Deep Learning for COVID-19 Pneumonia Detection

Narayani Wagle

Background

In the ongoing coronavirus pandemic, hospitals are encountering an increased number of cases of pneumonia, a common result of COVID-19. Around 75% of hospitalized COVID patients develop some form of pneumonia and several develop more severe respiratory conditions shortly after the initial onset of pneumonia. Consequently, chest radiographs are often taken to initially diagnose COVID among patients by identifying the presence of pneumonia in the lung. Rapid and early detection of pneumonia in COVID patients allows for early isolation and treatment. The current approach uses deep learning to speed up COVID diagnosis and augment physician diagnosis and/or reduce physician error.

Approach

The current approach uses deep learning to detect COVID pneumonia using chest x-rays. The dataset consists of labelled chest x-rays of 4 classes: healthy lung, bacterial pneumonia, viral pneumonia, and COVID pneumonia. However, viral pneumonia is virtually indistinguishable from COVID pneumonia without extensive radiology training, thus purely viral pneumonia images are disregarded for the purposes of this project. The remaining dataset consists of 227 images from the remaining 3 classes, split into sets of 200 and 27 for training and testing, respectively. The approach plans to build off models from literature (See Works Referenced) using transfer learning due to the demonstrated success for pneumonia detection in the past. VGG16 has shown promise in detecting pneumonia, so this model is used as a baseline, with additional fully connected layers added to yield output for distinguishing between 3 classes. Due to limitations in a small dataset, 10-fold cross-validation will be used to determine an average final test accuracy. Model performance will be evaluated by the average final validation accuracy of each fold as well as the average accuracy on the test set. All implementation is done using Keras and Tensorflow in Python, with processing done using Google Colab.

Model Architectures

The Baseline model utilizes an implementation VGG16 included in the Keras library, pretrained using ImageNet weights and using a 150 by 150 by 3 input size. The fully connected layers of the original model are not used, and instead two fully connected layers with 256 and 3 nodes are added to allow categorization of 3 classes. These additional layers are trainable, while the remainder of the base model is frozen. To compensate for the small dataset, several data augmentations are applied before training as well, including shears, flips, shifts, and feature-wise normalization. The following models tested are all derived from the baseline model using transfer learning.

Model 1 maintains the modified VGG16 network with 3 added fully connected layers from the baseline as well as its hyperparameters. However, the extraneous data augmentations from the baseline were removed (e.g. vertical flips, major shifts), as these transformations are unlikely to be present within x-ray data.

Model 2 employs multiple additions to the baseline model. The first addition is of two sets of convolutional layers with 512 feature maps, akin to the last convolutional layers of the baseline. Channel-wise normalization layers after each convolution were also tested, but this resulted in a vanishing gradient due to the depth of layers and an overall loss of information, leading to less generalizability and a stagnant validation accuracy. To remediate this, normalization was removed, and a single skip connection was added from the end of the base model to the end of the additional layers. Dropout layers were also added after the max pooling of each convolution to mitigate overfitting.

Model 3 is an extension of Model 2 with additional modifications to convolutional layers and activation functions. Data transformations and hyperparameters from the baseline model

were once again kept, but the dropout layers of Model 2 were removed, and all activation functions for added convolution layers were changed to leaky ReLU instead of ReLU to reduce the dying ReLU problem. The skip connection of Model 2 was also preserved.

Simpler models trained from scratch not using VGG16 or other pretrained networks were initially considered as another solution to the issues described, such as the vanishing gradient problem. However, these models performed very poorly overall, and their descriptions are thus omitted, and they are discussed in the following section very generally.

Results

The training dataset of 200 x-ray images was split into 10 folds (of 20 images), each stratified to have roughly equal proportions of each image class. Each model was trained on each set of 9 folds for 20 epochs, with the weights of the model being reset before rotating the validation fold. In addition to observing final validation accuracy for each fold, each iteration of the model was subsequently evaluated on the test set of 27 images for a more consistent measure of generalizability.

Although each model performed well, efforts to improve the baseline model through transfer learning was largely unsuccessful, as shown by Table 1. Additionally, validation accuracy across epochs in the training of each model varied greatly even as training accuracy consistently increased, possibly indicating overfitting. Sample comparisons of loss and accuracy during training is given in the Appendix. Numerous attempts were made to prevent overfitting and erratic behavior using simpler architectures and various forms of normalization, but these attempts failed to generalize, yielding consistent 30-40% validation fold accuracy and 30-50% test accuracy, even after achieving 80-90% training accuracy.

Table 1: *Shown are the average final validation accuracies and average test accuracies across all ten folds of training.*

	Avg Final Val Acc	Avg Test Acc
Baseline	92.0%	77.8%
Model 1	92.0%	79.3%
Model 2	89.5%	78.5%
Model 3	86.5%	80.4%

Conclusions

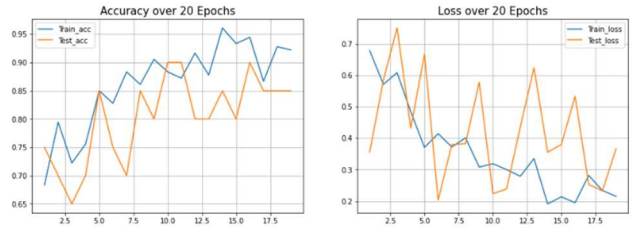
All models derived from the baseline generally performed the same as the baseline model. Tested architecture features include relevant data transformations and the preservation of information by implementing skip connections, additional convolutional layers, a leaky ReLU activation function, normalization, and the use of simpler models. However, these features failed to surpass baseline performance, and all models generalized sub-adequately to validation and test data. Lack of improvement in test accuracy may indicate the need for a more complex model structure, but due to a limited dataset leading to overfitting in training and variability in results, this cannot be determined for certain. Overall, modifications of the baseline model demonstrated generally on par COVID-19 pneumonia detection capabilities to pre-existing models.

Future Considerations

Several factors can improve COVID pneumonia detection rates in these models. Firstly, the models presented above classify three types of pneumonia (healthy, bacterial, COVID); however, a differentiation between COVID pneumonia and other viral pneumonias is yet to be completed, which may require more complex networks due to feature overlap. Secondly, the dataset used for training and testing only consisted of 227 images. Though the above models handled this issue with data transformations and cross-validation, an increase in the size of data would greatly improve training and classification. Finally, an ensemble of multiple models would greatly improve classification accuracy in more diverse settings and is an approach to be tested in the future.

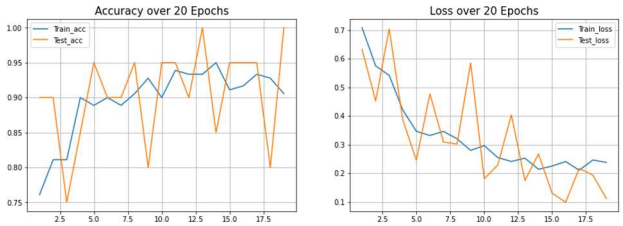
Appendix

Figure 1: Baseline Plot



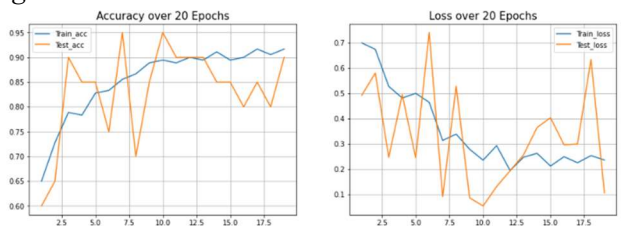
Training and validation loss and accuracy during Baseline training, Fold 8

Figure 2: Model 1 Plot



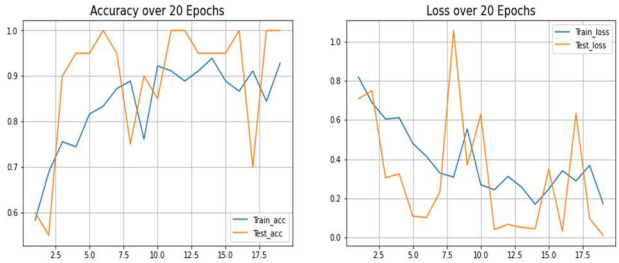
Training and validation loss and accuracy during Model 1 training, Fold 8

Figure 3: Model 2 Plot



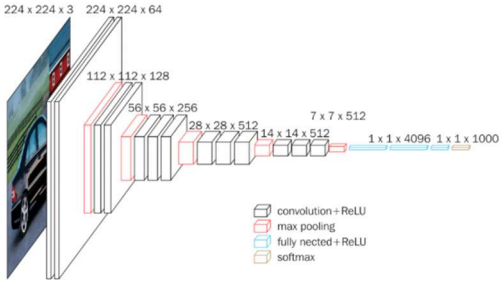
Training and validation loss and accuracy during Model 2 training, Fold 4

Figure 4: Model 3 Plot



Training and validation loss and accuracy during Model 3 training, Fold 6

Figure 5: VGG Architecture



Architecture of original VGG16 model designed for ImageNet dataset

Works Referenced

1. Xu, Adrian Yijie. "Detecting COVID-19 induced Pneumonia from Chest X-rays with Transfer Learning: An Implementation in Tensorflow and Keras." *Towards Data Science,* Medium, 21 March 2020.
2. Hurt, Brian et. al. "Deep Learning Localization of Pneumonia: 2019 Coronavirus (COVID-19) outbreak." *Journal of Thoracic Imaging*, Wolters Kluwer, 20 March 2020.
3. Abbas, Asmaa et. al. "Classification of COVID-19 in chest X-ray images using DeTraC deep convolutional neural network." *medRxiv,* 1 April 2020.