Automated Classification Using End-to-End Deep Learning

*Shobhit Sandeep Jaipurkar[1,3], Wang Jie[1,3], Zeng Zeng[1,2,4], Teo Sin Gee[4], Bharadwaj Veeravalli[4], Matthew Chua[3]

Abstract*— According to a study [1] by the Ministry of Health in Singapore, since 2009, cancer, ischaemic heart disease and pneumonia together accounted for approximately 60% of the total causes of death. It has been 9 years, and Pneumonia and other Acute Upper Respiratory Infections still is one of the top 10 conditions of hospitalization. In cases of respiratory diseases such as Chronic Obstructive Pulmonary Disease (COPD), it has been found that close to 55% of cases are misdiagnosed. This is shocking as, an early diagnosis of respiratory diseases can lead to an earlier treatment intervention, ultimately lessening symptoms, slowing the progression, and improving overall quality of life. With the advent of Deep Neural Network architectures which have shown phenomenal results in the field of Computer Assisted Diagnosis (CAD), we hope to implement a Lung Classification Model using End-to-End Deep learning to classify Chest X-Ray images into one of 14 primary classes of lung diseases. Using our implementation of the Densely Connected Convolutional Neural Network model architecture, we aim to increase existing model accuracy in Lung Disease classification by iteratively reducing the search space and region of interest for different. We shall experiment on a 14-class classification model and compare the results with a binary classifier as well, to understand the performance of DenseNets on Chest X-Ray (CXR) Data with a reduced search space.

I. Introduction

With the introduction of deeper architectures in Deep Neural Networks, there has been a massive uptick in the implementations of complex image classification in the medical field. We believe that the point that takes Computer-Aided Diagnosis to the next level, will rely solely on the optimization methodology used, and not the model construction. With this in mind, we propose a method to efficiently optimize the accuracy of a DenseNet trained to classify major lung diseases from X-Ray images.

In medical imaging, we notice that most images have highly heterogeneous class values. As described by this survey on deep learning methods for medical imaging [18], normal categories contains a multitude of normal tissue, as well as important benign details. This leads to models being able to accurately classify common subclasses, but failing on rare diagnostic details.

One possible solution would be to improve the dataset with detailed annotations of all possible subclasses and running a multiclass end-to-end model. However, the limited availability of expert time for providing annotations leads to this approach being infeasible. This imbalance has been tackled earlier by incorporating intelligence into the training process itself, instead of applying data pruning and augmenting processes. However, we cannot use these methods as, while, these strategies work on an ideal noiseless database, they usually fail when there is substantial noise in real-world data.

With our proposed model architecture using Densely Connected Convolution Networks, we hope to address this issue with the enhanced feature reuse and feature propagation capabilities of the DenseNet. Furthermore, to improve classification accuracy, we will experiment with a method to iteratively reduce the search space in an attempt to improve speed of computation, classification accuracy and reduce the dimensionality of the data.

II. Related Work

Research in machine learning has always been heavy in the medical imaging sector with constant work done across multiple disciplines. We can observe the massive uptick in research through recent years from Figure 1 taken from A Survey on Deep Learning in Medical Imaging Analysis [18]

The earliest implementation of a deep neural network in the field of CXR can be traced back to Rajkomar et al., who used a GoogLeNet architecture with image augmentation and pre-training on ImageNet to classify CXR images as either frontal or lateral with 100 percent accuracy [2]. While the implementation massively varies from the work we are keen on doing, this can arguably be defined as the point which demonstrated the usability of deep learning on CXR images. Other relevant research includes using deep learning in similarity ranking framework. Published by Anavi et al., the paper sought to create a network that could, given a query image, rank the other CXR images in its database by similarity to the query. They found that a 5 layer convolutional neural network was much more effective than similarity based on Image descriptors [3]. Such a network could be used to help clinicians search for past cases easily and help inform teir current or future diagnosis.

*The work was supported by Singapore-China NRF – NSFC Grant [No. NRF2016NRF-NSFC001-111]

[1] Co First Authors
[2] Corresponding Authors
[3] National University of Singapore, Singapore
[4] Agency for Science, Technology and Research (A*STAR), Singapore

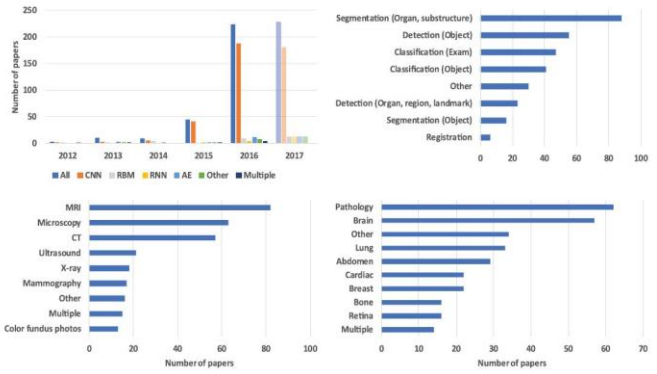

Fig. 1: Breakdown of the published papers included in the 2017 survey by Geert Litjens et al.[18]

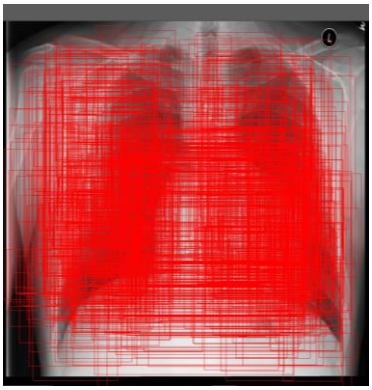

Fig. 2: Plot of all Bounding Boxes on a sample image.

However, the major breakthrough in diagnostic deep learning in research came in 2016 when Shin et al. used a Convolutional Neural Network (CNN) to detect specific diseases in CXR images and assign disease labels. They then used a Recurrent Neural Network to describe the context of the annotated disease based on the features of the CNN and patient metadata [4]. While the validation accuracy was only 0.698, they were still able to provide the relevant groundwork for future research. This performance may be largely due to their relatively small data set size of 7470 images, the challenges of multi-class classification, and incorporating textual data from patient records. Most recently, in 2017, Wang et al. successfully designed a CNN to diagnose specific diseases by detecting and classifying lung nodules in CXR images with high accuracy [15].

Once the ball got rolling, there have been plenty of advancements with deep learning in CAD, the most notable one yet being the CheXNet by Pranav Rajpurkar et al [6]. In this paper, they have implemented densely connected convolutional neural networks to identify and detect pneumonia from a large CXR dataset of close to 112,120 images, with an accuracy which is comparable to an expert radiologist. The implementation of CheXNet is currently the gold standard in CXR disease classification tasks.

III. Methodology

A. *Network Selection and Overall Architecture*

As discussed earlier, the classification tasks associated with CAD require a healthy mix of both, local and global information during feature learning, since CXR images and their respective diagnoses cannot be accurately classified based on a particular region of interest. Hence, we require a network that shows a high amount of information flow and feature reuse.

Upon comparing ResNets [5], HighwayNets [9] and DenseNets [19], we observe that, in addition to strengthening feature propagation and encouraging feature reuse, Densely Connected Convolutional Networks, or DenseNets, are ideal for our use-case because they handle the vanishing gradient problem while simultaneously reducing the overall number of parameters. The key characteristic of DenseNet is the fact that it creates shorter connections from early layers to later layers, thus promoting the ease of feature reuse. This is the main motivation behind using the DenseNet as the classifier in our model.

The overall training and testing pipeline can be seen in Figure 3. We make use of data defining Bounding Boxes for a sample size of 984 images to determine a maximum bound for the region of interest. Iteratively, we then crop images across each direction by a certain unit and retrain our DenseNet to evaluate the model accuracy for the particular crop.

For our model implementation, we use a 121 layer Keras Implementation of DenseNet with the growth rate hyper parameter set at 32, with a batch size of 16 over 10 epochs per iteration. For the cropping iteration process, we currently have 40 iterations in each direction with a crop unit size of 30 pixels. The metrics we shall be measuring will mainly be validation accuracy, precision, recall and F-Score.

B. *Data Description:*

We use the Chest X-ray Dataset of 14 Common Thorax Diseases from the National Institute of Health. This ChestX-ray dataset comprises 112,120 frontal-view X-ray images of 30,805 unique patients with the text-mined fourteen disease image labels (where each image can have multi-labels), mined from the associated radiological reports using natural language processing. The text-mined disease labels are expected to have accuracy >90%. The data set also contains bounding boxes of 984 images, which we shall use in our cropping iterations.

The data we have was found to be highly imbalanced with respect to the number of entries per class, as we can see from Figure 5. Hence, we apply some data pruning steps to truncate

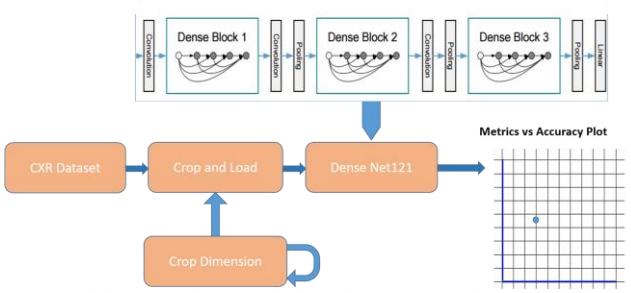

Fig.3 : Model Pipeline. The DenseNet breakdown is an example, and is not implicative of the usage of just 3 dense blocks

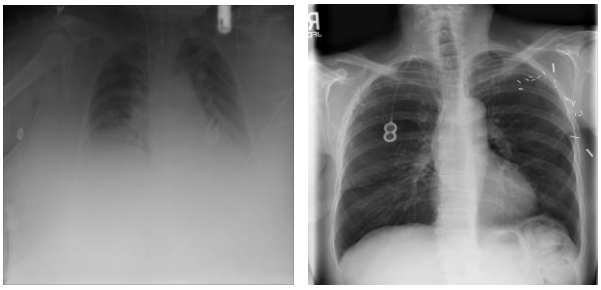

Fig. 4: Comparison of an image with Low SNR (left) to one with a high SNR (right)

the database entries to avoid a class imbalance problem during classification.

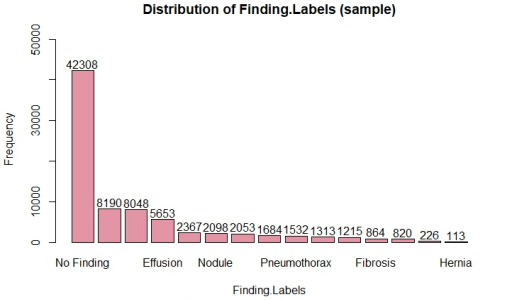

Fig 5: Unbalanced Class distribution of the CXR Dataset

C. *Preprocessing and Image Loading*

In the case of multiple diagnoses, we use a naïve approach to select the first diagnosis as the primary class identifying label. In an attempt to increase the overall quality of images we feed to the network, we calculate the Signal to Noise Ratio (SNR) of all the images in the database and use the image with the maximum SNR, for items with the same value of Patient ID.

In the case of medical image data, the ability to detect an aberrant object (lesion) in a radiograph is related to the ratio of the differential intensity to the ambient noise level. This ratio is called the absolute contrast to noise ratio, or the image signal to noise ratio. We then apply a base-10 logarithm to the ratio to obtain a normalized value of SNR.

$$SNR = \frac{\Delta \bar{I}}{\sigma_1}$$

The differential intensity of our image would be close to 255 as we are using a grayscale image. The ambient noise level can be approximated to the standard deviation of the background intensity due to noise processes. The SNR of every image is calculated and appended to the corresponding row as a new column. In figure 4, we see a comparison of an image with a high SNR to that of a low SNR

We also remove all entries of images with Finding Labels of 'No Findings'. In doing so, and applying the selection of one image per Patient ID based on SNR, we are able to prune the database from 112,120 class unbalanced images, to 13,499 images with relatively better class balance and good SNR, as we can see from Figure 6.

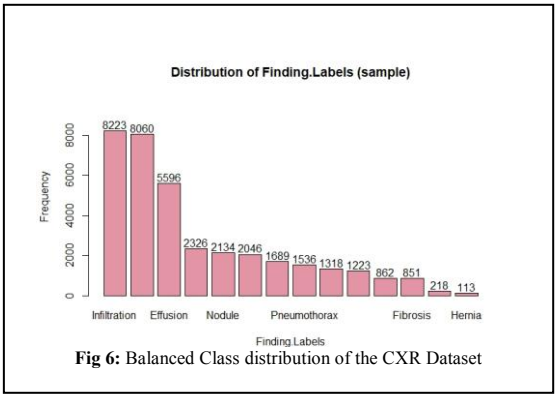

Fig 6: Balanced Class distribution of the CXR Dataset

D. *Freeze Layers and Data Augmentation*

Since we used pre-trained weights from the ImageNet database for our DenseNet training, we need to fine-tune these weights for our specific use case. However, since the first few layer weights trained on the ImageNet dataset accurately capture the universal features such as curves and boundaries, we shall choose to keep those weights intact.

We, then fit the model by minimizing the cross-entropy loss function using stochastic gradient descent using findings from a paper by T. Vatanen et al.[10] From the study on the class imbalance problem in Convolutional Neural Networks by Mateusz et. Al [20], we notice there is a chance for overfitting the data due to a class imbalance problem. Hence we implement a data augmentation method wherein, we horizontally flip the images to increase our data set for training.

IV. EXPERIMENTAL RESULTS

To test out our model, we have used the Keras library with a Tensorflow backend. A single iteration including loading, cropping and training the model takes close to 45 minutes. We first tried to use a 14-class classifier to gauge the results. However, we were not able to get accurate results for that experiment. This is probably due to a problem of class imbalance between the 14 different labels, and also due to the

fact that we are not provided with any prior information on the patient's history.

Model which tried multi class classification had the validation accuracy plateau at 31% and did not increase with subsequent crop iterations. Different loss functions or learning rates did not show much of change in the results.

We then implemented a binary classifier for each of the 14 classes to see if there was an observable difference in accuracy. With binary classifiers, we were able to see a massive upgrade in validation accuracy in 12/13 available classes, as compared to the CheXNet approach. In this current tabulation, we simply focus on the baseline accuracy, which is the accuracy we get when the image is cropped to the maximum rectangle as derived from the Bounding Box data

Diagnosis	CheXNet Accuracy	Max Crop Accuracy
Cardiomegaly	0.9248	**0.9314**
Consolidation	0.7901	**0.9538**
Edema	0.8878	**0.9744**
Effusion	0.8638	0.8465
Emphysema	0.9371	**0.9694**
Fibrosis	0.8047	**0.9642**
Hernia	0.9164	**0.9958**
Infiltration	0.7345	**0.7413**
Mass	0.8676	**0.9420**
Nodule	0.7802	**0.9177**
Pleural Thickenining	0.8062	**0.9691**
Pneumonia	0.7680	**0.9927**
Pneumothorax	0.8887	**0.9719**

Table 1: Results from the 14 binary classifier models, where we compare the our validation accuracy to that obtained from CheXNet.

V. Conclusion

We deployed a pipeline to conduct automated chest disease detection and classification. The performance on cropped images with same model architecture has higher accuracy than the baseline and previous state-of-the-art result in 13 of the 14 classes.

Limitations with this approach include the fact that a diagnosis does not depend solely upon the images, but can also be inferred from patient history, age, gender, living conditions etc. While we use high quality images with a great Signal to Noise Ratio, we need to employ other input variables to categorize images with a lot of noise. The class imbalance problem also exists greatly on the current data set, which might be a cause for relatively lesser accuracy.

References

[1] https://www.moh.gov.sg/content/moh_web/home/statistics/Health_Facts_Singapore.html

[2] Rajkomar, A., Lingam, S., Taylor, A.G., Blum, M., Mongan, J., 2017. High-throughput classification of radiographs using deep convolutional neural networks. J. Digit. Imaging 30, 95–101. doi: 10.1007/s10278- 016- 9914- 9 .

[3] Y. Anavi, I. Kogan, E. Gelbart, O. Geva, and H. Greenspan. A comparative study for chest radiograph image retrieval using binary texture and deep learning classification. Conf Proc IEEE Eng Med Biol Soc, 2015:2940–2943, Aug 2015.

[4] Kirk R. Le L. Dina D.F. Jianhua Y. Shin, H. and Ronald M. S. Learning to Read Chest X-Rays: Recurrent Neural Cascade Model for Automated Image Annotation. Conference on CVPR.Kunio Doi. Computer-aided diagnosis in medical imaging: historical review, current status and future potential. Computational Med Imaging Graph, 2015

[5] Kaiming He, Xiangyu Zhang, Shaoqing Ren, Jian Sun. Deep Residual Learning for Image Recognition. Microsoft Research, 2017

[6] Pranav Rajpurkar , Jeremy Irvin, Kaylie Zhu, Brandon Yang, Hershel Mehta. CheXNet : Radiologist Level Pneumonia Detection on Chest X-Ray with Deep Learning, *arXiv: 1711.05225,* 2017

[7] Tony Duan, Daisy Ding, Aarti Bagulm Robyn L. Ball, Curtis Langlotz, Katie Shpanskaya, Matthew P. Lungren, Andrew Y. Ng, Yupeng Chen, Jianan Li, Huaxian Xiao, Xiaojie Jin, Shuicheng Yan, Jiashi Feng. Dual Path Networks. *arXiv:1707:01692v2,* 2017

[8] K. He, X. Zhang, S. Ren, and J. Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In ICCV, 2015.

[9] R. K. Srivastava, K. Greff, and J. Schmidhuber. Highway networks, *arXiv:1505.00386,* 2015

[10] T. Vatanen,T.Raiko, H.Valpola, and Y LeCun. Pushing stochastic gradient towards second order methods-backpropagation learning with transformations in nonlinearities. In *Neural Information Processing,* 2013

[11] Medical image classification with convolutional neural network," *2014 13th International Conference on Control Automation Robotics & Vision (ICARCV)*, Singapore, 2014, pp. 844-848. doi: 10.1109/ICARCV.2014.7064414

[12] Liang-Chieh Che, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Tuille. Deeplab: Semantic Image Segmentation with deep convolutional nets, atrous convolution, and fully connected crfs, *arXiv preprint arXiv:1606.00915,* 2016

[13] Tianqi Chen, Mu Li, Min Lin, Naiyan Wang, Minjie Wang, Tianjun Xiao, Bing Xu, Chiyuan Zhang and Zheng Zheng, Mxnet: A flexible and efficient machine learning library for heterogeneous distributed systems *arXiv preprint arXiv:1512.01274,* 2015

[14] Geoff Pleiss, Danlu Chen, Gao Huang, Tongcheng Li, Laurens van der Maaten, and Kilian Q Weinberger. Memory efficient implementation of densenets. *arXiv preprint arXiv:1707.06990,* 2017

[15] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556.* 2014

[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition,* pages 770-778, 2017

[17] M. D. Zeiler and R. Fergus. Visualizing and Understanding convolutional neural networks. In *ECCV,* 2014

[18] Geert Litjens ∗, Thijs Kooi , Babak Ehteshami Bejnordi , Arnaud Arindra Adiyoso Setio , Francesco Ciompi, Mohsen Ghafoorian, JeroenA.W.M. van der Laak, Bram van Ginneken, Clara I. Sánchez. A survey on deep learning in medical image analysis. URL

[19] Gao Huang, Zhuang Liu, Laurens van der Maaten, Kilian Q. Weingberger. Densely Connected Convolutional Networks. *arXiv: 1608.06993,* 2018

[20] Mateusz Buda, Atsuto Maki, Maciej A. Mazurowski. A Systematic Study on the Class Imbalance Problem in Convolutional Neural Networks. *arXiv:1710.0538,* 2017