Pneumonia Detection Using CNN

Final Report

Group 41

Sherry Li, Yu Xin Li, Grace Xu

University of Toronto

APS360

April 13, 2022

Words count: 2385

1 Introduction

The prevalence of coronavirus started in 2019 also increases the diagnosis rate of diseases related to respiratory systems and the most frequent diagnosis for severe COVID-19 is severe pneumonia. Besides, pneumonia is still one of the top 10 leading causes of death in Canada [1]. Currently, physicians are only able to detect pneumonia with 47% to 69% accuracy, thus a more accurate and efficient model is needed to fulfill the increasing demand.

The purpose of the paper is to explain the convolutional neural network (CNN) built by the design team used for pneumonia detection. The goal of the model is to successfully predict if patients have pneumonia or not based on their chest X-rays and achieve an accuracy of at least 80% on the testing data. The basic structure of the CNN is shown in Figure 1, which imports the X-ray films and outputs one prediction: normal or pneumonia.

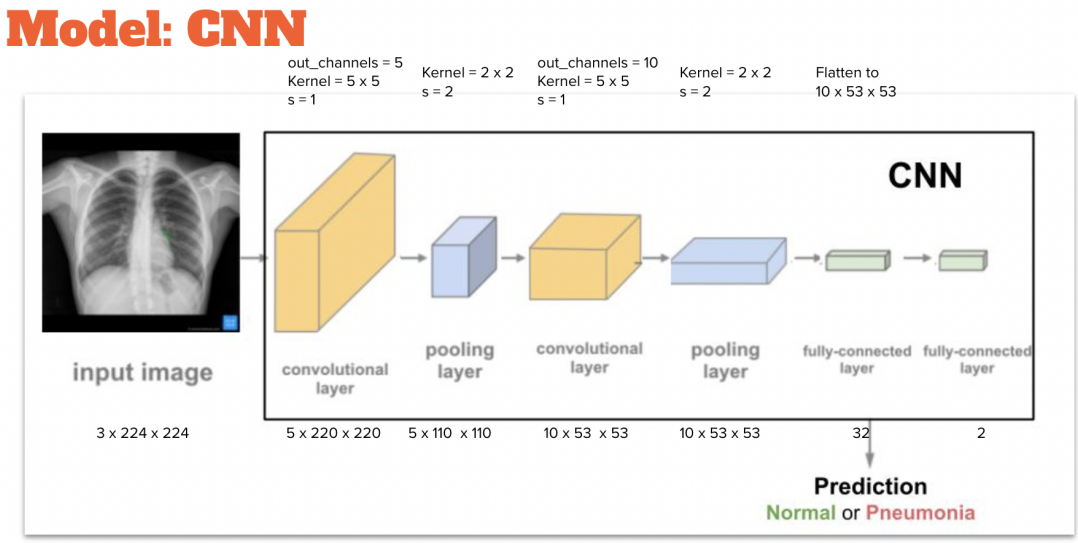

Figure 1. General Structure of the CNN

Port Score which is a screening tool that is used for clinical diagnosis of pneumonia. It will ask for patients' age, gender and some other True/False questions related to past history of similar diseases [2]. Depending on the answers, the number of points will differ and the sum of

the points will be recorded to decide the risk levels of the patients based on the specific metrics[2]. Without the help from machine learning and an actual physician, this screening tool cannot provide sufficient information for the patients to be cautious and thus the accuracy and efficiency of the convolutional neural network are more important.

2 Data Processing

Our data comes from the Kaggle dataset "Chest X-Ray Images Pneumonia" which includes 5,863 X-Ray images of anterior-posterior chest X-Rays with two classification labels collected from Guangzhou Women and Children's Medical Center's retrospective cohorts of children aged one to five years old.

A manual quality control inspection is conducted where the chest X-rays images are manually checked with infection identified depending on whether infiltrates appear as white spots in the lungs so that the datasets are confirmed to have the correct label.

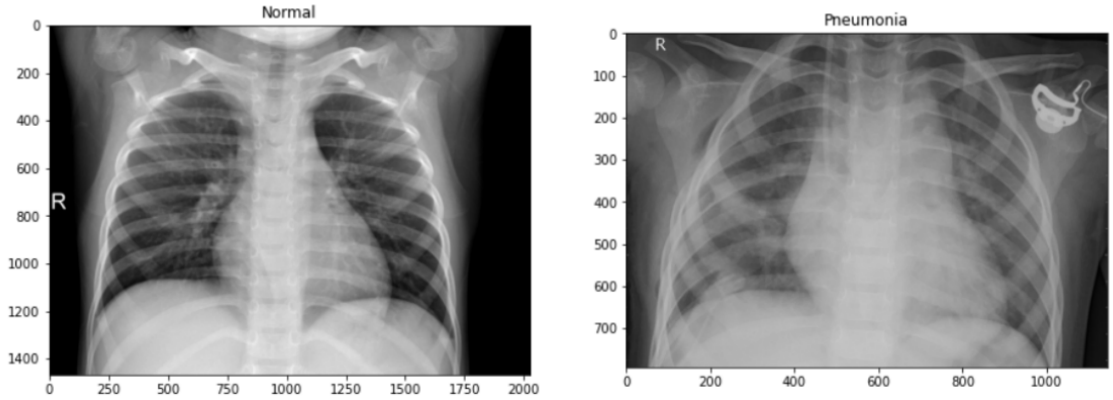

Figure.2 A pair of sample data with normal and pneumonia labels respectively.

After the classification of the dataset is checked, the zip file is downloaded and uploaded to google drive. Mounting the drive to collab, the datasets are unzipped in the roots/datasets directory and copied in the host’s local storage.

It is observed that the data were split into three folders with 5216 images in training, 16 in validation and 624 in testing originally. The images are reassigned to meet a ratio of 6:1:1. The testing folder will not be accessed until the model is finalized and ready for the last test.

Additionally, it is found that in the training set, the data provided for each label is extremely imbalanced where 73 percent of the images were from Pneumonia patients. The images with normal labels are then duplicated to improve training accuracy.

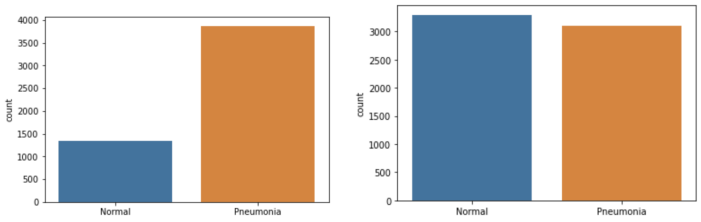

Figure.3 Label percentage of training dataset before (left) and after (right) balancing

Finally, data augmentation is applied where more training samples are generated by horizontally flipping, shifting, shearing and zooming the existing data to decrease the risk of overfitting and boost the performance of deep networks. The images are also converted to grayscale as well as normalized from [0, 255] to range [0,1] for minimized illumination effect and faster CNN convergence.

3 Architecture

As shown in Figure 1, the final convolutional neural network contains 6 layers: 2 convolutional layers, 2 pooling layers and 2 fully-connected layers. The parameter values (ex. kernel size, input color channel etc.)for each layer are shown in Figure 1. The image size and the kernel size have been shown in the diagram as well. After the optimization, the final model has a batch size of 128, 20 epochs and a learning rate of 0.001.

The list below contains other details during training process:

- Activation Function: **ReLU** (easy derivatives)

- Loss Function: **CrossEntropyLoss** (the model is dealing with a classification problem)
- Optimizer Function: **Adam** (best known optimizer)

Due to the large amount of processing data, it is suggested to enable cuda during training thus the training time can be shortened.

4 Baseline Model

The baseline model used for comparison is ANN which contains only two fully-connected layers as shown in the Figure below. Since it is a simple model and has only a few layers, the baseline model is easy to build and train. The specific number of input size, kernel size and final output are labeled as well and the functions used during training are the same as the CNN model. As shown in the Figure(right), the model stops training after 4 iterations and thus is not able to abstract details from the input films. Consequently, this baseline model has a much lower final training accuracy of 51.52% compared to our CNN model which has an accuracy of 84.3%. This comparison also highlights the high compatibility of convolutional neural networks while training on images.

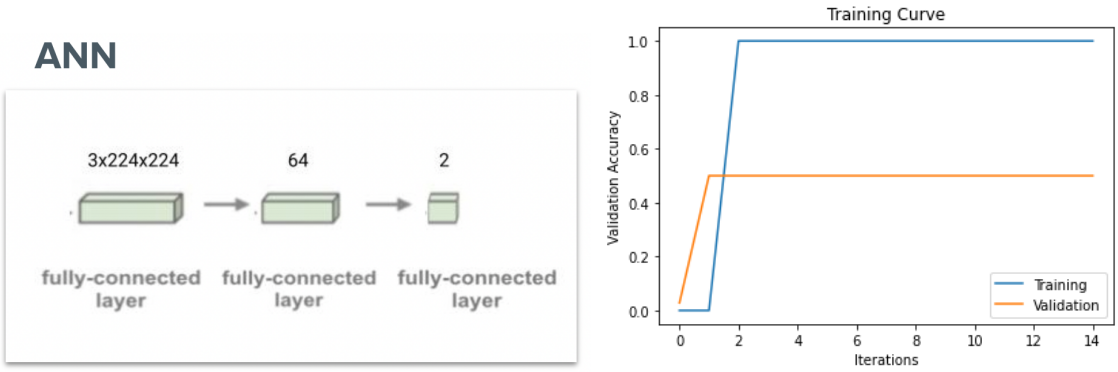

Figure.4 General Structure of ANN Model (left) and Training Curves (right)

5 Quantitative Results

For our model, the final hyperparameters that we used are batch size = 128, learning rate = 0.001, and number of epochs = 20.

To demonstrate how well our model performed, we plotted the training loss and accuracy curves for both the validation and testing datasets, shown in figures [] and [].

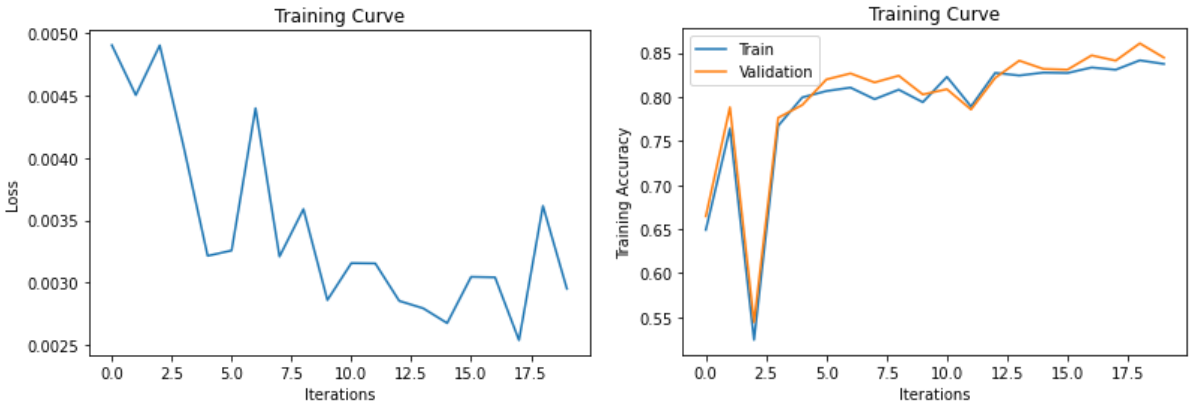

Figure.5 Training loss and accuracy curves for the validation dataset.

Looking at the loss curves for validation, the loss has an overall decreasing pattern for the first 10 epochs, and it stabilizes and maintains a steady loss after 10 epochs. This is reflected in the accuracy curve, as the maximum accuracy is reached at around 10 epochs. A few more epochs were run, just to confirm that the model is indeed not learning anymore. Furthermore, the validation accuracy has also remained constant, which suggests that the model is not overfitting, as the training and validation accuracy are always fairly close.

Another quantitative aspect measured was the time it took to train the model. It varied based on the hyperparameters used. For the final model, both the validation and testing data took about 90 minutes to train.

6 Qualitative Results

Our data processing resulted in an image as brightest as possible without altering any content of its information. The features of the image are then learned with a combination of kernels which includes vertical and horizontal edge detections.

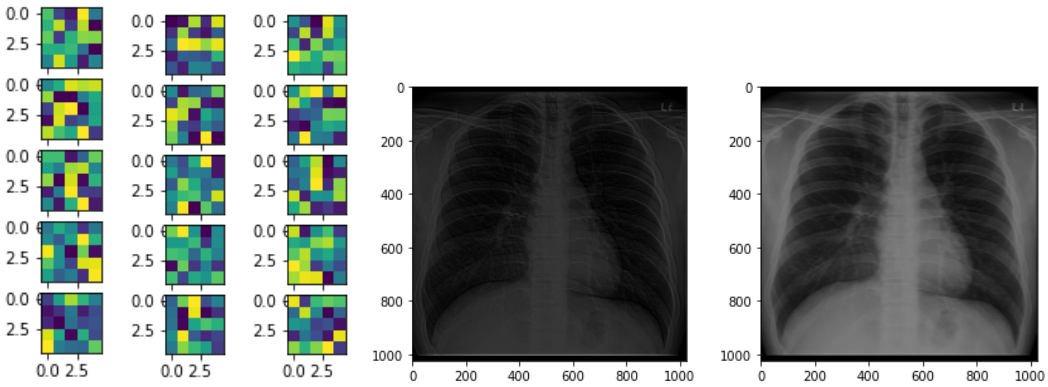

Figure.6 Kernel Visualization

Among 1172 test cases, our model is able to classify normal images with an accuracy of 87.5%, identifying pneumonia labels with an accuracy of 76.8%. Generally, our model's performance on normal images input is better than pneumonia images. Detailed confusion matrix is included in the *Discussion* Section.

7 Evaluation on New Data

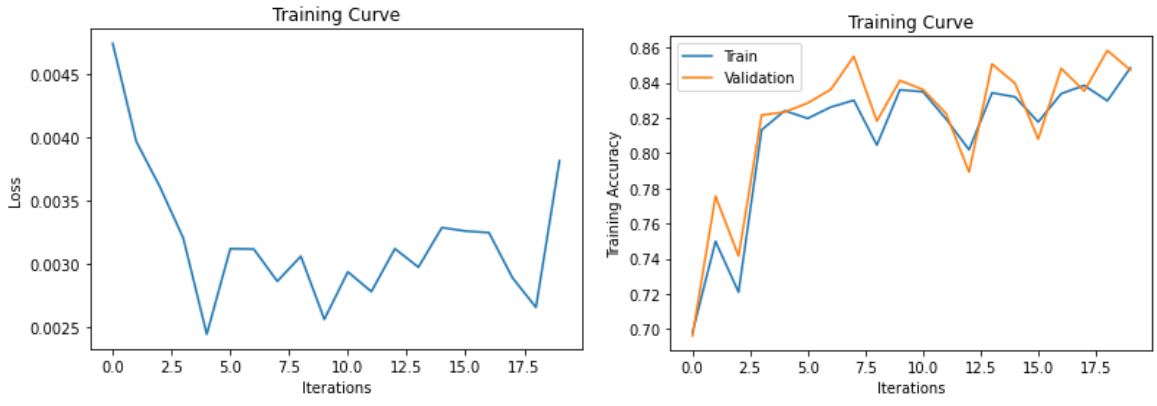

Figure.7 Training loss and accuracy curves for the test dataset.

To test out our model, we sectioned off some images at the beginning that we did not use during training and validation. These images were used at the very end to test the accuracy of our model. A similar pattern to the validation dataset was observed for the test dataset, except the model was able to learn all of the data in around 5 epochs. Afterwhich the loss remained constant and the accuracy reached its maximum. The same accuracy was achieved for both the validation and testing data (84%). This suggests that pneumonia can indeed be predicted through chest scans, and the model has not overfit to the training data. This makes sense because our dataset is not very biased, and the testing dataset is not much different than the training and validation datasets.

Compared with the ANN which has a testing accuracy of around 50%, our baseline model mentioned in Section 4, the CNN model has a much higher accuracy as it was able to learn to abstract the image features whereas the baseline model stopped learning new features after 5 iterations.

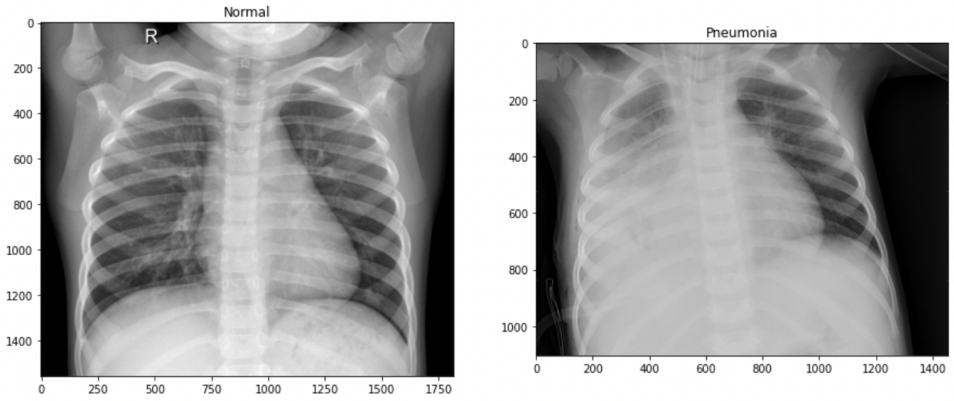

Figure.8 Random Samples of X-Rays

Random samples on the new data were chosen to test the accuracy of our model. For the figures shown above, our model predicted that a possibility of 87.5% for the left chest film is normal and a possibility of 76.5% for the right chest film is pneumonia.

8 Discussion

As the final test accuracy for our model is 84%, we believe that the model is performing well since it was able to predict with a greater accuracy than by eye (47% to 69%). This means that the model is useful and performs better than humans.

The results from the test dataset were then put into a confusion matrix as shown in figure 9. This is used to better understand how the data is segmented and which sections need the most improvements. A false positive may cause the patient to have unnecessary stress about their current situation and thus negatively impact their lives. On the other hand, a false negative is even more dangerous, because the patient would have thought they were safe when they are actually at a risk.

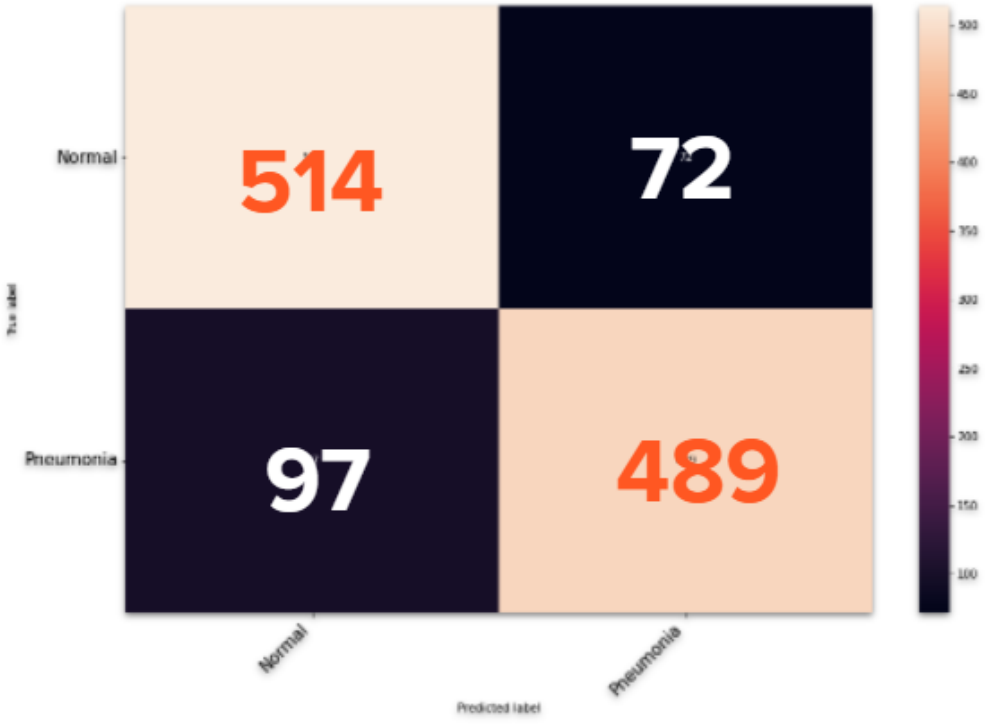

Figure 9: Confusion matrix with the true labels on the vertical axis and predicted labels on the horizontal axis.

At first glance, the model seems to perform well as true positives and true negatives contain the most amount of elements. However, something else to consider during training is maximizing the recall, which represents the proportion of all pneumonia cases that the model has accurately predicted. Numerically, it is the number of true positives over the number of all positive cases.

$$precision = \frac{TP}{TP+FN} = \frac{489}{489+72} = 0.87$$

$$recall = \frac{TP}{TP+FP} = \frac{489}{489+97} = 0.83$$

For our current model, our reall is lower than our precision, suggesting that the model could be made to be more harsh, predicting more positive cases. Harsher model is more prefered because we would rather have someone healthy diagnosed with pneumonia, than one who is sick and does not get the proper treatment because they have not been diagnosed. In other words, while maximizing the true positives and negatives, we should also make sure false positives are greater than false negatives.

Something we have learned is that finding the correct model to use can be difficult, and is the most crucial step of any machine learning problem. We originally started with AlexNet, as we were told that it was one of the best neural networks. However, with AlexNet, we found it interesting how we were only able to achieve an accuracy of 50%. This is likely because AlexNet contains the most common features, but chest scans are fairly niche and the features were not learnt through the AlexNet. Thus, in the future, we must be able to adapt accordingly depending on the size and nature of the project.

Although we spent a large amount of time developing the model, it is definitely not reliable enough to be implemented in the real world. We cannot trust machine learning 100%, because it is near impossible to build a model that predicts total accuracy. When diagnosing

patients, we can have both false positives and negatives that are dangerous for the users. Thus, the model should only be used as a quick test to narrow down the pool, but an experienced doctor should still check over the results using past cases and other medical tests.

9 Ethical Considerations

Although chest scans do not discriminate between race and/or gender, there are still some biases that exist within the dataset we chose. First, there may be some measurement bias with the equipment used, as different machines can produce different images of the same person. Additionally, there may also be some representation bias, as the demographic is for children one to five years old in China. Other demographics that are underrepresented may have different structures within the chest, and thus do not fit into the model produced.

The unbalanced dataset is the main limitation for our model. This may cause some ethical issues if it is used to predict pneumonia for another demographic. This is as the model likely works better for the group of people that makes up the majority of the training set, thus there is a disparate impact and the model is not fair towards all demographics. To produce a more fair model, more data can be collected that has equal representation from all groups. Or, to make it clear that the model should only be used for young children in China, to avoid ethical issues that may arise due to the biased model.

10 Project Difficulty Quality

Our model presents an infection identification that is capable of separating pneumonia patients' x-ray images from the normal ones depending on whether infiltrates appear as white spots in the lungs. Compared to an expected accuracy of 47% when predicted by human eyes, our model performs with a better accuracy of 84%. To increase the complexity of this project, we tried different models such as simple ANN, CNN, AlexNet. Surprisingly, our models turned out to

have better performance without the implementation of AlexNet. We also applied different augmentations composed of normalization, flipping, shifting, zooming and tested various combinations of these until a satisfying behavior was reached. In terms of project management, one of our members decided to audit the course during the project. Fortunately, this situation is anticipated in the risk register of the proposal which allows us to make timely adjustments following the guide, reassigning the responsibility of the member who dropped out.

11 References

[1]Regunath H, Oba Y. Community-Acquired Pneumonia. [Updated 2021 Aug 11]. In: StatPearls [Internet]. Treasure Island (FL): StatPearls Publishing; 2022 Jan-. Available from: https://www.ncbi.nlm.nih.gov/books/NBK430749/

[2]"PSI/PORT Score: Pneumonia Severity Index for CAP," MDCalc. https://www.mdcalc.com/psi-port-score-pneumonia-severity-index-cap.

Lastname, W. (2009). *Title of webpage*. Site Name. Retrieved July 3, 2019, from http://www.example.com