Pneumonia Detection using Convolutional Neural Network(CNN)

By: Temesgen Tesfay

September - 6th, 2021

Introduction

Pneumonia accounts for 15% of all deaths of children under 5 years old, killing 808,694 children in 2017. Being better able to predict pneumonia quickly and accurately and at a low cost could have a large impact on healthcare related to pneumonia and on patient outcomes. In this project, I seek to use transfer learning to build a model that can accurately diagnose pneumonia based on chest X-rays.

Description of the Pneumonia Dataset

The dataset is organized into 3 folders (train, test, and validation) and contains subfolders for each image category (Pneumonia/Normal). There are 5,863 X-Ray images (JPEG) and 2 categories (Pneumonia/Normal). From the data description: "Chest X-ray images (anterior-posterior) were selected from retrospective cohorts of pediatric patients of one to five years old from Guangzhou Women and Children's Medical Center, Guangzhou. All chest X-ray imaging was performed as part of patients' routine clinical care […] For the analysis of chest x-ray images, all chest radiographs were initially screened for quality control by removing all low quality or unreadable scans. The diagnoses for the images were then graded by two expert physicians before being cleared for training the AI system. In order to account for any grading errors, the evaluation set was also checked by a third expert."

- **Source**

  https://www.kaggle.com/paultimothymooney/chest-xray-pneumonia

  Description

  - The data have 1.2 GB of images, and files in .jpg containing the respective labels.
  - Classes labels, image is labeled as :
    - Pneumonia

- Normal

The dataset has 5856 images, the images are classified among train, test, and validation. 87% allotted to train set, 10% to test set, and the remaining 3% is to the validation set. There is a data imbalance between the training and test.

Exploratory Data Analysis(EDA)

The dataset contains over 5000 training images and additional 624 images for testing. Each image is stored in .jpg files The Distribution of the label instances in the training set can be seen in the bar graph below:

The distribution of dataset group between normal and pneumonia classes

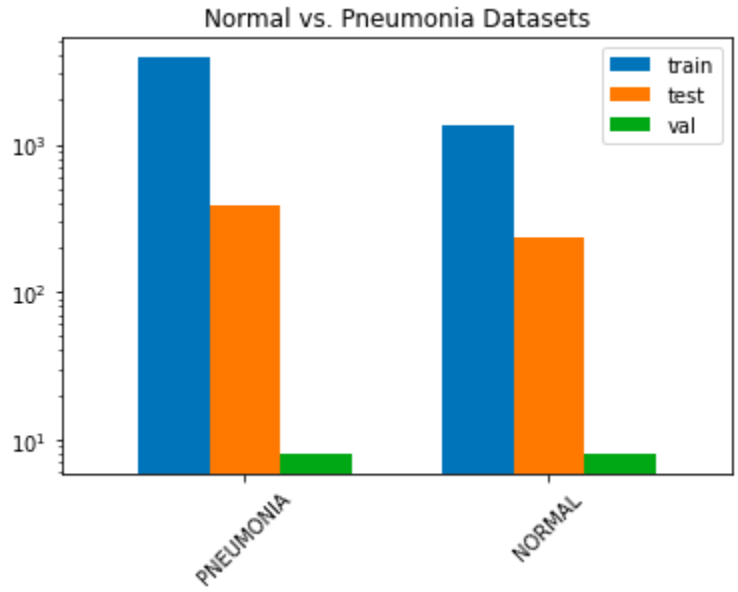

Fig 1.1 dataset proportion between normal and pneumonia

Fig 1.1 the sample size among the group has a huge difference hence, used log distribution to resize the bar graph for visualization. There is a clear indication of data imbalance.

The Distribution of Normal and Pneumonia images

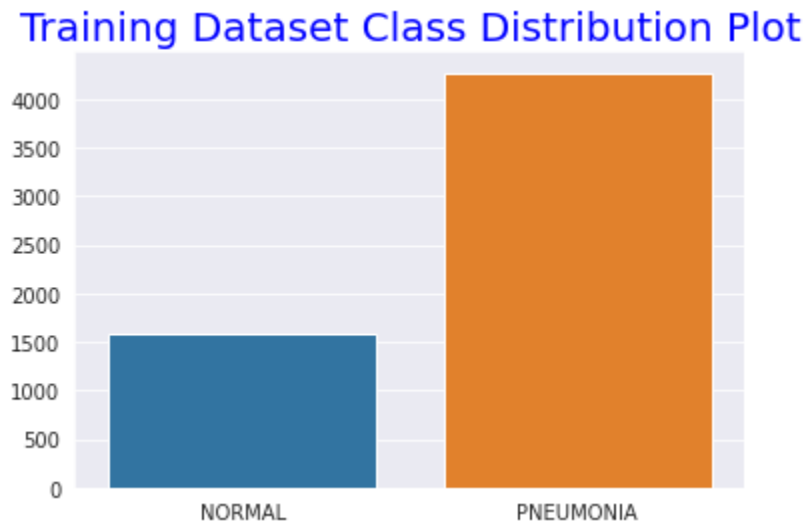

Fig 2.1 normal and pneumonia distribution

In the training dataset, we find that 75% of images show a patient with pneumonia.

A class of pneumonia versus normal lung images

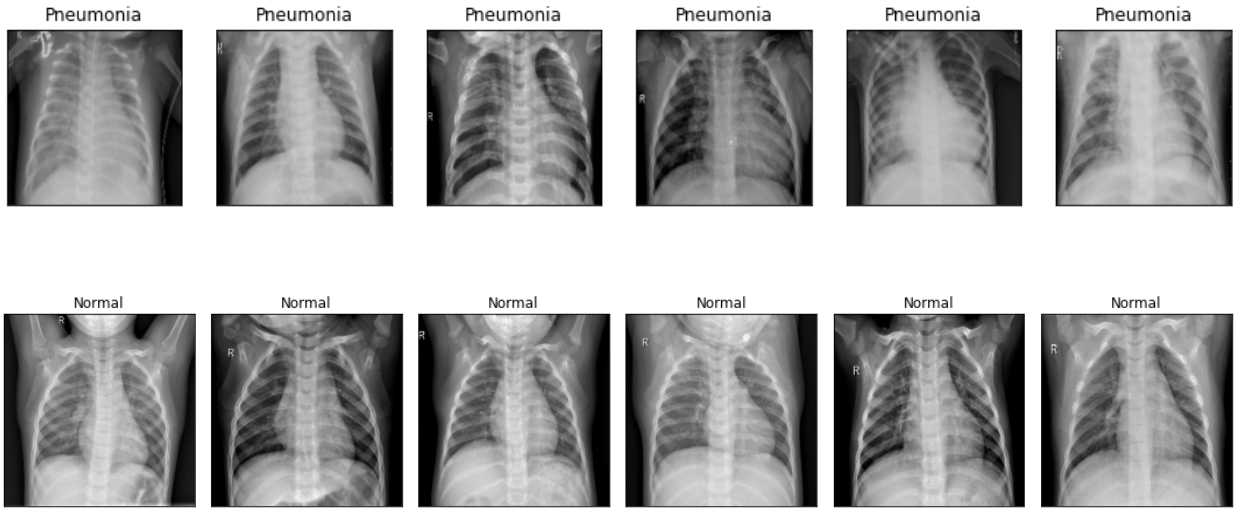

Fig 3.1 Sample images of pneumonia and normal chest-Xray image

Many Of the pneumonia-positive lungs show a wide irregular whitish area versus the normal dark brown regular pattern.

Data Augmentation

Data augmentation techniques deal with cases where the training data is limited. The techniques allow us to modify or even artificially synthesize more data thereby boosting the performance of a machine by reducing overfitting. The idea is to alter the training data with small transformations to reproduce the variations. These approaches that alter the training data in ways that change the array representation while keeping the label the same are known as data augmentation techniques. Using random rotation and random flip, However, model performance dropped in accuracy implies, despite the data misproportion the model trained well with the negligible effect of bias and overfitting problems. On the other hand, I have learned that the application of data augmentation should rely on the model performance, in particular to imbalance the dataset might make the difference worse likewise, the model performance.

Transfer Learning Primer

Transfer learning (TL) is a research problem in machine learning (ML) that focuses on storing knowledge gained while solving one problem and applying it to a different but related problem. For example, knowledge gained while learning to recognize cars could apply when trying to recognize trucks. From a practical standpoint, reusing or transferring information from previously learned tasks for the learning of new tasks has the potential to significantly improve model performance.

Model Building

For this task, we used the ResNet50V2 model which was previously trained on a subset of ImageNet (a huge database of 14 million images manually labeled with over 22,000 categories) as part of the ImageNet Large Scale Visual Recognition Challenge. While much of the data are very different from our pneumonia data, the abstract features that are created can still be very useful for us. In this case, I froze all the layers and added a new top layer to be trained on these data and output pneumonia predictions.

The base model of CNN architecture was created using ResNet50V2 Keras applications, and its parameters are as follows:

- Image resolution input shape = (height = 220, width, 220, and 3 channels = (220,220,3)
- Training set size: 5100
- Validation set size: 124
- Test set size: 624
- Batch size = 32
- Metric: Recall or Sensitivity, and F1 score

The model's Training and Validation accuracies and losses plots:

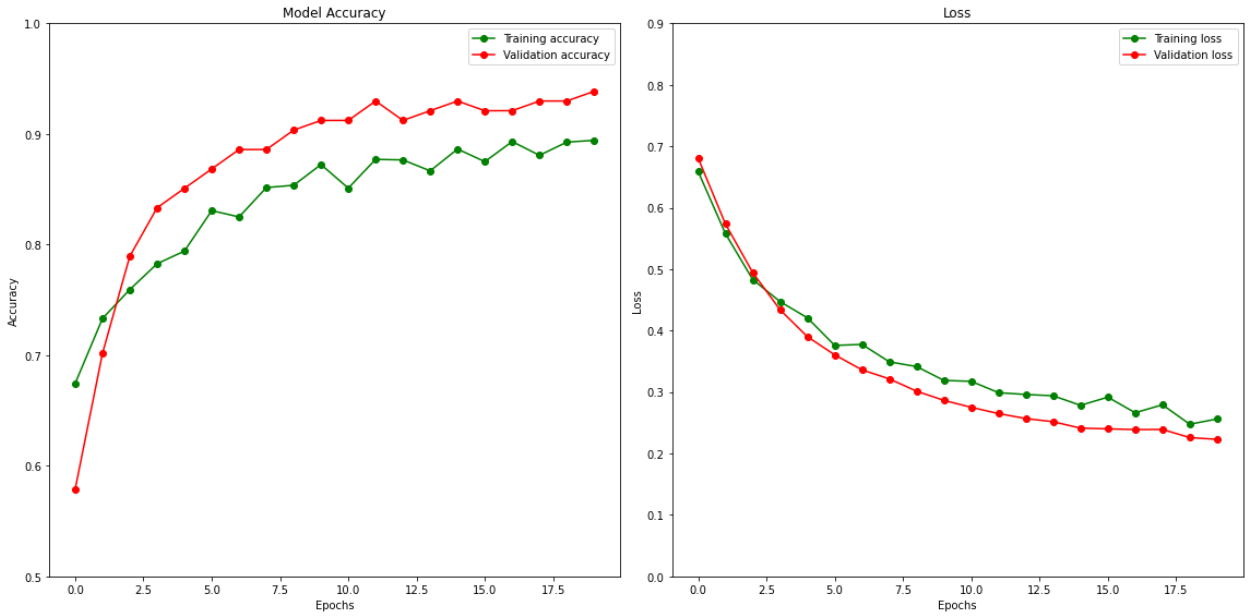

Fig 4.1 Model Accuracy vs Loss Distribution

In the first few epochs, transfer learning performs better accuracy but, as it runs to more epochs the model leads to overfitting because the validation dataset is infinitesimal for the model versus the training dataset.

Generally, Transfer learning provides high accuracy in the early stages of the epoch, it performs better with less computational time compared to traditional deep learning.

Confusion Matrix from the Transfer Learning

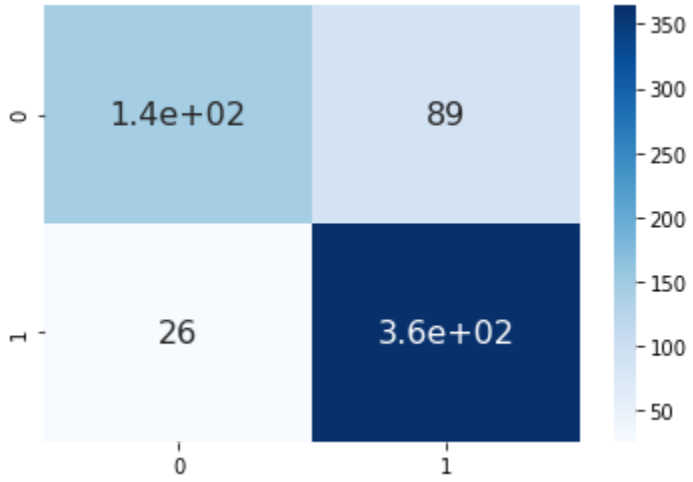

Fig 5.1 Confusion Matrix

The confusion matrix classification report metrics return the following:

	Precision	recall	f1-score	support
0	0.86	0.55	0.67	234
1	0.78	0.95	0.85	390
accuracy			0.80	624
macro avg	0.82	0.75	0.76	624
weighted avg	0.81	0.80	0.78	624

The focus of this project is to minimize false negatives or incorrectly classified patients. Therefore, Recall will be prioritized in evaluating model performance. However, the

impact of the high false positives (89 out of ~624) that result from optimizing F3. Those poor patients would be subjected to more tests, and there would be a lot more expense involved.

The project did experiments to improve recall outcomes. Some of the experiments implemented:

- Implemented fully connected convolutional neural network
- Executed various regularization techniques
- Transfer Learning from pre-trained data images stored in the file named-imagenet.

Since we want to prioritize recall, we decided to use an F-Beta score to allow us to weigh precision and recall separately. In F-Beta, a Beta score of > 1 weighs recall greater than precision, and the opposite is true for F < 1. In our case, we decided to set Beta equal to 3 since we want to highly prioritize recall.

Precision, recall, and F3 score

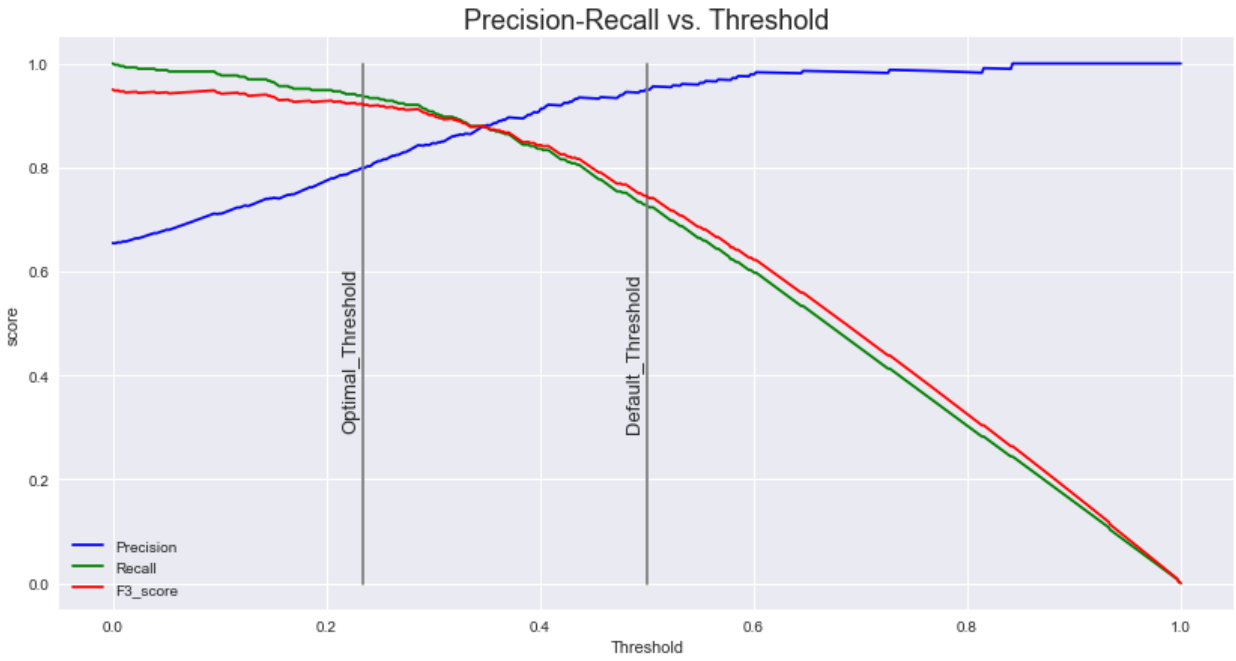

Fig 6.1 Precision, recall, and F3 score value as a function of model threshold

The model returns the highest recall at 0.95, when f_score is 3, the corresponding optimal threshold at 0.215. To keep the false negative error below 5% choosing the threshold range between 0.0 to 0.215 provides a recall above 95%. Generally, the choice of the threshold depends on the nature and the magnitude of the problem.
In order to get recall > 95% required a corresponding sacrifice to precision. This is not without cost as we have 89 false positives out of ~624 patients. Those poor patients would be subjected to more tests, and there would be a lot more expense involved.

Conclusion

In this project, transfer learning was used to predict pneumonia in images of chest X-Rays. The model generated was able to perform classification with a recall of ~.95 and a precision of .78. While an F3 score was optimized here, it would be worthwhile to speak with stakeholders to better understand the business need and whether or not a higher recall would be worth the associated sacrifice in precision.

Insufficient data size was a significant obstacle here, though the issue was minimized using data augmentation. However, in the future having more labeled data could be very useful to improve the predictive power of the model.