

Capstone Project



March 30, 2020

By

Avinash Pulichinthala

Contents

1. Summary of Problem Statement
2. Overview of the final process
3. Step by step walk through of solution
4. Model evaluation
5. Comparison to benchmark
6. Visualizations
7. Implications
8. Limitations
9. Closing reflections

**Summary of Problem Statement:**

The main goal of the project is to build a pneumonia detection system to detect presence or absence of a disease, if pneumonia is present locate the position of inflammation in an image by placing the boundary box using CV algorithms

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. In 2017, 920,000 children under the age of 5 died from the disease, Tissues with sparse material, such as lungs which are full of air, do not absorb the X-rays and appear black in the image. Dense tissues such as bones absorb X-rays and appear white in the image.

Automating Pneumonia screening in chest radiographs will be business value add, providing affected area details through bounding box. This assists physicians to make better clinical decisions or even replace human judgement in certain functional areas of healthcare (e.g.: radiology).

To solve this problem there are many way we can automate the process of detecting the pneumonia like traditional machine learning like KNN , SVM , these traditional methods work really well on data that are not images and that is mainly because of the Computational time it needs to process thousands of images very large images

One other rapidly growing field is Deep learning which is heavily implemented on Image related Use case and this will be ideal to solve this use case , Deep learning has been around for quite a sometime but it has become really popular over the last few years because of the latest advancements in Technology like the vast amount data and great computational powers we have available

To solve this Use case we particularly will be using Convolution Neural Networks which is basically analyzing the influence of nearby pixels by using something called filter , after the filter is passed over the image , a feature map is generated for each filter and these are then taken through an activation function which decides whether a certain feature is present in given location in image , we can do lot of things such as adding more filtering layers and creating more feature maps , we can also use pooling layers in order to select the largest values on the feature maps and use these as inputs to subsequent layers .

**Overview of Final Process:**

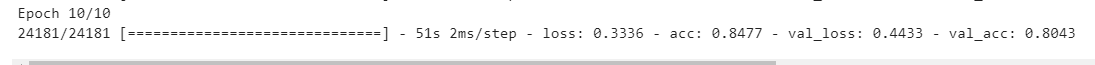
The training data that has been provided has set of patient Ids and bounding boxes. Bounding boxes are defined as follows: x-min y-min width height. There is also a binary target column, Target, indicating pneumonia or non-pneumonia. There may be multiple rows per patient Id.

All provided images are in DICOM format, medical images are stored in the format of DICOM, these images contain different number of features about a patient from their gender to age to image pixel values.

There are total of 30,227 patients in the data we have received and out of that the unique patients are 26,684. We have a total of 3 classes out of that one is lung opacity and one is no lung opacity but there's a third class where the image not normal but there's no opacity as well, this is the class that has more patients when compared with other two

🡪 {'No Lung Opacity / Not Normal': 11821, 'Normal': 8851, 'Lung Opacity': 9555}

We have created three different types of models to see how they have performed for different parameters, initial model is a basic CNN without any hyperparameters, the model was able to achieve 84 % training accuracy and 80 % validation accuracy after training for 10 epochs

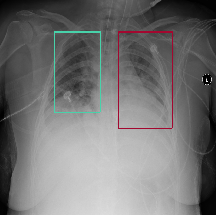
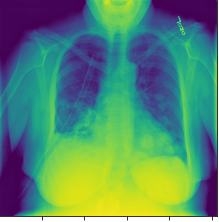


Similarly, we have created two other models one with few hyper parameters and another with completely different Architecture and finally we have created two output files with predicted class on the test images, one with just the class information and another with Bounding box information

**Step by Step Walk Through:**

By going over the data set that has been provided, images are of size 1024 by 1024 of grey scale, The very first thing we have done as part of EDA is to see how an image looks for a patient with Pneumonia and we have created a function to draw the bounding box over the image as shown

**Image with Pneumonia Image without Pneumonia**

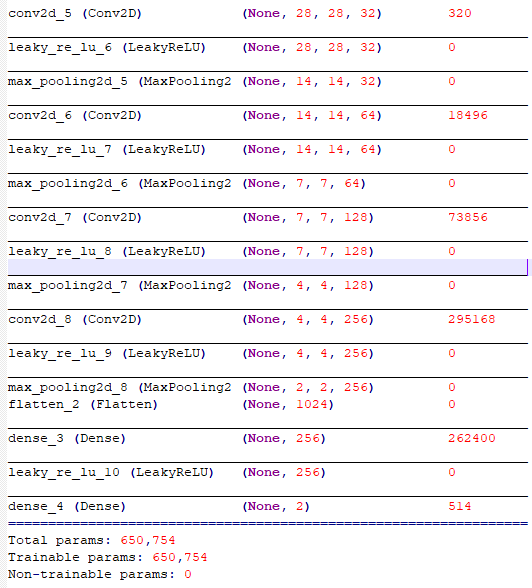
 

Another interesting thing I have checked for is there any useful information that can be extracted from the DICOM data that can help with EDA which consisted of lot of different features , out of all the features that are present Patient Age, Patient Gender, View Position, Body Part Examined and Modality are the feature caught the interest , to study the data set in more details we have extracted the details these details and combined with the bounding box data set and performed few simple visualizations to see how the data is spread across , for example population of male patients is higher and they are also the ones with more lung opacity as well

As part of the preparing the data for model input we have come across number of hurdles , one of the main thing being the size of the images and the time it took to read each image and create a array with pixel values of each image , as we have tried with 1024 to keep the original size as is but we couldn’t load more than 2000 images because of computational limitations hence we have reduced the pixel size to 28 by 28 for out initial models , we have first converted he DICOM images to JPEG and loading the each image with 28 \* 28 size and divided each of the image by 255 to normalize the data and stored all the images as an Array

Once we have prepared the training set data in, we have converted the Target variable using the categorical encoding Next, we have split the data set into Training and validation by using sci kit learn Train test Split method by allocating the 20 % of the data to validation

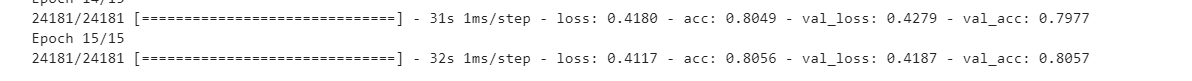
As part of the first model we have built a very simple CNN model without any Dropout , it consisted of Conv2D and Max pooling layers, first we have create a model which is Sequential as this is the easiest way to build model in keras, it allows to build model layer by layer, the first layer is a Conv2D layer of 32 filters of 2 by 2 followed by Activation function of Leaky Relu and Max pooling layer of size 2 by 2 and padding of same , we have added 3 more layers similarly and flattened the layer and applied Softmax as final activation function to determine the probability of image belonging to lung opacity or not , below is the model summary



Next we have compiled the model with three parameter as optimizer, loss, and Metric Optimizer controls the learning rate and we have used Adam as out optimizer, this is a important parameter as the learning rate determines how fast the optimal weights of the model are calculated And we have used categorical cross entropy as out loss function , the low score for this indicates that model is performing better and accuracy is the metric we have used to interpret the model performance as we don’t have the class label given for test images

As mentioned above we have trained the model for 10 epochs with a batch size of 50 and we have achieved the training accuracy of 84 % and validation accuracy of 80 %

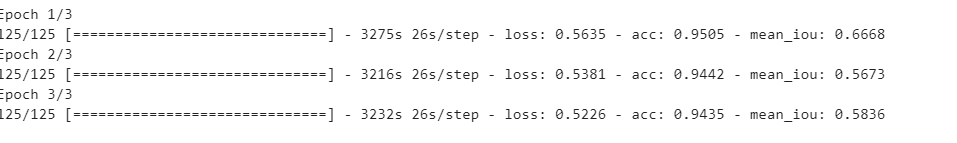
For the second model we have implemented the similar architecture but we have also added the drop out layer to the network and trained the model for 15 epochs with a batch size of 64 and we were able to achieve training accuracy of 80 % and validation accuracy of 80 % and also used this same model to predict the class of the Test images , I have attached the predicted test labels for the test data set , as we don’t have the actual value we couldn’t compare it with predicted value



For the both the first two models we were only predicting the class label of the test images and we wanted to build a complete different model with different method where we can also predict the bounding box of the predicted output as well , for this model instead of reducing the image size to 28 by 28 I have used the original image size but the number of images I have used for training are also less because of the computational resource limitations

First a CNN is to segment the image using the bounding box details directly as mask, secondly connected components is used to separate the multiple areas of predicted pneumonia, finally a bounding box is simply drawn around every connected component

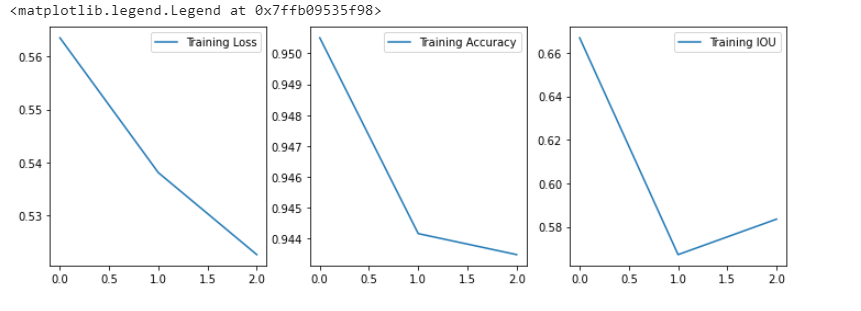
The network consists of a number of residual blocks with convolutions and down sampling with max pooling , as the end of the network single up sampling layer converts the output to same shape as input , with this network we have able to achieve 94 % training and validation accuracy and we were able to predict the probability of image belonging to lung opacity or not along with bounding boxes, we have only trained the model for 3 epochs as the each epoch was taking more than 2 hours



**Model Evaluation:**

Final model has consisted of three main components, data generator, model building and defining the loss functions

we have created a class to load the Training images and by loading the pixel data and created a mask by using the bounding box data and also defined a function with in to load the test image set as well , and then we have created functions to down sample the data and residual blocks and up sampling , the model consists of batch normalization and convolution layers as well , finally to measure the model we have implemented loss function as Inter section over union, Accuracy as mean IOU and we have used Adam Optimizer , the most important parameters are loss and network architecture , we have evaluated the model using training accuracy , below is the graphical representation , as we can see from the graph as the number of epochs are increased the training accuracy has been stabilized and Training loss has been reducing



**Comparison to Bench Mark:**

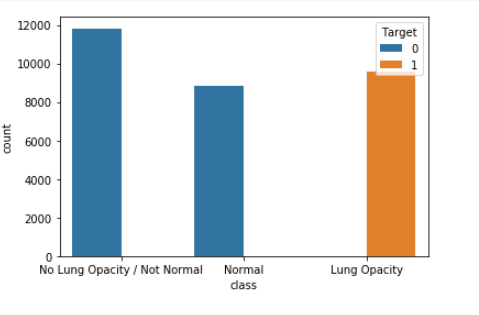
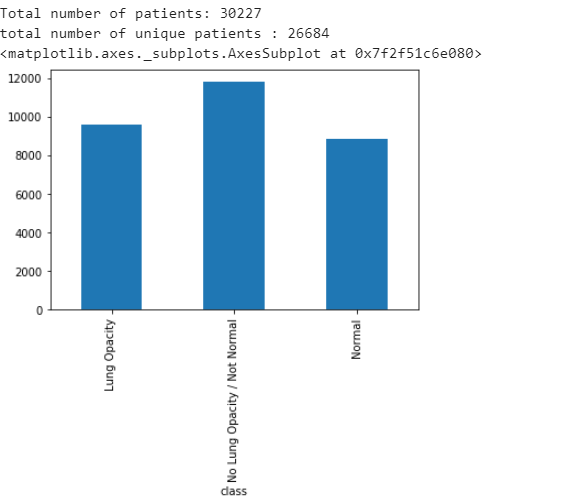
The benchmark we have set for out model has been 95% accuracy, with out third model we have come close to achieving that bench mark, we would probably will be able to achieve it if we could train the model on more images and for more epochs with more computational resources

the Initial model we have trained had an accuracy of 84 % for over 10 epochs this is also another thing we have considered to keep as out benchmark while deciding on benchmark, as this model is trained with basic details any model that is created later should surpass this accuracy

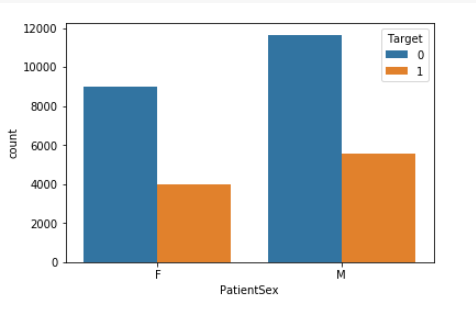
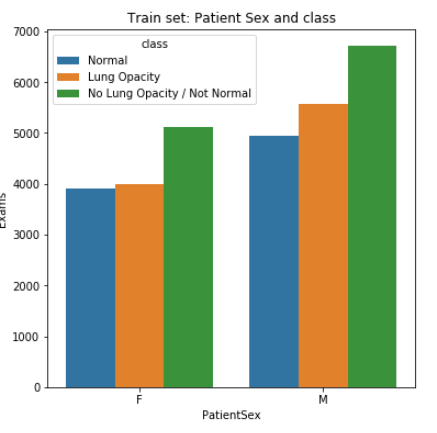
**Visualizations:**

Here are some of the visualizations as part of data pre processing

1. Graphical Representation of portion of each class and distribution of each class



1. Graphical Representation of patient Gender based on Target and Target type

Based on above all three different visualization here are the conclusions

1. Population of male patients is higher and they are also one with more cases of lung opacity and Total number of unique patients are 26684
2. There are three different classes one with class as lung opacity which is what we are trying to predict

**Implications:**

Pneumonia accounts for over 15% of all deaths of children under 5 years old internationally. In 2017, 920,000 children under the age of 5 died from the disease. 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 (atelectasis 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. When available, comparison of CXRs of the patient taken at different time points and correlation with clinical symptoms and history are helpful in making the diagnosis.

**Automating Pneumonia screening in chest radiographs, providing affected area details through bounding box. Assist physicians to make better clinical decisions or even replace human judgement in certain functional areas of healthcare (e.g., radiology). Guided by relevant clinical questions, powerful AI techniques can unlock clinically relevant information hidden in the massive amount of data, which in turn can assist clinical decision making.**

**Limitations:**

As the traditional method of identifying the pneumonia by doctor is still needed even if we automate this process because there’s always a risk of mis classification where when the patient has lung opacity but the model predicted as no lung opacity , this case is more important than misclassifying someone who doesn’t have lung opacity as opacity because this will lead to death of the person , I would use this model as reference and leave the final decision to doctor and some of the limitations include the data is always changing and we have to keep on training the model with new data and build a more deep network and train for more epochs

**Closing Reflections:**

Some of the learning from this capstone are handling the image data is really difficult and now that we can see firsthand that how the computation has been one of the biggest road block while training the images , with more learning experience I would find a way to load the data set fast by using GPUS and build different models using architecture like RCNN or mask CNN which is really heady in terms of coding standards