Pneumonia Classification Through X-Ray Imaging

Max Enderlein
University of Minnesota – College of Science and Engineering

Abstract

Healthcare is one of the most interesting fields to apply machine learning to. It can be implemented in many ways. For this project I chose to see if I could detect pneumonia from X-Ray images. The 2 models I chose to use for this problem was a random forest developed using Sci-Kit Learn and a convolutional neural network using PyTorch. My random forests were able to successfully detect pneumonia and had an accuracy of a little over 90%. My convolutional neural network had an accuracy of 86%.

Introduction

Being able to efficiently detect and diagnose illnesses is incredibly important in healthcare. Many different illnesses can have terrible impacts if not caught early enough. For this project I attempted to detect pneumonia from X-Ray images.

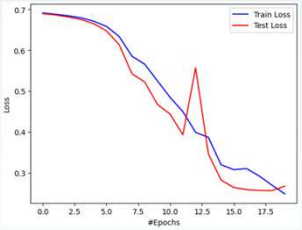
I used a dataset with around 6000 rows of data containing images of chest X-Rays and a binary label signifying if they have pneumonia or not. I then evaluated the performance of a random forest, testing different tree numbers, against a convolutional neural network.

Methodology

This project was done following the step-by-step process below:

I began with the random forest as I have more experience using them. I preprocessed the data using the helper functions from homework 4 (preprocess for sklearn) I then used a 75/25 split on the data and stratified for balance.
Next, I built the 3 random forest models the differences being the amount of trees. One had 100 trees, another had 500 trees, and the last had 1000 trees.
The 100 tree model did best. It had the highest accuracy and highest macro f1 score. Adding more trees did not improve the model's ability to detect pneumonia.
Overall random forests are fantastic at this task. They were easily able to detect pneumonia at a very high rate and impressed me in their accuracy of detection. Full results shown on the right.

The next model I created was a convolutional neural network. After a very long trial and error process I finalized my model with 3 convolutional layers followed by max pooling operations, for nonlinearity I used ReLU activation functions between each convolutional layer. The model was then trained using stochastic gradient descent with a learning rate of .01 over 20 epochs.



There was high class imbalance so I used a weighted cross entropy loss function to penalize misclassification more which was a major issue as before it would predict pneumonia every time. At the end I was able to get 86% accuracy and turned out to be a decent model, but there is still a lot of room for improvement.

The random forest model was better in this situation, but I am sure the CNN would be better with more resources and tuning.

Results

	precision	recall	f1-score	support
0	0.93	0.80	0.86	396
1	0.93	0.98	0.95	1068
accuracy			0.93	1464
macro avg	0.93	0.89	0.91	1464
weighted avg	0.93	0.93	0.93	1464

100 trees: 93% accuracy

	precision	recall	f1-score	support
0	0.93	0.78	0.85	396
1	0.92	0.98	0.95	1068
accuracy			0.92	1464
macro avg	0.93	0.88	0.90	1464
weighted avg	0.93	0.92	0.92	1464

500 trees: 92% accuracy

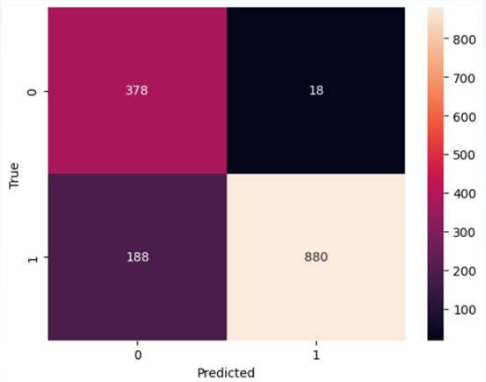
	precision	recall	f1-score	support
0	0.93	0.78	0.85	396
1	0.92	0.98	0.95	1068
accuracy			0.92	1464
macro avg	0.93	0.88	0.90	1464
weighted avg	0.92	0.92	0.92	1464

1000 trees: 92% accuracy

```
Train Set: Accuracy: 3770/4392 (85.8%)

Test Set: Accuracy: 1258/1464 (85.9%)
```

Convolutional neural network: 86% accuracy



Precision: .98 Recall: .82 F1 score: .89

Conclusion

Overall, both my models were very successful at detecting pneumonia from X-Ray images. The random forests were best at classifying pneumonia and the model with 100 trees was best out of the three with an accuracy of 93% and an F! score of .93. The convolutional network was not bad by any means as it still reached an accuracy of 86% and an f1 score of .89. Out of these two the random forest was better, but I still believe overall the convolutional neural network would be better with a little more time and resources in creating it. If I had more time, I would have a higher number of epochs as I have heard that 20 epochs is typically the bare minimum for a healthcare related machine learning task. I would also test around more with different numbers of layers and try some different activation functions as well. I was also curious on implementing ResNet-50 which has been used for this task in the past but due to time constraints and high runtimes causing my troubleshooting to be more dragged out I was not able to implement this. In the future I would like to extend this project and test my models against some proven models and see if I can come close in accuracy.

Final rankings of my created models:
1. Random forest (100 trees)
2. Random forest (500 trees)
3. Random forest (1000 trees)
4. My convolutional neural network.

Libraries

- SciKitLearn
- PyTorch
- MatplotLib
- SeaBorn
- NumPy
- Pandas
- PIL Image Processing

Acknowledgements

Professor Bianco Prado
Teaching assistants
College of Science and Engineering Staff
Hugging Face