Deep-Learning Based Predictive Models for Chest X-ray Diagnosis

Ali Borhan[*]
Stanford University
hborhan@stanford.edu

Abstract

In this study, predictive multi-class models are trained for chest x-ray diagnosis of 14 observations using different deep learning architectures and a large dataset of chest x-ray images called CheXpert. First, three different deep learning models including VGG-16, ResNet-50, and DenseNet-121 are trained on an Amazon AWS EC2 GPU instance. For DenseNet-121, both transfer learning and full training are applied. While a good accuracy is achieved on testset data, the F1 scores on a few observations were low. This was an indication of model robustness issue for a few class predictions. Further analysis of the data indicates an unbalance between available data for those observations with low F1 scores. An up-sampling approach is applied to balance the training data. This results in a significant improvement in both accuracy and F1 scores over the testset data. Finally, a gradient weighted Class Activation Map is applied to localize the highest probability observation for a given x-ray image input.

1 Introduction

Initially, the project was proposed to develop automatic detection models for manufacturing defects during production process. However, the available data was assessed to be limited at the time of the project. Therefore, a similar important problem from Chest radiography is selected for this study to train deep learning models for automated multi-classification of chest x-ray images for 14 observations. Furthermore, a class activation map is applied to localize a selected positive observation. A schematic view of the modeling framework developed in this study is shown in Figure 1. The codes and results of this study are shared in [1].

2 Related Work

In [2], a 121-layer convolutional neural network model is trained on ChestX-ray14 chest image dataset containing observations for 14 diseases. In this reference, the model predicts the probability of the observations along with a heatmap localizing the areas of the image most indicative of pneumonia. Similar to this study, the output of CNN is modified to predict 14 outputs with sigmoid activation. In [3], the authors extend this work to CheXpert, a larger dataset. In this reference, a few CNN architectures are trained and DenseNet121 is shown to achieve the best performance. Both these references provide the background for this study. Incorporating these models and dataset along with a method to deal with unbalanced dataset using up-sampling approach is the contribution of this study.

[*] Hoseinali (Ali) Borhan is a Technical Project Leader in the area of control systems research at Cummins Inc. He is doing his certificate of artificial intelligent at Stanford University (hborhan@gmail.com)

CS230: Deep Learning, Spring 2019, Stanford University, CA.

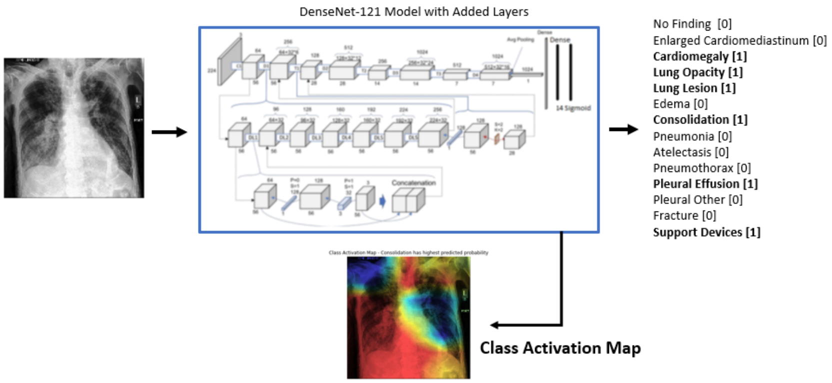

Figure 1: Schematic view of the developed model

3 Dataset and Features

The data is based on CheXpert dataset for chest radiograph interpretation, consisting of 224,316 chest radiographs of 65,240 patients. This data is available in reference [3]. In this study, only frontal view of chest x-ray images are considered. Furthermore, "Age", "Sex" and "AP/PA" features are not considered in training. Since the main goal of this study is a model to be used as a first detection tool for the doctors to do further assessment, the cost of false negative is high and all uncertainties are set to positive in the dataset.

The training data are split into training set, test set and validation set with 10% of data for validation and 10% for test set.

4 Methods

To develop the predictive model, three deep learning model architectures including VGG-16 [5], ResNet-50 [6], and DenseNet-121 [7] are trained. The pre-defined models from Keras [4] is employed as the base model and the last layers are modified to add a global spatial average pooling layer, a fully connected layer with Relu activation, and finally a logistic layer for 14 outputs with sigmoid activation. The loss function is defined to be binary cross entropy. Adam optimizer with $r = 0.0001$ (learning rate), $\beta_1 = 0.9$, and $\beta_2 = 0.999$.

5 Performance Results and Analysis

5.1 DenseNet-121 with transfer learning

First, the DenseNet model with the added layers are trained using transfer learning method. For this purpose, all layers of the base model were set to be frozen with the weights trained on ImageNet dataset. With this approach, $1,063,950$ of the parameters out of total of $8,101,454$ parameters are trainable. The training of the model results in accuracy = 0.80 on test data. Furthermore, the model results on false positive and negative is summarized in table 1. It is observed that, while the prediction accuracy is relatively acceptable, the F1-score results are overall low.

5.2 DenseNet-121 Full Training

Since transfer learning approach with lower number of tuneable parameters led to low F1 score results on testset data, this might be an indication of over-fitting. Therefore in this section, all parameters of the model (about 8 million parameters) are set to be tuneable without freezing any layer. This model trained over the dataset and in 0.85 test set accuracy. Furthermore, the F1 score results are shown in Table 2. The results indicate improved performance on testset data. This indicates that the overfitting was part of the low performance of the model. However as seen in Table 2, the F1 scores are still low on a few observations including Enlarged Cardiomediastinum, Lung Lesion, Pneumonia, Pleural

Table 1: F1-score on testset with DenseNet model trained with transfer learning

```
                            precision    recall  f1-score   support

                No Finding       0.45      0.19      0.27      1679
Enlarged Cardiomediastinum       0.00      0.00      0.00      1880
              Cardiomegaly       0.26      0.24      0.25      2991
              Lung Opacity       0.65      0.37      0.47      9908
               Lung Lesion       0.00      0.00      0.00       810
                     Edema       0.59      0.14      0.22      6203
             Consolidation       0.00      0.00      0.00      3824
                 Pneumonia       0.00      0.00      0.00      2049
               Atelectasis       0.00      0.00      0.00      6047
              Pneumothorax       0.00      0.00      0.00      2084
          Pleural Effusion       0.64      0.46      0.53      8740
             Pleural Other       0.00      0.00      0.00       431
                  Fracture       0.00      0.00      0.00       787
           Support Devices       0.65      0.85      0.74     10848

                 micro avg       0.60      0.32      0.42     58281
                 macro avg       0.23      0.16      0.18     58281
              weighted avg       0.42      0.32      0.34     58281
               samples avg       0.49      0.30      0.34     58281
```

Table 2: F1-score on testset with DenseNet model - full learning

```
                            precision    recall  f1-score   support

                No Finding       0.77      0.06      0.11      1683
Enlarged Cardiomediastinum       0.75      0.00      0.00      1928
              Cardiomegaly       0.66      0.49      0.56      3089
              Lung Opacity       0.63      0.87      0.73      9920
               Lung Lesion       0.67      0.04      0.08       799
                     Edema       0.64      0.74      0.69      6200
             Consolidation       0.52      0.03      0.06      3727
                 Pneumonia       0.56      0.08      0.14      2112
               Atelectasis       0.58      0.32      0.41      5879
              Pneumothorax       0.77      0.24      0.36      2053
          Pleural Effusion       0.69      0.91      0.79      8679
             Pleural Other       0.50      0.01      0.01       433
                  Fracture       0.57      0.02      0.03       755
           Support Devices       0.89      0.73      0.80     10844

                 micro avg       0.69      0.57      0.63     58101
                 macro avg       0.66      0.32      0.34     58101
              weighted avg       0.69      0.57      0.56     58101
               samples avg       0.63      0.54      0.55     58101
```

Other, and Fracture. Before we do an error analysis on the results, 2 more model architectures are trained on the dataset in the next section.

5.3 Other model architecture: VGG-16 and ResNet-50

In this section, the VGG-16 and ResNet-50 architectures from Keras are trained. The detailed results are available in [1]. The summary of the average F1 scores are listed in Table 3. It is concluded that the model structure did not improve the F1 scores further.

Table 3: Model Structure Impact (DenseNet-121 is selected for error analysis and further improvement)

	DenseNet-121	VGG-16	ResNet-50
micro avg	0.63	0.60	0.61
macro avg	0.34	0.30	0.33
weighted avg	0.56	0.52	0.55

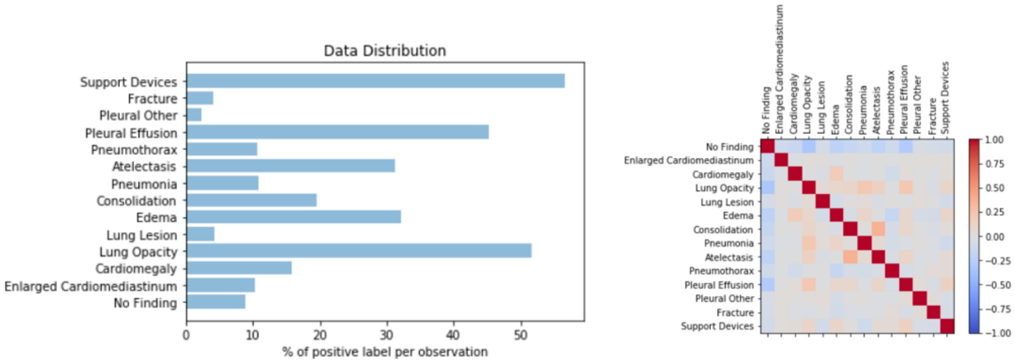

Figure 2: Data distribution

6 Balancing CheXpert train data to improve the selected DenseNet-121 robustness performance

Further exploration on the training data is done in this section to identify methods to improve the low F1 scores of the few observations. The distribution of train data and the correlation matrix are shown in Figure 2. The results indicate that the size of training data for the observations with low F1 scores is much lower comparing to the ones with high F1 score. This indicates the issue of unbalanced data for this multi-class classification problem. To fix this issue, an up-sampling approach is applied on the class observations with low size. The distribution of the re-sampled data are shown in Figure 3 showing more balanced train data. This approach has similar effect as having a weighted loss function with higher weight for the observations with low training size. The DenseNet-121 model was re-trained with the balanced dataset and the results on the same testset data are shown in Table 4. The accuracy of the model on testset data is 0.88. The results indicate significant improvement in the performance of the trained model particularly on F1 scores for all observations.

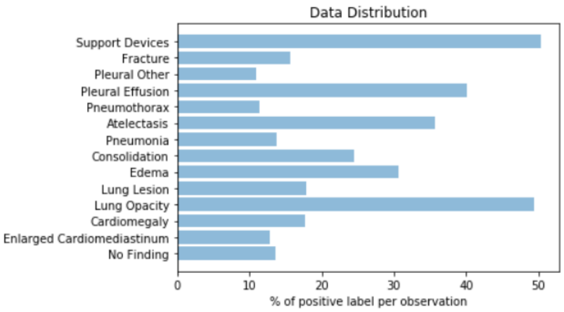

Figure 3: Data distribution after balancing through up-sampling approach

Table 4: The results of the DenseNet-121 model trained on the balanced data (significant improvement is achieved with balancing data)

```
                            precision    recall  f1-score   support

                No Finding       0.72      0.84      0.77      2291
Enlarged Cardiomediastinum       0.63      0.32      0.42      2100
              Cardiomegaly       0.58      0.70      0.64      2662
              Lung Opacity       0.74      0.71      0.73      9283
               Lung Lesion       0.71      0.95      0.81       861
                     Edema       0.69      0.73      0.71      5619
             Consolidation       0.43      0.76      0.55      3643
                 Pneumonia       0.64      0.45      0.52      2042
               Atelectasis       0.66      0.65      0.66      6581
              Pneumothorax       0.64      0.51      0.57      1439
          Pleural Effusion       0.77      0.76      0.77      6775
             Pleural Other       0.59      0.96      0.73       314
                  Fracture       0.52      0.99      0.68       880
           Support Devices       0.85      0.83      0.84      9736

                 micro avg       0.69      0.72      0.70     54226
                 macro avg       0.66      0.72      0.67     54226
              weighted avg       0.70      0.72      0.70     54226
               samples avg       0.66      0.69      0.65     54226
```

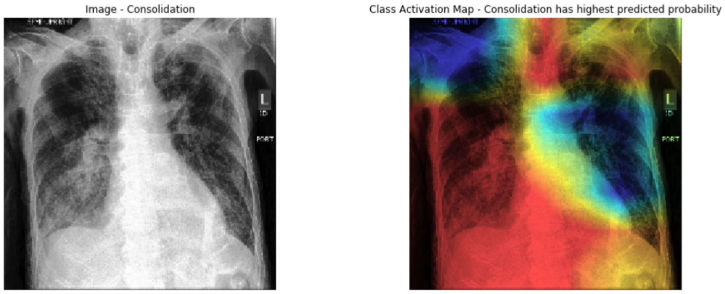
Image - Consolidation
Class Activation Map - Consolidation has highest predicted probability
L
PORT
L
PORT

Figure 4: A sample of Class Activation Map application

7 Weighted Gradient Class Activation Map

Finally, the weighted gradient class activation map is applied on the last convolution layer of the DenseNet model trained on the balanced data [8]. The results are shown in Figure 4 for the predicted observation with highest probability.

8 Conclusion and Future Work

A predictive model for chest x-ray observation over large amount of data is developed in this study. The DenseNet-121 model is shown the best performance after balancing training data for all class observations. This indicates the importance of both model architecture design and data processing for successful application of deep learning techniques. For the next steps, further hyper-parameter tuning is proposed to be done particularly on positive class threshold after sigmiod activation. Furthermore, the CheXpert data is proposed to be combined with other x-ray image data including MIMIC-CXR dataset. Finally, an image segmentation can be applied by developing corresponding dataset based on methods such as Class Activation Map applied in this study.

9 Contributions

This work is done solely by the author. The author would like to thank Aarti Bagul from Stanford University for her mentoring during the project. The author also would like to thanks Prof. Andrew Ng for cs230 course and great teaching of deep learning methods and applications.

References

[1] **Project github Repository**, *https://github.com/hborhan/DL-CNN-CheXpert-data*

[2] Rajpurkar, Pranav, et al. "Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning." arXiv preprint arXiv:1711.05225 (2017).

[3] Irvin, Jeremy, et al. "Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison." *Thirty-Third AAAI Conference on Artificial Intelligence.* 2019.

[4] Keras applications website, *https://keras.io/applications/*

[5] Simonyan, Karen, and Andrew Zisserman. "Very deep convolutional networks for large-scale image recognition." arXiv preprint arXiv:1409.1556 (2014).

[6] He, Kaiming, et al. "Deep residual learning for image recognition." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016.

[7] Huang, Gao, et al. "Densely connected convolutional networks." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017.

[8] Selvaraju, Ramprasaath R., et al. "Grad-cam: Visual explanations from deep networks via gradient-based localization." *Proceedings of the IEEE International Conference on Computer Vision.* 2017.