Pneumonia Detection using deep learning techniques

Submitted by

Name of Student

KANNU PRIYA
MUHAMMED ALI JOWHER
SPANDANA KUMBHA
VIMOCHANA KUMMUR

University at Buffalo
Buffalo, NY 14260

Abstract

Pneumonia is an infection in one or both lungs. Bacteria, viruses, and fungi cause it. The infection causes inflammation in the air sacs in your lungs, which are called alveoli.

In the proposed work an automated Pneumonia detection screening system is used to detect the severity of the disease. Medical images of patients lungs of the format DICOM(dcm) is used for various pre-processing techniques, data visualisation and EDA. Model is built using deep learning model. Model is tested and fine-tuned with different hyper parameters, by trying different optimizers, loss functions, epochs, learning rate, batch size, checkpointing, early stopping etc..

Contents

Chapter 1

Introduction

Pneumonia is an infection in one or both lungs. Bacteria, viruses, and fungi cause it. The infection causes inflammation in the air sacs in your lungs, which are called alveoli.Pneumonia accounts for over 15% of all deaths of children under 5 years old internationally.It requires review of a chest radiograph (CXR) by highly trained specialists and confirmation through clinical history, vital signs and laboratory exams. Pneumonia usually manifests as an area or areas of increased opacity on CXR. However, the diagnosis of pneumonia on CXR is complicated because of a number of other conditions in the lungs such as fluid overload (pulmonary edema), bleeding, volume loss (or collapse), lung cancer, or post-radiation or surgical changes. Outside of the lungs, fluid in the pleural space (pleural effusion) also appears as increased opacity on CXR.CXRs are the most commonly performed diagnostic imaging study. A number of factors such as positioning of the patient and depth of inspiration can alter the appearance of the CXR, complicating interpretation further. In addition, clinicians are faced with reading high volumes of images every shift. Now to detection Pneumonia we need to detect Inflammation of the lungs. Medical images are used to train and build model for Pneumonia detection.

In this project, we are building an algorithm to detect a visual signal for pneumonia in medical images. Specifically, the algorithm will automatically locate lung opacities on chest radiographs. There are various techniques and methodology used in image processing and machine learning to identify abnormal digital image.

Chapter 2

Problem Definition

To design a pneumonia detection system using appropriate pre-processing techniques and deep learning to locate the position of inflammation in an image.

Chapter 3

Work Done and Implementation

3.1 Design and Implementation

3.1.1 Dataset

The proposed automated diagnosis system of Pneumonia is evaluated by using a set of 30000 training images(26684 after droping duplicates) and 3000 test images from the downloaded dataset having high-resolution and taken under a variety of imaging conditions[1]. Images in the training data is provided as a set of patientIds and bounding boxes. Bounding boxes are defined as follows: x-min y-min width height, There is also a binary target column, Target, indicating pneumonia or non-pneumonia.There may be multiple rows per patientId.We configured our train class info csv file as classes and our training image set info as train .Train has patient id's of patients along with bounding box coordinates of the opacity in the xray. target column is a binary and tells us if that particular patient had pneumonia or not. Classes also has patient id's along with the class that it belongs to The classes in dataset are having either of the three labels below:
1- No Lung Opacity / Not Normal
2- Normal
3- Lung Opacity

Medical images are stored in a special format called DICOM files (*.dcm). They contain a combination of header metadata as well as underlying raw image arrays for pixel data. These images are provided as input to the system. Data pre-processing and analysis is done on the input data images to explore the given Data files, Deal with missing values, visualizing different classes. The aim of performing pre-processing and analysis on images is to obtain an improvement of the image data and to understand more about the input data that will be used to train our model. The pre-processing techniques and Data analysis performed in the proposed system are:
1)Handling missing values
2)Data analysis to understand overall distribution of classes
3)Visualise the images of different classes

3.1.2 Image Pre-processing and Data Analysis

The following pre-processing techniques was performed on the images using Python.

- Handle missing values
- Data analysis to understand overall distribution of classes
- Visualise the images with basic dicom metadata

Handling missing values

Missing data present various problems. First, the absence of data reduces statistical power, which refers to the probability that the test will reject the null hypothesis when it is false. Second, the lost data can cause bias in the estimation of parameters. Third, it can reduce the representativeness of the samples. We have wrote a function which checks the missing values and the % number compared to full data. Checking missing values in our train dataframe, there are an ample amount in co ordinates since xray images with no Pneumonia needs no bounding boxes,hence H,W,X,W would be nan for class 0 We have not changed it and let it be, because these will be handy further, since this can help easily to separate classes by filtering with np.nan. The classes dataframe is free of null values.

Data analysis to understand overall distribution of classes

A function which gives us the % of data in each class against the total was created, using this we got our target class distribution as below image (figure 3.1).No Lung Opacity or Not Normal is approx. 39% which is negative class along with another negative class Normal which is approx. 30%.Our positive class Lung Opacity is approx.31%. We next merged the dataframes into a single train dataframe for better usage and checked overall distribution of our classes positive and negative (figure 3.2). Positive is almost 69% and negative class is 31%. From these analysis We were able to get an idea that class imbalance is not a case here. We then checked the distribution of our x,y,h,w (figure 3.3), Y, H, W seem to be normally distributed but X seems worrysome with bimodal distribution which may cause problems further.

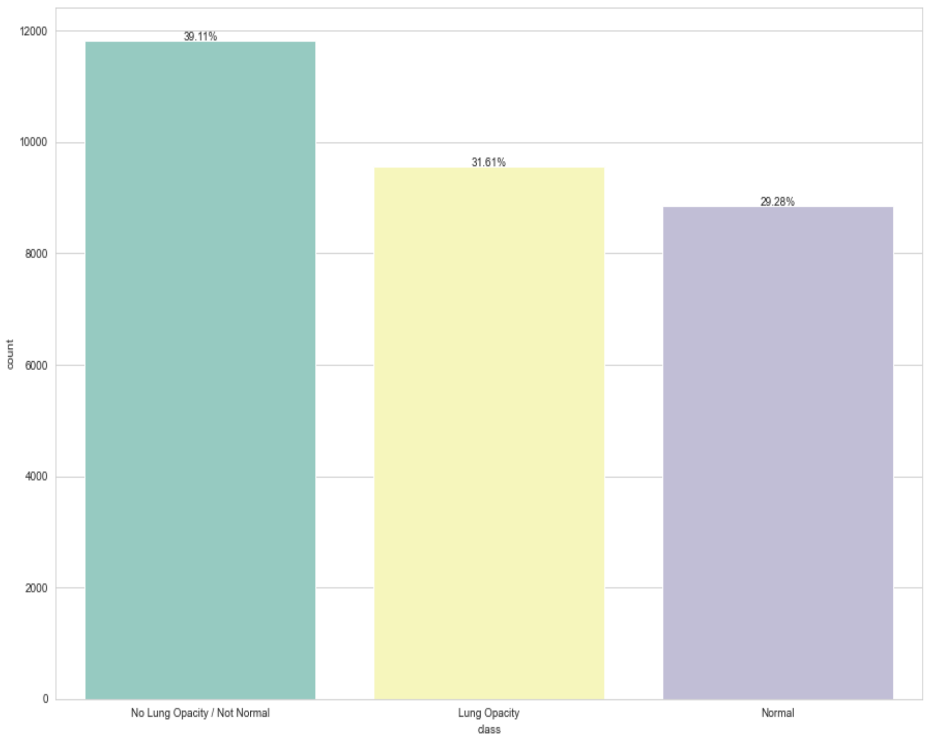

Figure 3.1: target class distribution

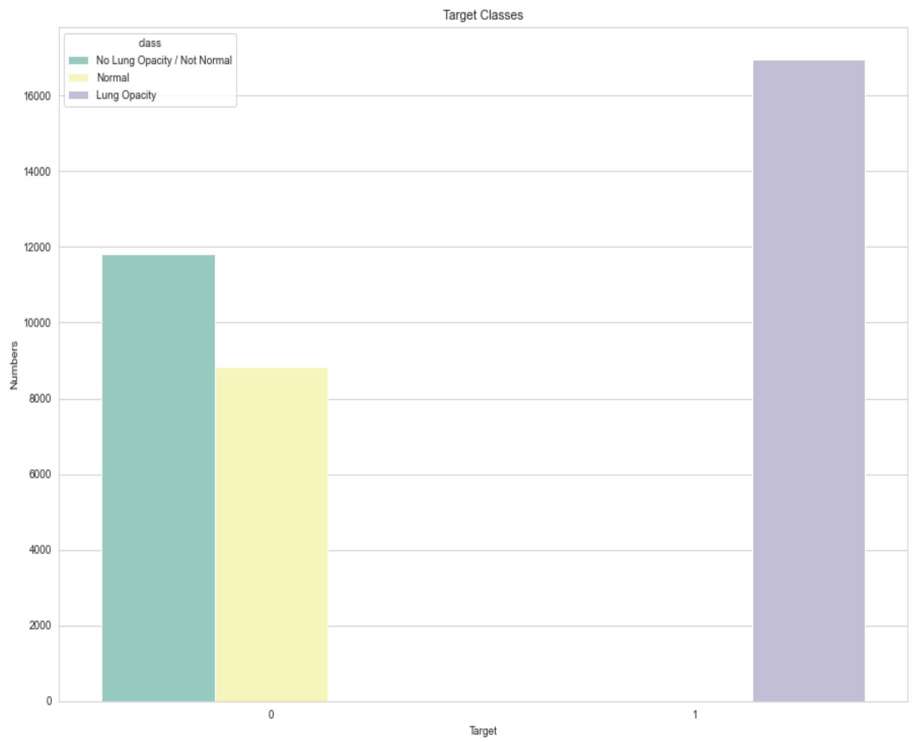

Figure 3.2: overall distribution

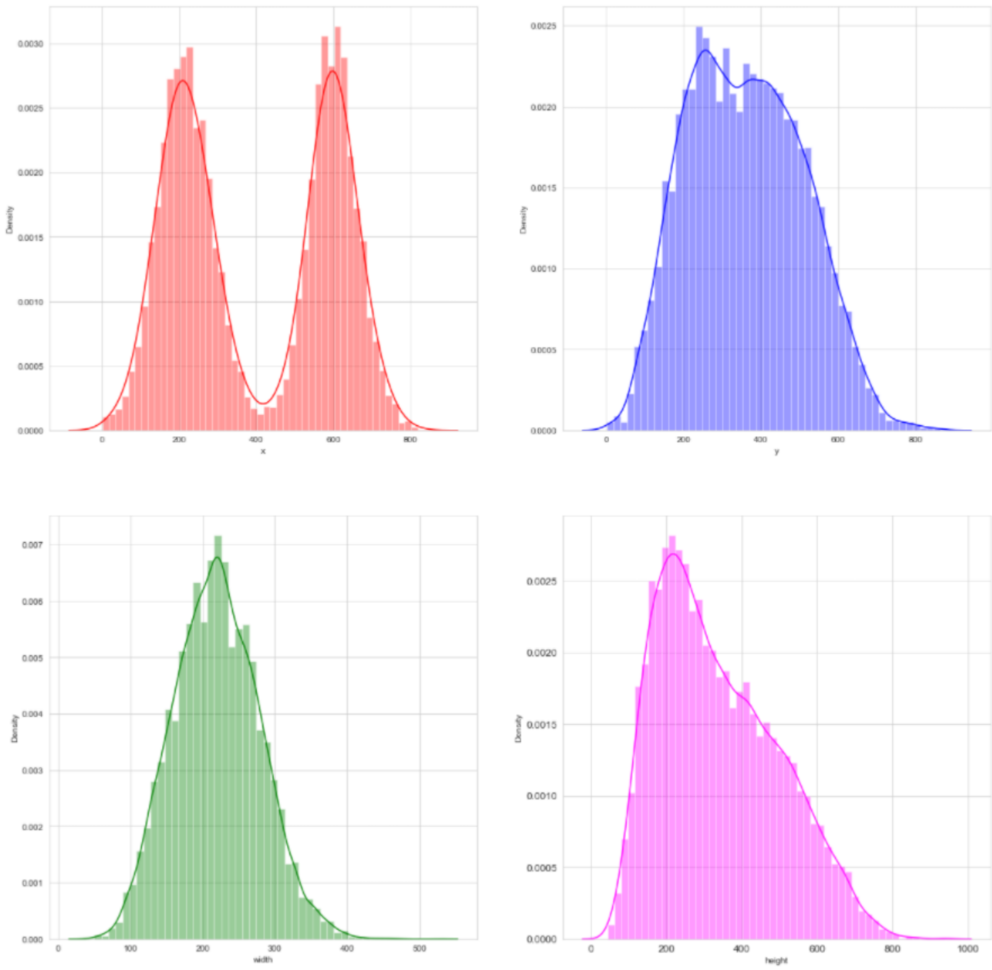

Figure 3.3: x,y,h,w distribution

Visualise the images with basic dicom metadata

Here we consider 2000 samples of pneumonia data and jot the center of the binding boxes. The red patches are the entire boxes and black points are the center of those boxes(figure 3.4)

With this we get an very important idea that no specific part of the lung may be prone to opacity. Almost we have every part of lung in 2000 samples having opacity, We then checked the train and test set, train set excluding repeated images is 26684 and test has 3000 samples. This data was good to train a Deep learning model from scratch with decent accuracy and more than enough to feed pre trained model. Our images are numpy arrays with dimension 1024x1024, These dimensions are perfect for training model as they are AI friendly, not extremely high dimension nor extremely pixalated. This saves us from image transformation or any sort of data compression .

We then visualised the images along with basic dicom metadata to see positive class as well as negative class and also with binding boxes and were able to understand from the images is the opacity is not visually identifyable.

Figure 3.4:

3.1.3 Modelling using Residual Networks (ResNet)

After the first CNN-based architecture (AlexNet) that win the ImageNet 2012 competition, Every subsequent winning architecture uses more layers in a deep neural network to reduce the error rate. This works for less number of layers, but when we increase the number of layers, there is a common problem in deep learning associated with that called Vanishing/Exploding gradient. In order to solve the problem of the vanishing/exploding gradient, this architecture introduced the concept called Residual Network. In this network we use a technique called skip connections . The skip connection skips training from a few layers and connects directly to the output. The approach behind this network is instead of layers learn the underlying mapping, we allow network fit the residual mapping. The advantage of adding this type of skip connection is because if any layer hurt the performance of architecture then it will be skipped by regularization. So, this results in training very deep neural network without the problems caused by vanishing/exploding gradient.

Model Implementation

We extracted all the positive samples into a dictionary and split them as train and validation. We created generator class which has load data, get data and predict data functions, and also created functions to down sample the data(Since positive class is very less than negative class), create ResNet layers and complete the network. After that jaccord loss metric was calculated. Then a model with 28 Convolution layers, 21 batch Normalisation layer, with clubbed with 21 leaky relu and one upsampling, output layer and one input layer contributing overall to 72 layer. Finally the model is trained and tested.

Next we try our hands on transfer learning by using pre-trained models.
1- VGG 19
2- Inception v3
3- MobileNet

VGG 19

VGG-19 is a trained Convolutional Neural Network, from Visual Geometry Group, Department of Engineering Science, University of Oxford. The number 19 stands for the number of layers with trainable weights. 16 Convolutional layers and 3 Fully Connected layers. We tested this model with our test data samples which gave us a test accuracy of 70%.

MobileNet

As a lightweight deep neural network, MobileNet has fewer parameters and higher classification accuracy. In order to further reduce the number of network parameters and improve the classification accuracy, dense blocks that are proposed in DenseNets are introduced into MobileNet. In Dense-MobileNet models, convolution layers with the same size of input feature maps in MobileNet models are taken as dense blocks, and dense connections are carried out within the dense blocks. The new network structure can make full use of the output feature maps generated by the previous convolution layers in dense blocks, so as to generate a large number of feature maps with fewer convolution cores and repeatedly use the features. By setting a small growth rate, the network further reduces the parameters and the computation cost. Two Dense-MobileNet models, Dense1-MobileNet and Dense2-MobileNet, are designed. Experiments show that Dense2-MobileNet can achieve higher recognition accuracy than MobileNet, while only with fewer parameters and computation cost. MobileNet gave us an accuracy of 73%.

Chapter 4

Results and Analysis

We have trained three popular models and these are the analysis and result

Resnet Trained from scratch	81%
VGG19, pretrained	70%
Mobilenet	73%

From the above table it is clear that the proposed cnn resnet model trained from scratch we developed as part of this project gives us a better accuracy.

Considering the model is built from scratch and we were able to achieve a decent accuracy. Below table shows result of the Model we developed :

Training Accuracy	97.16%
Validation Accuracy	83.91%
Testing accuracy	81.24%

Below are few test images by our prediction algorithm in action, the Red box is ground truth, blue box is prediction(figure 4.1)

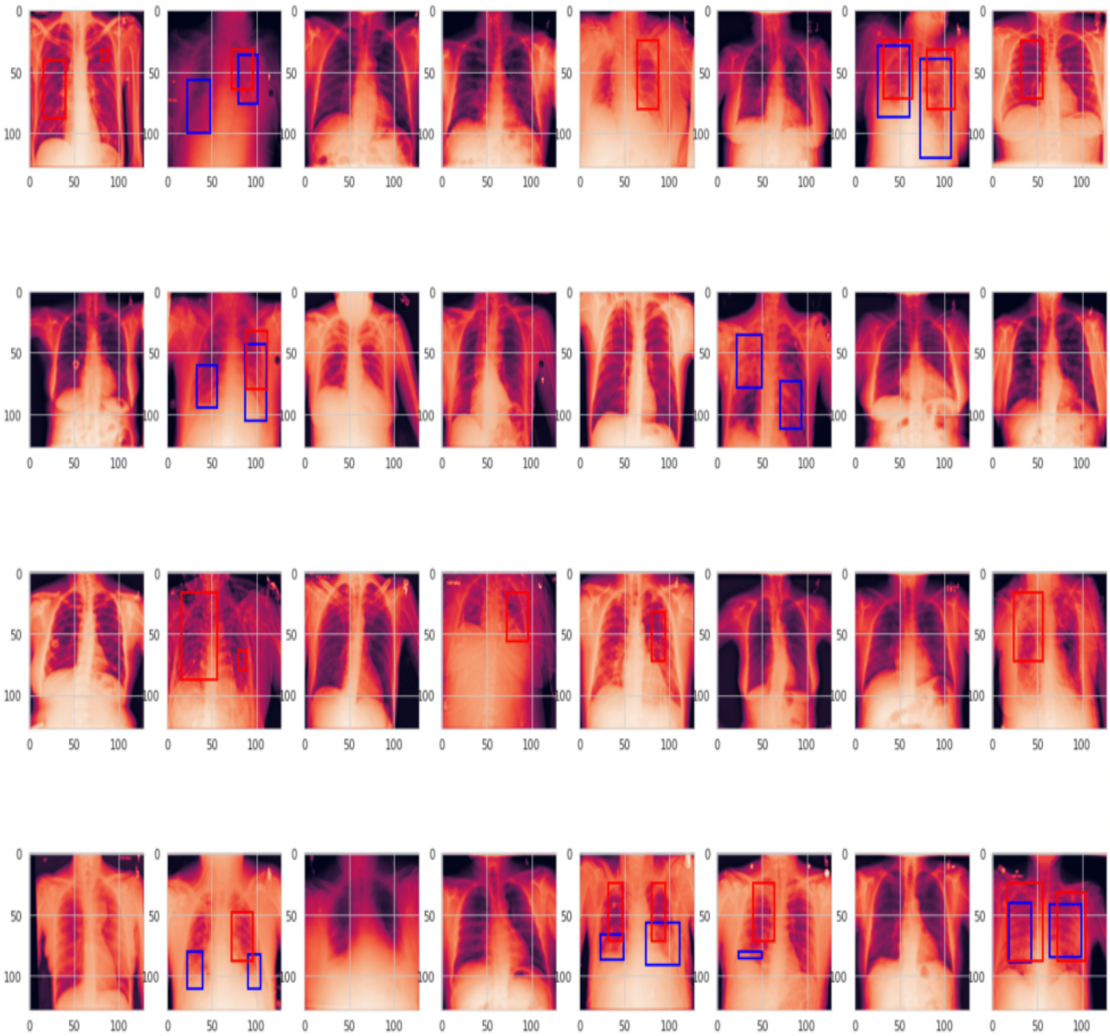

Figure 4.1: test images by our prediction algorithm

Chapter 5

Conclusion

An automated pneumonia screening system was developed to determines the presence of No Lung Opacity / Not Normal, Normal, Lung Opacity in a patient and is presented.In this work we have used image processing and data analysis techniques on DICOM images taken by the use of medical image cameras under different conditions. A computer based screening system was used for grading the severity of pneumonia in a patient. This screening system helps in determining the presence of pneumonia in its early stage. The Proposed system demonstrated a classification with 81 percentage accuracy.

We trained three pre trained models and thus arrived at the conclusion we will choose our cnn resnet model developed from scratch above the other models for screening the disease. The reason are listed below :

- A lot of hyper parameter tuning can be made if necessary.
- Gives us good accuracy since our train data set is properly fit on it.
- Pre trained models can be used if computation is a barrier.
- Mobilenet has promising results if faster running is desired(Since it is light)
- Considering the model is built from scratch and we were able to achieve a decent accuracy and because of above reasons we decided to choose this as a better model for prediction.

References

[1] Pneumonia Detection dataset. <https://www.kaggle.com/c/rsna-pneumonia-detection-challenge/data>

[2] TJ Bowerman, J Zhang, LM Waite *"Antibacterial treatment of aspiration pneumonia in older people: a systematic review.The aims of this paper were to systematically review the literature on the antibacterial treatment of aspiration pneumonia in elderly patients and identify the microbiology of aspiration pneumonia.*, Vol.3: Issue 1, January 2015.

[3] Shickel, B.; Tighe, P.J.; Bihorac, A.; Rashidi, P. Deep EHR. *A survey of recent advances in deep learning techniques for electronic health record (EHR) analysis. IEEE J. Biomed. Health Inform.*, *Journal of Intelligent Learning Systems and Applications*, August 2013.

[4] Litjens, G.; Kooi, T.; Bejnordi, B.E.; Setio, A.A.A.; Ciompi, F.; Ghafoorian, M.; van der Laak, J.A.W.M.; Ginneken, B.; Sánchez, C.I.*A survey on deep learning in medical image analysis.*, *Med. Image Anal. 2017.*,

[5] Scott, J.A.; Brooks, W.A.; Peiris, J.S.; Holtzman, D.; Mulholland, E.K *"Pneumonia research to reduce childhood mortality in the developing world. J. Clin. Investig. 2008.*, Vol. 11:No. 1, October 2014