Report: Pneumonia Chest X-Ray Image Processing Model

Problem

Pneumonia is a type of infection that inflames the air sacs in one or both lungs. The air sacs can fill with liquid or pus. Infections can range from mild to severe and are caused by bacteria, viruses, and fungi. Bacterial pneumonia typically results in focal lobar consolidation. Viral pneumonia typically is more diffused and results in a pattern present in both lungs. Anyone can develop this infection but individuals with weaker immune systems are more susceptible. Symptoms vary but some common symptoms are: chest pain, cough, fatigue, fever, nausea and shortness of breath.

Pneumonia is diagnosed in a variety of ways. One common way is through the use of a chest X-ray to find the infection and the severity of it's spread. When interpreting the results, radiologists look for white spots in the lungs (called infiltrates) to identify the infection. Interpreting the results of chest X-rays requires careful observation and a good understanding of chest anatomy.

Due to the increase in chest X-ray volumes, significant backlogs are piling up and patients are having to wait an average of 30 days for results. HUP's lung center is asking for an image processing model to help classify chest X-rays of suspected pneumonia patients as either normal or pneumonia detected. Radiologists will further review X-rays where pneumonia was detected.

Data

Guangzhou Women and Children's Medical Center released over 5,000 chest X-ray images (anterior-posterior). The images come from retrospective cohorts of pediatric patients one to five years old. The images are labelled either pneumonia or normal. In addition, they were previously screened for quality control and split into train, test and validation sets.

Approach

The image files were read by folder using a custom load function to import images as 150 by 150 grayscale arrays. The image arrays were then labelled, combined by sets, and saved as dataframes. Through exploratory data analysis, 5856 total images were found. The arrays are of type uint8, meaning the pixels range from 0 to 255. Training data accounts for approximately 89% of the data, while testing is about 11% and validation is about 0.3%. There is a class imbalance in the data with the majority of images having detected pneumonia. The data was preprocessed using shuffling, equalization, rescaling, and corrected for class imbalance. Data augmentation was used to avoid overfitting due to the small dataset. The model was made using keras. Four convolution layers were used; along with max pooling, batch normalization, dropout and two dense layers. The model was optimized using ADAM, loss was measured using binary cross entropy, and six epochs were used. Model evaluation showed a F1 score around 93%, precision around 91% and loss around 25%.

The Model

The model was built using a Keras convolution neural network. Finding the appropriate layers to use was a challenge. Starting out using an out-of-box model resulted in low model performance. The training set was overfit. Through trial and error, the out-of-box model was found to be too complex for the problem at hand. Fewer layers, less drop outs, and lower number of filters resulted in higher evaluation results. In addition, adding weights to the images based on labels greatly boosted model performance. The final model used four convolution layers, two with 32 filters and two with 64 filters. For each number of filters, one convolution layer was traditional 2D and the other was separable 2D. To save computational space, max pooling was added between each convolution layer. Batch normalization was also added after each convolution layer. Two dense layers were also used. The first was activated using relu and had 128 units. The second was activated using sigmoid and had 1 unit. A dropout layer with 30% node loss, added after the first dense layer, helped to decrease overfitting. Losing any of the above mentioned layers significantly decreased model performance.

Constraints

The dataset was fairly small for image processing. Adding more X-rays could help to improve model performance. The data was also pre-split into train, test and validation sets. Validation data only contained 16 images. This resulted in low validation scores. Redistributing the data so more images were contained in the validation dataset might help improve performance. The data was also not labelled by severity of infection. Relabelling the data to include this feature might lead to increased performance, in particular detecting less severe cases. The data also only looked at pediatric patients. Pneumonia might differ in fully grown individuals. Adding adult patients to this dataset might improve model performance.

Further Research

This model can be used for a variety of future studies. To begin, the model can be expanded to not only detect pneumonia but also classify the severity of the infection or even the cause of the infection (viral, bacterial, fungal). Another use could be to expand the model to classify multiple types of chest X-ray diseases. If all chest X-rays were first passed through an image processing model, radiologists could spend more time focusing on irregular X-rays and patients could receive results within days. A third future use of this model could be to add image segmentation to research which areas of infection result in the most severe cases. Are focal infections more likely to be severe than diffused ones?

Summary

The model was created using a Keras convolutional neural network. Evaluations were maximized using a less complex model. The dataset was fairly small. To prevent overfitting, data augmentation was used. There was also a class imbalance that was corrected by weighing images by label type. The final model found a F1 score around 93%, precision around 91% and loss around 25%. Constraints include: the small dataset as a whole, the small number of images in the validation set, the data not being labelled for severity of infection, and only including pediatric X-rays. Correcting these constraints could improve model performance.