CS584 – Theory/Application of Data Mining

Pneumonia Type Detection from X-Ray Scan Images using Convolutional Neural Networks

By - Sharukh Afridi
December 10th, 2021

Abstract

Pneumonia, a respiratory disease, is caused by filling up of the air sacs of the lungs. The detection of the pneumonia type is time critical, as the treatment can not begin until then. The project makes use of convolutional neural network model to classify the bacterial, viral pneumonia and normal cases from chest X-Ray scan images. The dataset available is not of very high volume due to privacy issues of accessing medical scans of patients and also of the available data, the majority of the images are of bacterial pneumonia. Hence a combination of oversampling, data augmentation involving various techniques such as random flips along axes, zoom in zoom out, slight rotations etc., undersampling was performed to achieved good results with more balanced data. ResNet50 model is used for classification, with the usage of pre-trained weights learned from ImageNet. Considering that the similarity between medical images and that of ImageNet isn't too high and yet good results are achieved, a key take away is that transfer learning is beneficial in the field of medical imaging field. Making use of this, the project develops a model which is beneficial in saving critical time of patients and money towards expensive trivial diagnostic procedures.

1. Introduction

Pneumonia is the inflammation of the air sacs of the lungs called alveoli which results in filling up of the alveoli with fluid. This causes difficulty in breathing in cases of late detection and may even prove to be fatal. The treatment for pneumonia by the medical practitioners is dependent on the type of pneumonia – bacterial or viral. For instance, antibiotics are used for the treatment of bacterial pneumonia, which cannot be used in the case of viral Pneumonia. Chest X-Rays, blood tests are common methods for the detection of pneumonia. These methods help in detecting whether the patient has pneumonia or not, but not readily helpful in determining the type of pneumonia. The existing methods to find out the type of pneumonia involves culture generations and further different type of blood tests and thus the microorganisms causing the pneumonia. These methods are time consuming and may take up to 48 hours for the reports to be generated after lab testing procedures. In cases of severity and where time is an immense factor, this delay in pneumonia type detection can lead to loss of lives, as treatment cannot begin until the type of pneumonia has been detected.

Hence, addressing this issue by developing a model that can predict the type of pneumonia readily and thus help in saving vital time is important. For classification of the images, it is essential to observe the presence of blurry lesions in the images. This blurry lesion is known as Ground-Glass Opacity or GGO in medical terms. The location of the GGOs in the lungs is deterministic in diagnosing the type of the pneumonia. For instance, in the case of viral pneumonia, the GGOs are present over the entire lungs whereas in the case of bacterial pneumonia, these GGOs are clustered in the center and lower part of the lungs. This project develops a model for detection and classification of pneumonia X-Ray images by using deep learned Convolutional Neural Networks (CNN). Convolutional Neural Networks and Image Processing methods are in existence used in diagnosis of various disease such as breast cancer, brain tumor, etc. Extending this idea to make use of Chest X-ray scan images, I make use of residual model ResNet50 along with various preprocessing techniques to develop a model which can detect the type of pneumonia. The below images are X-ray scan images of the three classes in consideration.

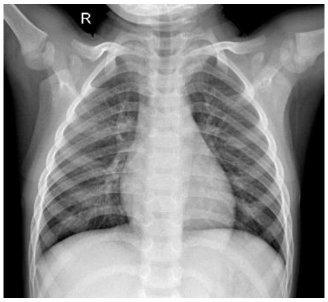


Fig.1 – Normal

Fig2. Bacterial Pneumonia

Fig3. Viral Pnneumonia

2. Problem Statement

The need for quick diagnosis of the pneumonia type to begin the treatment as soon as possible, saving the precious lives of critical patients and saving the money of the patients spent on the diagnosis is the problem aimed to be addressed. Below are two images showing the problem which the project aims to address.

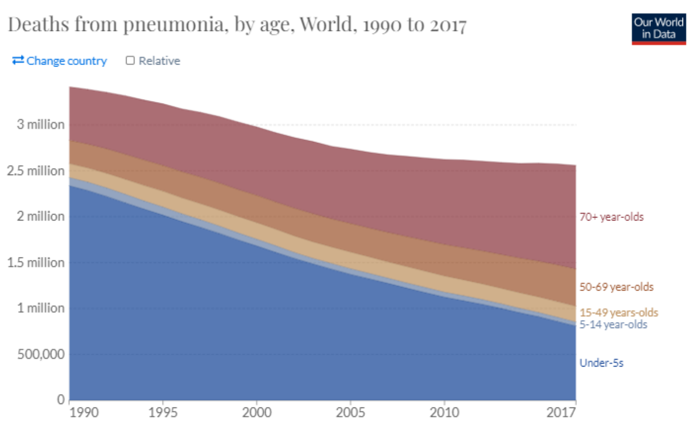


Fig4. Yearly deaths from Pneumonia

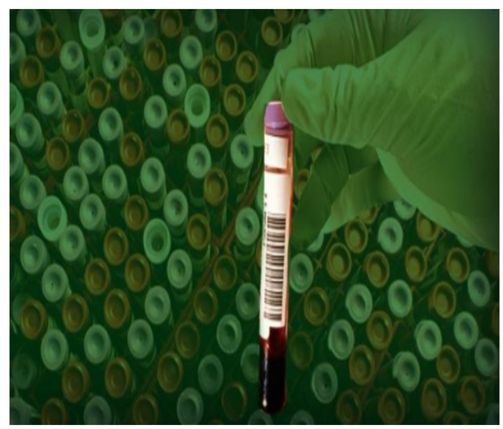


Fig5. Diagnosis costs

As can be observed in the Fig.4, the number of deaths annually from Pneumonia is over 2.5 million across the world. For US adults, pneumonia is the most common cause of hospital admissions other than women giving birth. About 1 million adults in the US seek care in a hospital due to pneumonia every year. The diagnosis costs involved through the trivial methods of culture generations and blood tests can cost anywhere up to $2000. Thus, there exists a problem in terms of saving lives and reducing costs which needs to be addressed, which the project tries to achieve.

3. Literature Survey

The idea for the project is based upon the IEEE paper titled ‘Pneumonia Detection Using Deep Learning based on Convolutional Neural Network published 2021, 25th International conference on information technology. One existing model I have come across is by the authors Apostolopoulos and Bessina, who have employed Deep transfer learning involving building a classifier on top of a pre-trained network to classify obtained dataset images using VGG19, MobileNet, models with weights trained on ImageNet. In my approach, the CNN based machine learning algorithm will be used consisting of few layers such as input, output and between them are hidden layers. These hidden layers perform most work in terms of calculation. Each neuron in the convolutional layer is only connected to a small number of neurons in the next convolutional layer. This approach allows for the focus on small low level feature in the first hidden layer and then enhance into higher level features in the next hidden layer. This approach accounts for CNNs being successful in image recognition. The ResNet50 model is used for the classification problem. Residual networks have higher relative performance when compared with other pre-trained network architectures such as VGG, Inception, etc. in terms of accuracy as well as computation requirements. Also, the storage requirements in this approach is much less as compared to other approaches.

4. Methods and Techniques

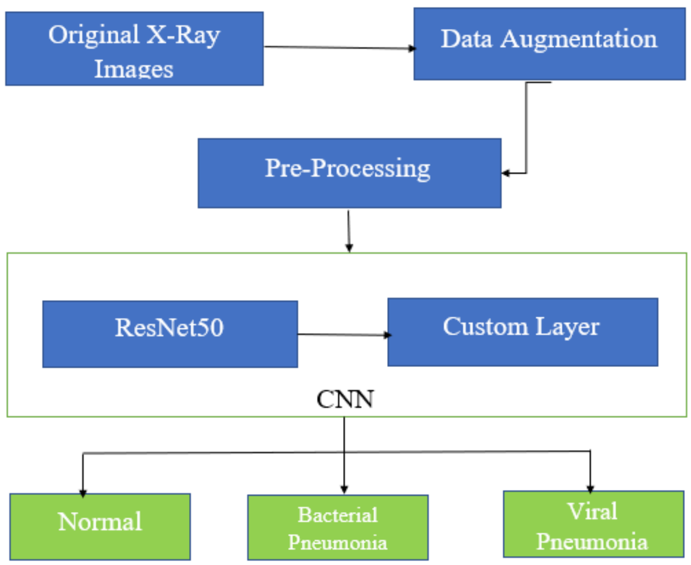


Fig6. Model Architecture

The overall methods and techniques used in the project is represented above as a Model Architecture. The original X-Ray images are initially subjected to Pre-Processing and Data Augmentation techniques. Once we have the data which is balanced and increased amount of data, is given as input to the CNN model. The CNN model consists of the ResNet50, which is a pre trained residual network, which is pre trained on ImageNet database. This is further added with custom layers to make up the CNN model. The CNN model is then able to classify the images into our three classes of Normal, bacterial and viral pneumonia.

Transfer Learning

Keras Appliacations' pre trained model – ResNet50 enables to use weights from ImageNet that are pre calibrated to make predictions. Feature extraction is achieved by using the attribute include_top = False to remove the last dense layers.

```
#Creating the base pre-trained model
 base_model = ResNet50( weights='imagenet', include_top=False, input_tensor=img_conc )
```

Due to the number of layers of ResNet50, we have large amount of parameters. By freezing the layers we make sure that those values do not change and thereby saving time and computational costs. In the model, all blocks, except for the last block of the ResNet50 are freezed. This is achieved by the following code-

```python
for layer in base_model.layers[:144]:
    layer.trainable = False
```

Result -

```
136 conv4_block6_1_relu - False
137 conv4_block6_2_conv - False
138 conv4_block6_2_bn - False
139 conv4_block6_2_relu - False
140 conv4_block6_3_conv - False
141 conv4_block6_3_bn - False
142 conv4_block6_add - False
143 conv4_block6_out - False
144 conv5_block1_1_conv - True
145 conv5_block1_1_bn - True
146 conv5_block1_1_relu - True
147 conv5_block1_2_conv - True
```

This is followed by connected the pre trained model with new layers. Global spatial average pooling layer is added followed by fully connected layers and a logistic layer.

Learning rates and ReduceLROnPlateau

Learning rate is the most vital hyper-parameter while training the network. The weights of the network are updated after each epoch depending on the learning rate. The weights (θj) are updated according to the formula below:

$\theta j + 1 = \theta j - \alpha\, [\partial J(\theta)/ \partial (\theta j)]$,where

$j \rightarrow$ number of epochs
$J(\theta) \rightarrow$ loss function
$[\partial J(\theta)/ \partial (\theta j)] \rightarrow$ gradient of the weight θj
$\alpha \rightarrow$ learning rate
$\theta\ j+1 \rightarrow$ value of updated weight

Having an optimal learning rate is very essential as a higher value of learning rate causes premature convergence of weights and thus inaccurate results whereas a smaller learning rate can cause gradient descent problem. ReduceLROnPlateau from tensorflow.keras.callbacks is used to get a new and reduced learning rate whenever the loss does not improve which indicates that the current learning rate is not working. The below snippet shows the reduction in learning rate 'lr' when the loss does not improve in epoch 43.

```
Epoch 00043: loss did not improve from 0.38835
949/949 [==============================] - 1145s 1s/step - loss: 0.3887 - accuracy: 0.8531 - lr: 2.5000e-05
Epoch 44/50
949/949 [==============================] - ETA: 0s - loss: 0.3852 - accuracy: 0.8537
Epoch 00044: ReduceLROnPlateau reducing learning rate to 1.249999968422344e-05.
```

5. Discussion and Results

5.1 Data Analysis

The dataset is downloaded from Kaggle – Chest X-Ray images (Pneumonia). This data set has total of 5863 images. The total size of the dataset is 1.24 GB. The training data is highly imbalanced, with a greater number of bacterial pneumonia images and much lesser images of normal patients and viral pneumonia images. Also, the image sizes were not same. These issues needed to be addressed in preprocessing. The below table shows the data imbalance.

	Training Dataset
Bacterial Pneumonia	2530
Viral Pneumonia	1329
Normal	1349
Total	5208

Fig7. Data imbalance

5.2 Data Pre-Processing

Image resizing was performed using cv2 library using the method - cv2.resize(). All images were resized to 224 * 224 pixels. RGB reordering is performed, giving us the final input images of (224 * 224 * 3). Data augmentation is the major technique used in the preprocessing of images apart from oversampling and under sampling techniques. Data augmentation is used to avoid creating a class imbalance. A class imbalance can cause the model to be overfitting on training and result in false positives, leaning towards class with more data, which in my case would be bacterial pneumonia. In-order for the model to avoid false positives, it is critical to have balanced data. For the implementation of image augmentation, the python library Albumentations is made use of. The techniques made us of to created augmented images are zooms, shifts in X & Y axes, Random rotations in the range of +15 to -15 degrees, Random flips along axes, Contrast, brightness changes, distortion & scaling changes.

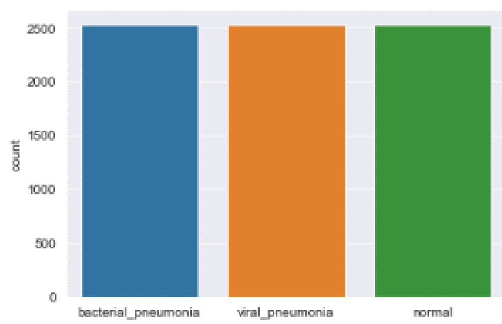


Fig.8 Oversampled data to avoid Data imbalance

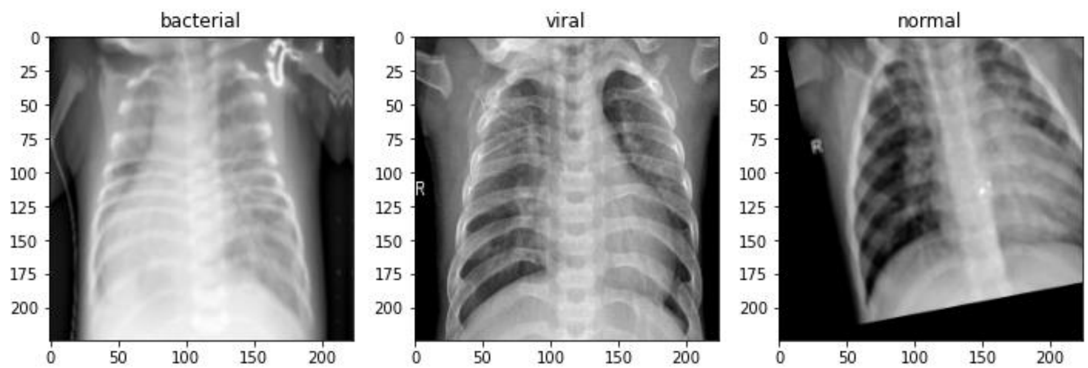


Fig9. Sample augmented images created from the techniques discussed above

5.3 Evaluation Metrics

Accuracy - The evaluation metric used is accuracy, which is the ratio of true results to the total number of examined cases. Accuracy is a valid evaluation metric for classification problems which are well balanced and not skewed or having class imbalance.

The formula for accuracy can be realized as follows:

Accuracy = (TP+TN) / (TP + FP + TN + FN)

Where, TP is True Positives, TN is True Negatives, FP is False positives and FN is False Negatives.

Precision - Precicion gives the answer to the question, what percent of predicted positives is truly positive?

The formula for precision can be realized as follows:

Precision = (TP) / (TP + FP)

Recall – Recall is the term which defines what proportion of actual positives is correctly classified?

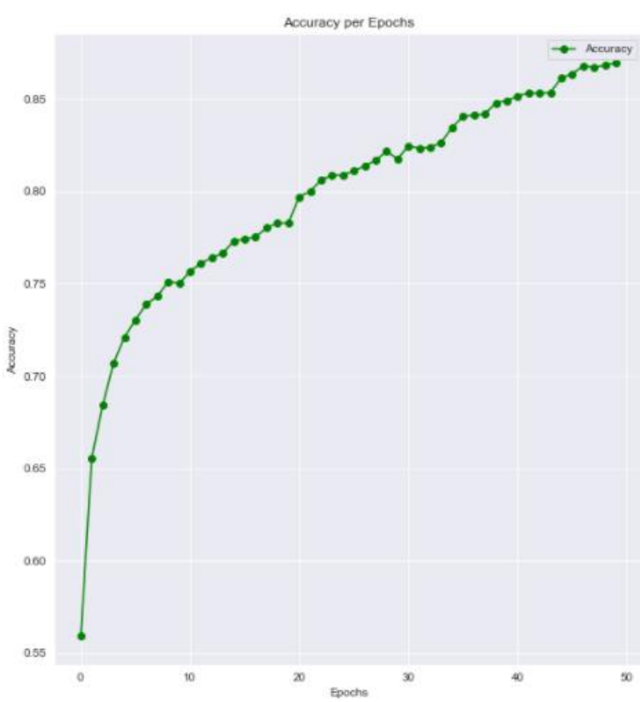
The formula for precision can be realized as follows:

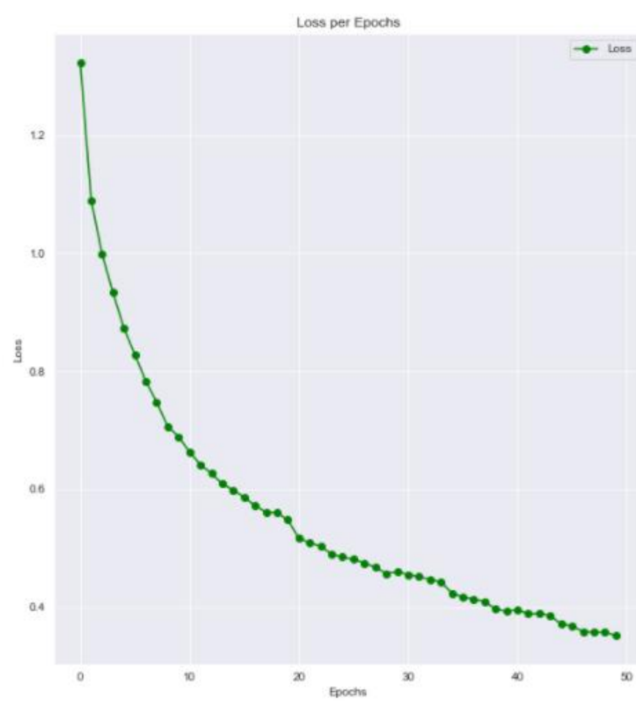
Recall = (TP) / (TP + FN)

5.3 Experimental Results & Cross Validation

The model is run with the setting configuartion parameters such as learning rate, epochs, batch_size etc . The results are evaluated through the evaluation metrics stated above which are accuracy, precision and recall.

Accuracy & Loss graphs





The model accuracy as can be seen from the graph improves with the increase in epochs and reaches overall accuracy of greater 0.85. The Loss per Epochs graph shows that the model loss decreases and reaches close to 0.35 with higher epochs

Precision & Recall

```
Precision_bacterial of the model is 0.82
Recall_bacterial of the model is 0.87

Precision_viral of the model is 0.79
Recall_viral of the model is 0.71

Precision_normal of the model is 0.95
Recall_normal of the model is 0.93
```

As can be seen from the above snippet, the precision & recall values are observed for each of the three classes and the model performs the best for the normal class. This could be due to the absence of lesions in the chest-xray scan images of normal patients, whereas the lesions are spread in different areas for bacterial and viral cases.

Cross Validation

Cross validation was performed on the training and validation dataset in a ratio of 8 to 1 in different scenarios of data pre processing. The best method which was adopted involved Oversampling the original dataset followed by augmenting the oversampled dataset 3 times. Also the original dataset was undersampled and finally all these datasets were merged.

The other approaches which were follwed were 1) Original dataset a) Only Augmentation for 3 times, b) Augmentation twice c) Oversampling only d) Oversampling and augmentation for 1 time.

The metrics were better in the cases a) & b) which clearly indicated towards Data Augmentation being effective in improving the performance of the model. In case c) though the performance improved, it wasn't as good as cases a) & b). The second best approach was case d). Thus by repeated evaluation of the performances, the best approach was followed which is stated earlier.

6. Conclusion

The model achieves the objective of identifying the type of the chest X-Ray scan image with an accuracy of greater than 85%. From the various cases tried, it was inferred that the model performs well the images are augmented repeatedly but after a certain level of augmentation, it had stopped making significant difference. The major hurdle in the project was the availability of less number of scan images due to privacy issues of patients. Also, from among the available images, the data was more for the bacterial pneumonia class as compared to viral pneumonia and normal cases. A lot of time and efforts was required for the dataset pre processing. The whole success of the project is based upon the pre processing techniques. The albumenations library was immensely helpful in the augmentation of the images as it had varied methods of creating slightly different images.The use of pre trained weights learned from ImageNet though ResNet50 being so very useful in achieving the goals is a major takeaway that, that transfer learning is beneficial in the field of medical imaging field. Making use of the pre trained model and adding additional layers helped in achieving greater accuracy.

Thus the goals of the project are satisfactorily met using the techniques stated above which will help save the lives of the patients by quickly detecting the presence of pneumonia including the type, so that the treatment can begin and there is no delay in the treatment, which could be fatal.

The learnings from the project were immense, starting from exploring the various pre-processing techniques, the neural networks working and the working of ResNet50 model in particular and also the evaluation metrics.

6.1 Directions for future work

The model can be further imrovement with the techniques of varying the learning rates, batch size, and other configurations of the Convolutional Neural Network model. Also the effect of adding other layers, freezing layers etc. can be explored to improve the performance of the model. Furthermore the augmentation techniques can be added or removed to see the effect of performance of the model. Another improvement area is to develop a simple UI web application for the same, so that any person can make use of the model with ease.

References

1) Introduction to Data Mining, 2nd Edition – Pang-Ning Tan, Michael Steinbach, Anuj karpatne, Vipin Kumar
2) IEEE paper - 'Pneumonia Detection Using Deep Learning based on Convolutional Neural Network published 2021, 25th International conference on information technology
3) Dataset - https://www.kaggle.com/paultimothymooney/chest-xray-pneumonia
4) Chest X-ray Image View Classification - https://lhncbc.nlm.nih.gov/system/files/pub9175.pdf
5) X-ray classification https://github.com/obendidi/X-ray-classification
6) https://arxiv.org/pdf/1712.07632.pdf preprocessing xrays by removing bones to increase accuracy
7) Wikipedia, StackOverflow, GeekforGeeks.