any disease, leading CheXNet to mis-classify. Interestingly, we observe that lower brightness to some extent improves the performance; we believe that this is a result of lowering intensity of moderately-bright regions that are not disease-indicative while disease-indicative regions that are relatively bright are affected less(Dobbins et al. 2003). We should note that this is not a result of increased contrast as increasing contrast merges nearby structures while darkening/brightening with neighborhood majority-voting (Graham et al. 2005). This is evident in the following graph where we see that increasing the contrast compromises the performance severely - in many diseases such as small (but not fine) nodules, increasing contrast directly causes them to disappear, and therefore result in low performance. At last, sharpness does not seem to affect the performance significantly as it does not erase or add any structures but only creates more obvious divisions between them, which are already sufficiently obvious with frontal chest X-rays that have a large black area (the lung).

XII. Conclusion and further areas of inquiry

Our model demonstrates that DenseNet is able to converge on the ChestX-ray14 dataset. But since we where unable to properly train the model, the reproducibility of the CheXnet multi-label classification algorithm remains inconclusive. The reproducibility of the pneumonia detection experiment which compares the results of ChexNet with the diagnoses of four physicians is not possible because the authors have not released the necessary data. Additionally, we find that the author's have not sufficiently described the implementation of the CheXnet algorithm in their paper in order for it to be reproduced in the time frame that we had.

Because X-ray classification has potentially life saving real-world significance, future studies aiming to improve the CheXnet algorithm would be valuable. Based on our robustness testing, we suspect that training the model with added Gaussian noise and slightly decreased brightness on the images might improve classification accuracy.

In future ablation studies of the CheXnet algorithm, we recommend testing several decisions which have been made by the original authors in their implementation of the CheXnet algorithm. First, it is worth testing the effect of different numbers of dense layers on classification accuracy as there are four different DenseNet models commonly available with different numbers of layers. Second, we recommend comparing results using Adam and SGD, since the CheXnet paper uses Adam but it has been shown that SGD usually leads to much better results than Adam (Wilson et. al. 2017).

Additionally, CheXnet only uses frontal X-rays, while physicians make their diagnoses using both frontal and lateral chest X-rays (Rajpurkar et. al. 2017). So for clinical purposes, it would be worth training a model using both of these, as well as perhaps including other useful patient data.

XIII. Contributions

Joanna, Talia and Peter worked on project conceptualization and planning. Joanna worked on the CheXnet reproduction code with contributions from Talia. Peter worked on the robustness study. Talia wrote the paper with contributions from Joanna and Peter.

XIV. Bibliography

Young, Cristobal, and Katherine Holsteen. "Model Uncertainty and Robustness: A Computational Framework for Multimodel Analysis." Sociological Methods and Research 46.1 (2017): 3-40.

Dobbins III, James T., and Devon J. Godfrey. "Digital x-ray tomosynthesis: current state of the art and clinical potential." Physics in medicine and biology 48, no. 19 (2003): R65.

Ulin, Matej, Jens Lundstrm, and Stefan Byttner. "Robustness of deep convolutional neural networks for image recognition." International Symposium on Intelligent Computing Systems. Springer, Cham, 2016.

Gravel, Pierre, Gilles Beaudoin, and Jacques A. De Guise. "A method for modeling noise in medical images." IEEE Transactions on medical imaging 23.10 (2004): 1221-1232.

Huang, Gao, et al. "Densely connected convolutional networks." Proceedings of the IEEE conference on computer vision and pattern recognition. Vol. 1. No. 2. 2017.

Graham, Richard NJ, R. W. Perriss, and Andrew F. Scarsbrook. "DICOM demystified: a review of digital file formats and their use in radiological practice." Clinical radiology 60, no. 11 (2005): 1133-1140.

Rajpurkar, Pranav, et al. "CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning." arXiv preprint arXiv:1711.05225 (2017).

Wang, Xiaosong, et al. "Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases." 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2017.

Dodge, Samuel, and Lina Karam. "Understanding how image quality affects deep neural networks." Quality of Multimedia Experience (QoMEX), 2016 Eighth International Conference on. IEEE, 2016.

Peng, Roger D. "Reproducible research in computational science." Science 334.6060 (2011): 1226-1227.

Wilson, Ashia C., et al. "The marginal value of adaptive gradient methods in machine learning." Advances in Neural

Information Processing Systems. 2017.

Xinyu Weng, Nan Zhuang, Jingjing Tian and Yingcheng Liu https://github.com/arnoweng/CheXNet

https://github.com/andreasveit/densenet-pytorch

XV. Appendix

Note:

We had originally planned to do an ablation study on CheXnet but then realized that the code we found online was not a full implementation of that algorithm, but rather just a pre-trained model that could be used by someone wishing to classify thoracic X-rays (and could be used to confirm the test-set results). Because of this we ended up having to reproduce the model anyways, so our project became a combination of our originally planned ablation study and a reproducibility study.

Applications of DenseNet for Medical X-ray Classification

Joanna Halpern, Talia Wise and Chenghao Liu

Abstract- Pneumonia is a potentially deadly disease that affects millions of people worldwide every year. It is commonly diagnosed through chest X-ray images and can often appear similar to other thoracic diseases, making automatic X-ray classification of pneumonia valuable to populations with inadequate accessibility to medical imaging specialists. CheXnet is a 121-layer dense convolutional neural network (Densenet-121) that classifies images of chest X-rays labeled with 14 different diseases from the ChestX-ray14 dataset. In this paper, we reproduce the CheXnet algorithm and also test DenseNet's performance on binary classification between pneumonia and other diseases as well as perform robustness testing on the model. Ultimately, we find that only some of the results of the CheXnet study are reproducible within the time and budget constraints that we have. Our study, however, affirms their hypothesis that DenseNet-121 classifies the diseases in their dataset with high accuracy. We also find that CheXnet is robust to small amounts of noise and changes in contrast.

Pneumonia detection- We trained DenseNet on both pneumonia vs. non-pneumonia images, and on pneumonia vs. edema images. In the pneumonia vs. non-pneumonia classifier, the model attributed all images to the pneumonia-free class, despite it being a balanced data set. We suspect that this is because the pneumonia-free class contained images with 13 different diseases leading it to be extremely varied. The Pneumonia vs. Edema model attributed all images to the majority class after only one epoch and did not change after that.

Multi-class Classification- We then trained the model to classify X-rays with all fourteen different diseases in our dataset, and found again that the model was only able to achieve majority-class classification accuracy. We suspect that further tuning of the model's parameters and hyper-parameters would have fixed this issue.

Model Robustness Testing- Various adjusted frontal chest X-ray images were tested on a already-implemented CheXNet model, available for X-ray testing (but not training) on GitHub. Since noise in X-ray images is commonly modeled with a Gaussian distribution, images in the test set were adjusted to include varying amounts of Gaussian noise, as well as varying amount of other types of noise as controls. Brightness and contrast also differ in X-rays due to differences in equipment and radiation doses, so these were altered to simulate varying amounts of radiation as well. Average AUROC was then recorded on all adjusted test sets. It was found that that the CheXNet model is robust to Gaussian noise (with mean 0 and variance 1) up to sigma=8, but not to other types of noise. Notably, at high noise levels of all types of noise distributions, the average AUROCs are still comparable to that of previous state-of-the-art performance (0.76-0.69 compared to 0.73). It was also found that the performance of the model can be slightly improved by using brightness-reduced images, but appears to be always negatively affected by increase in contrast. This shows that the model is fairly robust to differing qualities of X-rays.