ChestX-ray14 Project - Group 38

Ana Vitoria Rodrigues Lima *
Harvard University
anavitoria_rodrigueslima@
@fas.harvard.edu

Van Anh Le *
Harvard University
vananhle@g.harvard.edu

Albert Ge *
Harvard University
albertge@g.harvard.edu

Aloysius Lim *
Harvard University
aloysius_lim@g.harvard.edu

Abstract

In this project we aim to explore the recently released ChestX-ray14 dataset and utilize neural network architectures to classify a given chest X-ray image for one or multiple diseases among Atelectasis, Consolidation, Infiltration, Pneumothorax, Edema, Emphysema, Fibrosis, Effusion, Pneumonia, Pleural Thickening, Cardiomegaly, Nodule, Mass and Hernia. Also, we aim to utilize saliency maps to identify where in the chest this disease is present in a given patient. This project will also enable us to learn about loss re-weighting and transfer learning, as we will be continuing to train on existing architectures with pre-trained weights on the ImageNet dataset.

1 Introduction

Since images accounts for 90% of all medical data [5], healthcare is a field in which computer vision can provide significant benefits. One particularly popular application is using computer vision to perform diagnosis of X-ray images and detect diseases. In this project, we will build a model to predict common thorax diseases and pathologies using a large-scale database of chest X-ray images, ChestX-ray14. To do this, we will start from pre-trained image classification models and work on fine tuning model parameters. The focus of our project will be on designing an effective loss reweighing scheme to address the heavy class imbalance problem in the dataset. We plan to explore the evaluation our model performance at the macro level and class-level using difference metrics and model diagnostics tool such as saliency maps.

2 Literature Review

In the realm of radio-logical examinations, X-ray is a common screening process to utilize in screening for diagnosis of several lung diseases. Such images are rich of information and could be so valuable in training deep learning methods. In this sense, [11]'s contributions has been revolutionising for the field of medical imaging and deep learning. [11]'s contribution consists of the first large scale chest X-ray dataset, which initially was called 'ChestX-ray8' as it contained eight disease image labels but then was relabelled to 'ChestX-ray14' for fourteen diseases.

'ChestX-ray8' comprised of 108,948 frontal-view X-ray images of 32,717 unique patients with eight diseases: Mass, Nodule, Pneumonia, Pneumothorax, Infiltration, Effusion, Cardiomegaly and

*Equal Contribution

Atelectasis. Some patients could have one disease, multiple diseases or no disease. Out of these images, 983 images have been annotated with a box from the doctor assessing the patient and annotated in the picture where the disease is. In these 983 images there are 1600 annotated boxes but this information has not been used in the training process but rather only at testing.

As architecture, [11] proposed to use four popular architectures (AlexNet [4], GoogLeNet [10], VGGNet-16 [9] and ResNet-50 [2]) with pretrained weights trained on the Imagenet dataset [1]. On these, they performed network surgery in leaving out the fully connected layers and the final classification layers. Instead of these, they replace these by adding a transition layer, a global pooling layer, a prediction layer and a loss layer. Only the transition layers and the prediction layers are hence trained from scratch, whereas the rest of the model uses the pretrained weights and fine tunes them with the new current dataset.

Given the imbalance of the dataset, namely that 84,312 images had no findings of any disease in the dataset, they utilized a re-weighted cross entropy loss. Furthermore, given that some patients might have more than a disease, they implement these architectures to be multi-label architectures - meaning, each image is labeled to be a one hot encoding vector of zeros of length eighth (for eighth diseases, if healthy then it will be all zeros), where one appears in the disease found. With this structure, if multiple diseases are predicted, the output vector will be a eighth-length vector with multiple ones where those disease are found, and zero elsewhere.

[11] utilizes their global pooling layer and prediction layer for two tasks: classification and heat-map generation. The intent is not only to determine a diagnosis, but to also highlight where the disease is on the X-ray image. Where the saliency map passes a threshold chosen, they create bounding box around the most intense area of the heat-map, and compare it to the 983 images which have a ground truth box. The information of the boxes is hence used only in testing and in evaluation with the IoBB (Intersection over the detected B-Box area ratio) metric, not in training. To evaluate the classification in the testing dataset, they utilize ROC curves and AUC values for each of the four implemented models.

Further work has been made possible because of this dataset, namely CheXNeXt. Given this dataset, [6] has trained an architecture that far exceeds the accuracies of [11]. In [6], more than one disease has been correctly classified in testing with AUC values around 90%, and most of them perform at around 80%. Only three out of fourteen diseases have a classification AUC between 70% to 80% in the CheXNeXt model. On the other hand, none of the classification AUC from [11] exceeded 81% for any disease across any architecture but reached rather mostly around 50%, 60% or 70% at the most.

CheXNeXt achieves such impressive results by separating their structure in two steps. The first consists in relabelling the dataset, due to partially incorrect labels in the ChestX-ray 14 dataset. The relabelling has been done by an ensemble of NNets where the classification threshold has been set by the highest F1 score. In the second step, they train new networks (they chose a Denset architecture [3] of 121-layers pretrained on Imagenet [1]) on the relabeled training dataset - this led to an ensamble of 10 networks where the final predictions are the average of the prediction of each network . Also CheXNeXt utilizes a re-weighted loss due to the dataset imbalance, namely they utilize a per-class weighted binary cross entropy loss.

3 Exploratory Data Analysis

3.1 Data Description

Our dataset is the ChestX-ray 14 dataset, which consists of 112,120 frontal-view X-ray images of 30,805 unique patients. The data labels come from associated radiological reports and were mined using natural language processing. There are 15 distinct labels representing fourteen common pathologies and a "No Findings" label. The 14 common pathologies are Atelectasis, Consolidation, Infiltration, Pneumothorax, Edema, Emphysema, Fibrosis, Effusion, Pneumonia, Pleural Thickening, Cardiomegaly, Nodule, Mass and Hernia. Notably, a single image can have multiple labels.

The dataset is split into a training-validation set (77%) and a test set (23%). In addition to the images, there are two tabular data files. One of them contain image labels, patient ID, and patient characteristics (gender and age). The second tabular data contains x-y coordinates of bounding boxes on 984 images in the test set. The bounding boxes are created by doctors who hand-indicated the physical areas in the image that contain useful information for diagnosis.

For the purpose of Exploratory Data Analysis, we will rely only on the training-validation data set.

3.2 Data Summary and Visualization

3.2.1 Distribution of Labels

Out of 86,524 training-validation images, 85% of them have a single label and 15% of them have multiple labels. Figure 1 shows the distribution of labels in the training-validation images.

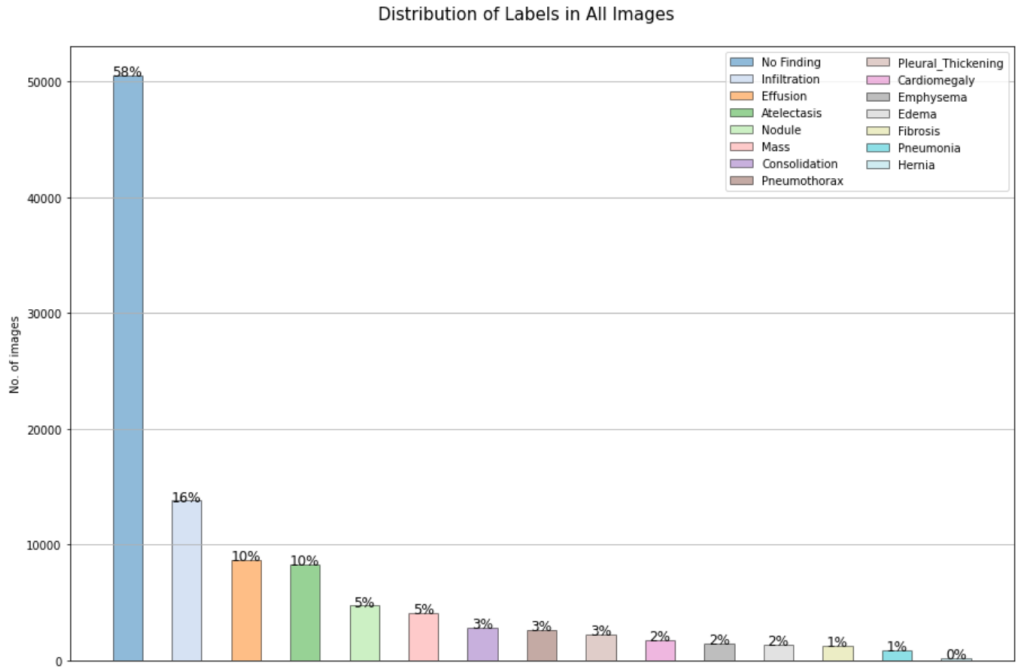

Figure 1: Distribution of Labels

Figure 1 demonstrates that our data exhibits significant class imbalance. More than half (58%) of the images in the training data set is "No Findings" while the other 14 classes are distributed amongst the other half. Most of the classes only occupy less than 5% of the data set. Hernia, the smallest category, accounts for less than 1% of the images in our data. Therefore, class imbalance is something that we should take into account as we built our model.

3.2.2 Distribution of Patient Characteristics

Next, we want to investigate the distribution of patient characteristics such as gender and age. This information could be useful to us when evaluating model fairness and performance after training and prediction.

(a) Patient Gender Figure 2 shows that 56% of the images belong to male patients while 44% of the images are from female patients. Gender distribution also tends to differ by labels, according to Figure 3. For example, the gender representation is more imbalanced in Atelectasis images (40% female and 60% male). For most classes, there are more male patients than female patients, except for Cardiomegaly and Hernia.

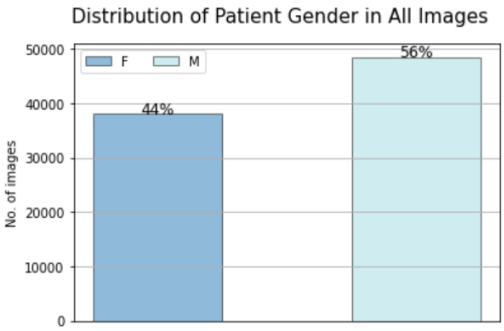

Figure 2: Distribution of Patient Gender

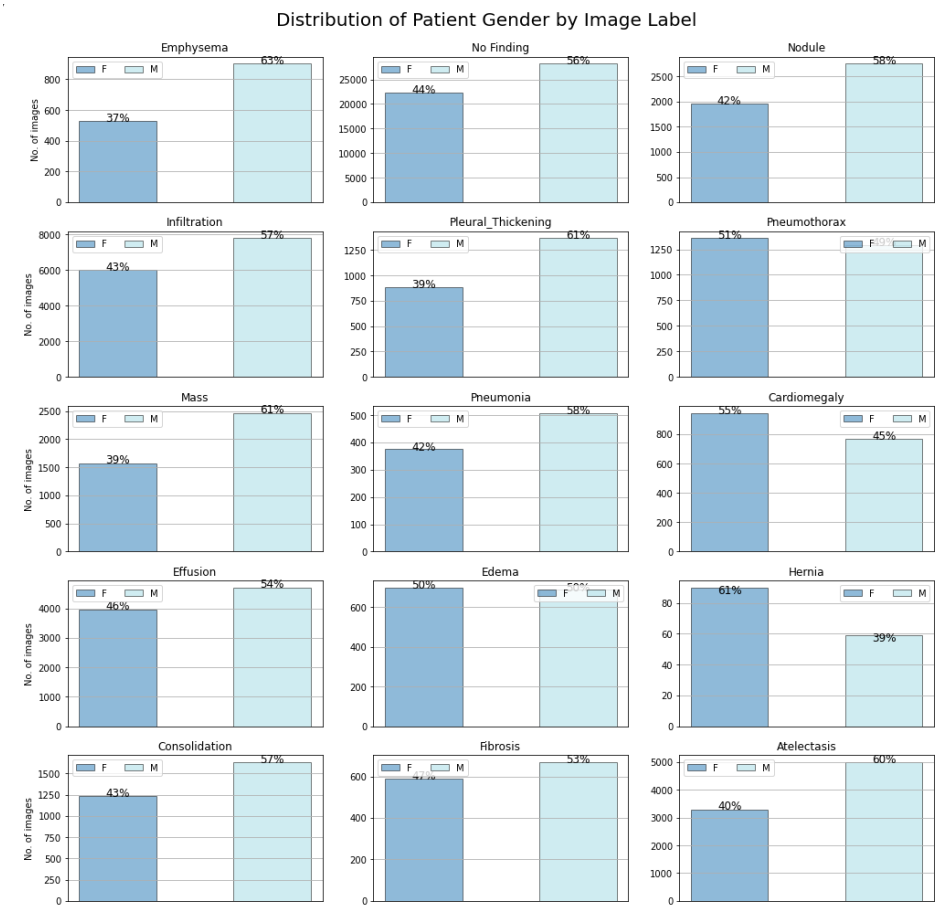

Figure 3: Distribution of Labels

(b) Patient Age

Figure 4 shows the patient age ranges from 0 to 94. The overall distribution resembles a normal distribution that is slightly skewed to the right with the mode value being 60 years old. In addition, age distribution of female and male gender do not differ from each other drastically. In Figure 5 we plot age distribution by label. We show that all classes have roughly the same age distribution with the largest number of images belonging to patients around 60 years of age. Some classes such as hernia and emphysema have notable lower number of patients on the left end of the distribution (lower than 40 year-old).

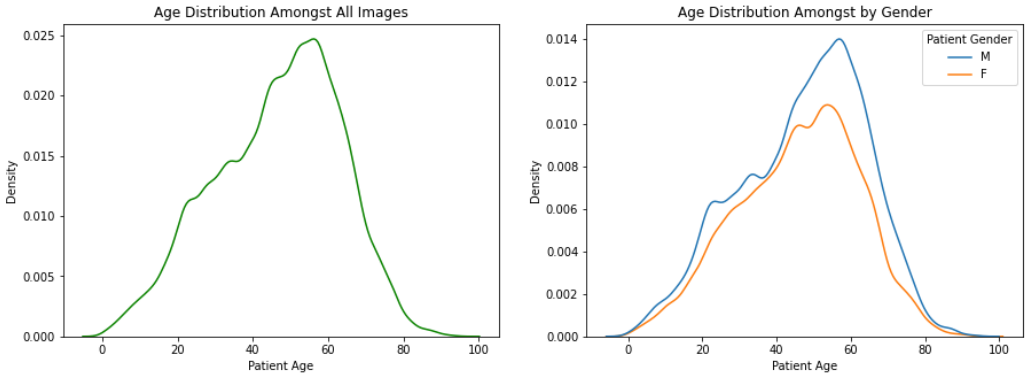

Figure 4: Distribution of Patient Age

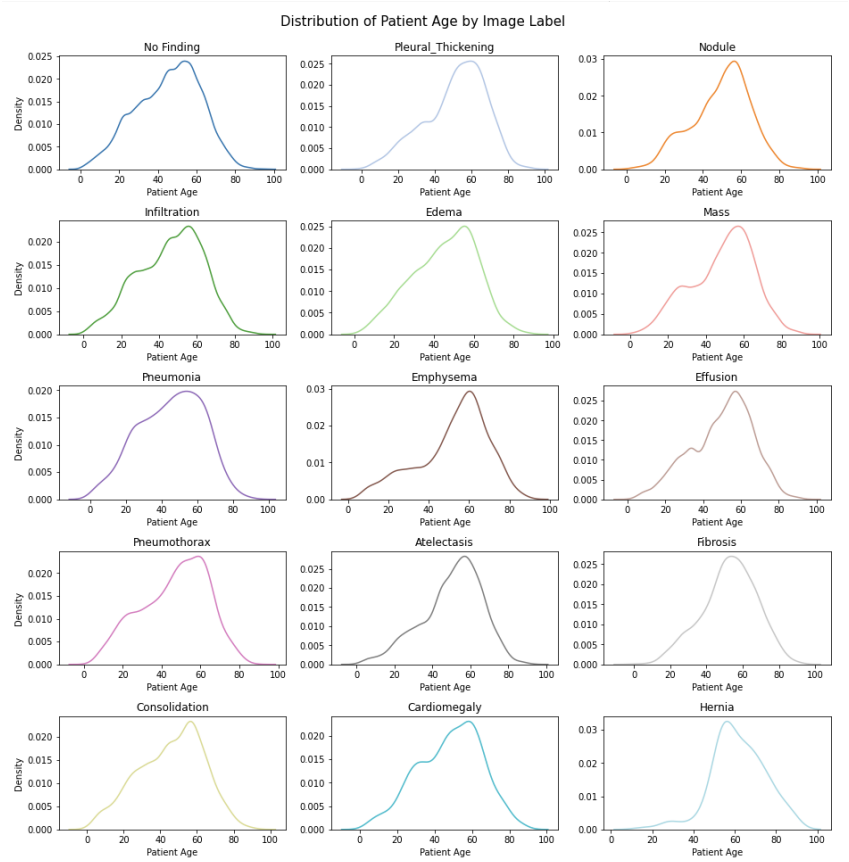

Figure 5: Distribution of Patient Age by Label

3.2.3 Visualization of Images

Lastly, we visualize random samples of the input images to get an idea of the image quality. Figures 6 7 shows there are different levels of contrast and brightness. Some images are blurrier and compose of more white pixels than others. The placement of the chest and the amount of space they take up in the frame are also different.

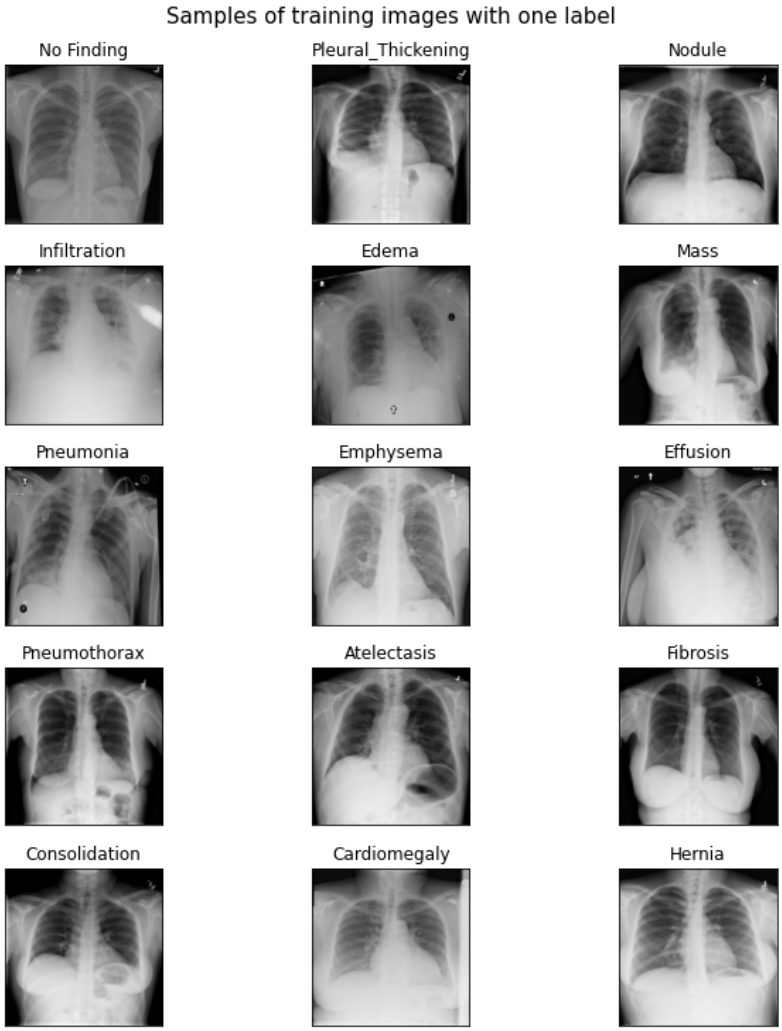

Figure 6: Training Images with One Label

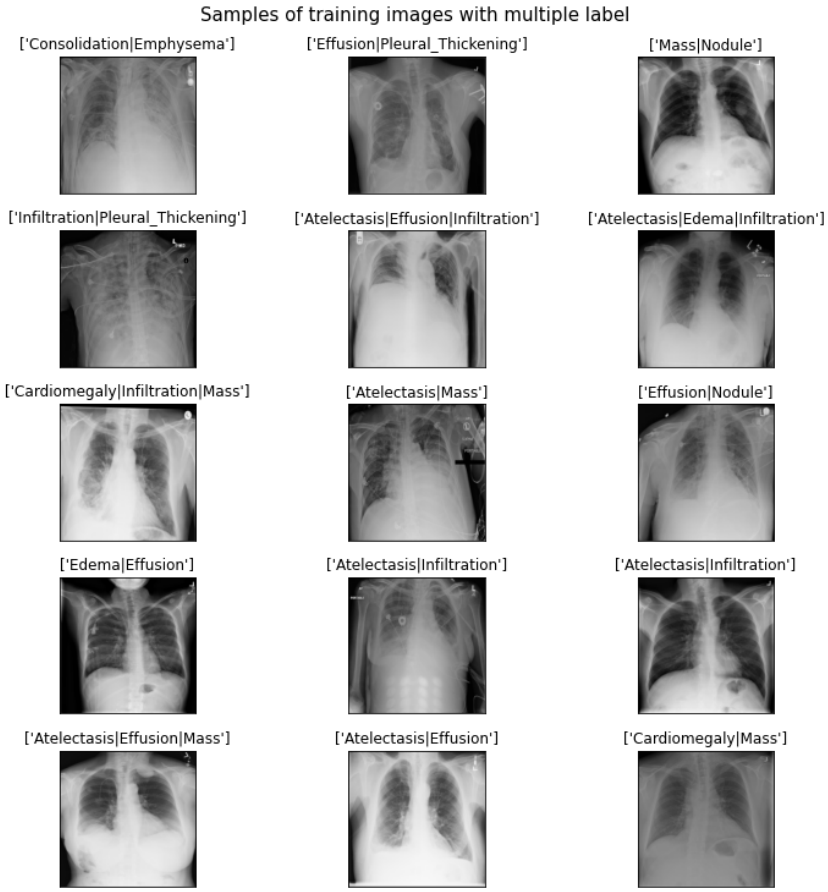

Figure 7: Training Images with Multiple Labels

4 Baseline model results

4.1 Approach

Our baseline model is based on a naive classifier, which predicts using the most frequent class label. Based on our data, the most frequent class label is "No Finding". For this purpose, we are looking at the within class performances of the following 10 classes.

Classes dictionary = 'Atelectasis': 0, 'Cardiomegaly': 1, 'Effusion': 2, 'Infiltration': 3, 'Mass': 4, 'Nodule': 5, 'Pneumonia': 6, 'Pneumothorax': 7 , 'No Finding':9, 'others': 10

4.2 Metrics

As highlighted in our data visualization, our data has a class imbalanced problem. If we were to use accuracy as a performance metric, it would erroneously report that our model is performing relatively well.

A better metric to consider would be the F1 score. F1 score combines both the precision and recall of a classifier. Precision would be more appropriate should we be concerned with false positives, while recall would be more appropriate with false negatives. At this stage, we are concerned with both false positives and negatives and F1 score is deemed more balanced and thus more appropriate. For overall model evaluation, we would rely on macro F1 which is an unweighted average of the F1 scores.

$$Precision = \frac{True\,positives}{True\,positives + False\,positives} \tag{1}$$

$$Recall = \frac{True\,positives}{True\,positives + False\,negatives} \tag{2}$$

$$F1 = 2 \times \frac{Precision \times Recall}{Precision\ +\ Recall} \tag{3}$$

4.3 Results

Figure 8 shows our reported metrics for our naive baseline model. Simply using the most frequent class as the prediction, our accuracy is at 58% despite a naive approach at prediction. We are however focusing on macro average of the F1 score which is showing a poor value of 7%, indicating that our naive classifier is not performing well on the imbalanced data. For the per class performance, the results shows great precision, recall and F1 score for class 9 ("No findings"), while the other classes performs badly which is expected from our approach of predicting class 9 ("No findings") for all data records.

	precision	recall	f1-score	support
0	0.00	0.00	0.00	3414
1	0.00	0.00	0.00	777
2	0.00	0.00	0.00	2788
3	0.00	0.00	0.00	7327
4	0.00	0.00	0.00	1696
5	0.00	0.00	0.00	2248
6	0.00	0.00	0.00	234
7	0.00	0.00	0.00	1241
9	0.58	1.00	0.74	50500
10	0.00	0.00	0.00	16299
accuracy			0.58	86524
macro avg	0.06	0.10	0.07	86524
weighted avg	0.34	0.58	0.43	86524

Figure 8: Naive baseline results

5 Plan for our modelling

Our plan for modeling is split into two sections: defining a weighted loss to address our data imbalance, fine-tuning existing model architectures. Finally, we will consider evaluating our trained model on a small of set images with manual bounding-boxes, to verify how well the model localizes on each disease.

We plan to use one hot encoded vectors of length 15, so that to enable a multi-label classification architecture as in [11]. In this sense, if a patient has 'No Findings', there will be a 1 in the corresponding index of the vector and 0 everywhere else. On the other hand, if a patient has multiple diseases, then there will be multiple 1 in the corresponding indexes of the 15-length vector, and 0 elsewhere.

5.1 Loss re-weighting

From our EDA, we notice that there is general class imbalance in our dataset. A majority of our images contain No Finding labels, whereas individual diseases make up a small fraction of the dataset.

To remedy this imbalance during training, we will employ a weighted cross-entropy loss which penalizes incorrect predictions from positive (disease) images more than negative (no finding) images. To formalize this, we can define the following loss function

$$L_w = w_P \sum_{y=1} -\log f(x) + w_N \sum_{y=0} -\log(1 - f(x))$$

where $w_P = \frac{|P|+|N|}{P}$ is the inverse fraction of positive samples in the batch, and $w_N = \frac{|P|+|N|}{N}$ is the inverse fraction of negative samples in the batch.

5.2 Architectures

We base our modelling on existing architectures AlexNet [4], GoogLeNet [10], VGGNet-16 [9] and ResNet-50 [2], trained on the ImageNet [7] dataset. To fit the model to our current task, we will remove the final fully-connected layers, and only retain the convolutional layers. We will replace them instead with a transitional layer, a global pooling layer, and a fully-connected prediction layer. When training on the ChestX-ray14 dataset, we will train all the layers of the architecture.

The purpose of the transitional and global pooling layers is to implement Grad-CAM [8], which is help visualize the saliency heatmaps for a given image. The transitional layer is to obtain a uniform activation shape for the various models evaluated. This is to obtain a consistent shape between the various models, so the heatmaps to be generated from each model are standardized. The global pooling layer, applied after the transitional layer, summarizes each activation map into a weights, which are combined with the transitional layer to produce the saliency map.

5.3 Boxes IoBB

Once the model is trained, we can generate saliency maps and compare them to the bounding boxes on a small set of of manually labelled images. To compare our heatmaps against bounding boxes, we will predefine a threshold for the pixel intensities, and then fit a bounding box to cover the region of those pixels.

References

[1] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009.

[2] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition, 2015.

[3] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q. Weinberger. Densely connected convolutional networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2261–2269, 2017.

[4] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In F. Pereira, C.J. Burges, L. Bottou, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc., 2012.

[5] Laszlo Papp, Clemens P. Spielvogel, Ivo Rausch, Marcus Hacker, and Thomas Beyer. Personalizing medicine through hybrid imaging and medical big data analysis. *Frontiers in Physics*, 6, 2018.

[6] Pranav Rajpurkar, Jeremy Irvin, Robyn L Ball, Kaylie Zhu, Brandon Yang, Hershel Mehta, Tony Duan, Daisy Ding, Aarti Bagul, Curtis P Langlotz, et al. Deep learning for chest radiograph diagnosis: A retrospective comparison of the chexnext algorithm to practicing radiologists. *PLoS medicine*, 15(11):e1002686, 2018.

[7] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. Imagenet large scale visual recognition challenge, 2014.

[8] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization, 2016.

[9] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition, 2014.

[10] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions, 2014.

[11] Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M. Summers. 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)*, Jul 2017.