Pneumonia Diagnosis via Chest X-ray Image Classification using Deep Learning and Transfer Learning

Akshat Gupta, Sanat Mishra, Zhen Yang

Introduction

Pneumonia is a respiratory infection caused by bacteria or viruses. Timely and accurate diagnosis of pneumonia plays a pivotal role in ensuring prompt therapeutic intervention, ultimately improving patient survival rates. Although chest X-ray imaging is a prevalent diagnostic tool, its interpretation relies on radiologists' subjective expertise, leading to variability. This challenge is amplified in regions with limited radiologist availability [1], hindering timely and accurate diagnosis. In response to these challenges, researchers have adopted deep learning to assist in pneumonia diagnosis through the analysis of X-ray images [2] [3]. Methods such as transfer learning train a neural network and show high accuracies [2]. By leveraging the capabilities of deep learning technology, we aim to contribute to advancing pneumonia diagnosis, particularly in regions facing a shortage of radiologists, and facilitate more effective patient care.

In our project, SZANet, a custom neural network, was developed and compared to the established pre-trained image classifier ResNet152V2 using chest X-ray data. Both methods achieved comparable accuracy in classifying images as 'healthy' or 'pneumonia,' showcasing transfer learning's ability to conserve resources by utilizing prior parameters.

Data

The project utilized chest X-ray images [4], as referenced by Alex McKeown et al. [2]. These images were collected from retrospective cohorts of pediatric patients aged one to five at the Guangzhou Women and Children's Medical Center, Guangzhou. To ensure quality, an initial screening process was conducted on all chest X-ray images, eliminating scans of low quality. Subsequently, two expert physicians labeled the images 'healthy' or 'pneumonia' classes [2]. Given that the images underwent rigorous quality control and were expertly labeled, we can utilize these labels to train our custom model and validate our results.

Our model processed 5,839 chest X-ray images, divided into 5,215 for training and 624 for testing. We further split the training data into 4,172 images for training and 1,043 for validation (80/20 split). Due to data imbalance (74% pneumonia, 26% healthy in training; 63% pneumonia, 37% healthy in testing), class weights were assigned to balance it: 1.94 for healthy and 0.67 for pneumonia. All 2D chest X-ray images, initially varying in size, were standardized to 224x224 pixels (Fig. 1a). Also, All the chest X-ray images are 2D matrices where entries represent the raw pixels of the image (Fig. 1b). To standardize the dimensions, we resize all images to a uniform size of 224x224, representing the smallest size among all the images.

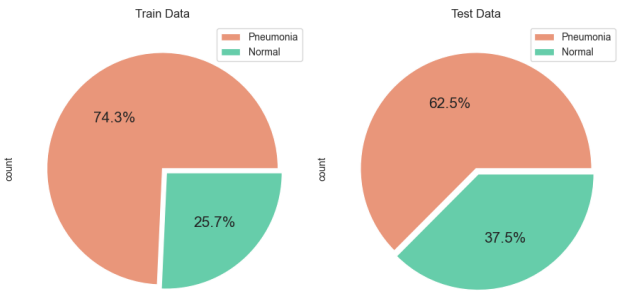

Figure 1a. Data Distribution

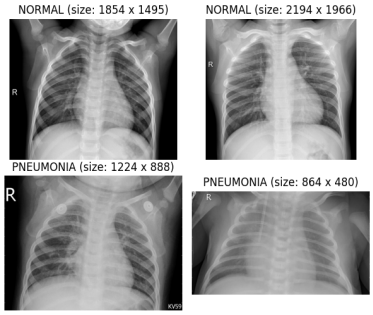

Figure 1b. Examples of X-ray Images

Finally, before feeding the input to train our model, we applied multiple data transformations to introduce variations in the input images during training, enabling our model to better generalize to new, unseen images. We initially normalized the image's pixel values, scaling them to a range between 0 and 1 by dividing each pixel value by 255. Following this, we applied shearing transformations with a maximum shear angle of 0.2 radians to the images. Next, we employed zoom transformations using a parameter of 0.2. Lastly, we applied horizontal flipping to the images.

Methods

Framework

We employed Keras and TensorFlow frameworks to leverage transfer learning and construct a custom deep learning model from scratch. We opted for Keras and TensorFlow owing to their comprehensive documentation, ease of use for building neural networks, and their robustness in handling complex computations. Additionally, we chose Python as our programming language due to its versatility and extensive libraries, further enhancing our capability to seamlessly integrate these frameworks into our development process.

Transfer Learning

Transfer learning is a prominent technique in deep learning and involves utilizing knowledge acquired by a model trained on one task to solve a related but distinct problem. The process starts by choosing specific layers from a pre-trained model that applies to the new task. These selected layers are frozen, maintaining their learned patterns and preventing changes during further training. New trainable layers are introduced on top of these frozen ones.

These new layers are designed to adjust the existing features to suit the prediction requirements of the new dataset.

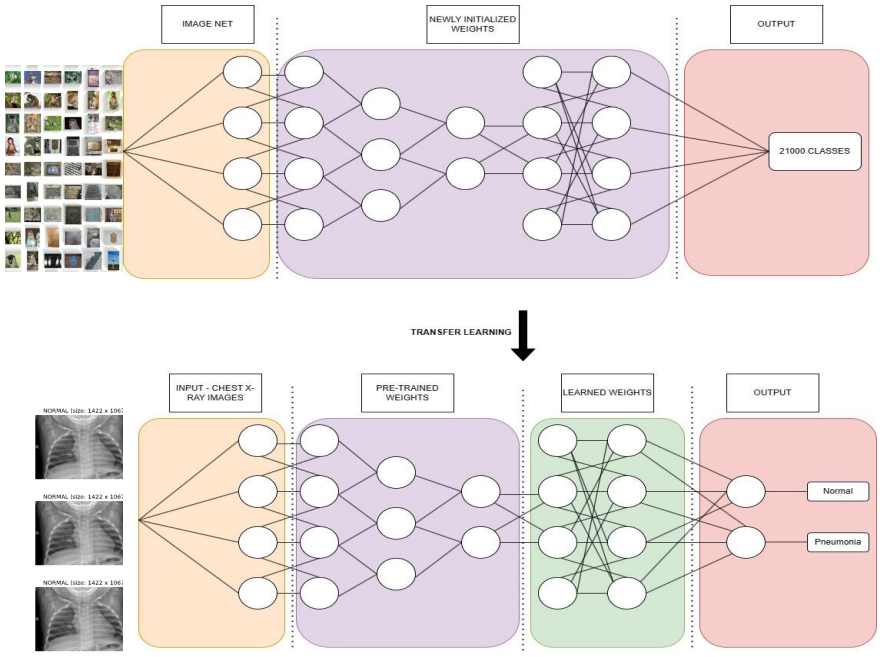

Figure 2: Transfer learning workflow for ResNet152v2

Training these added layers with the new dataset allows them to comprehend the extracted features and generate precise predictions tailored to the specifics of the new problem domain. Our model used the ResNet152v2 architecture, a pre-trained model on the ImageNet dataset, adapted for a specialized classification task. We

configured the input layer for images sized (224, 224, 3) and used the model primarily for feature extraction. Enhancements included global average pooling for spatial data condensation, a 128-unit dense layer with ReLU activation, and a dropout layer at a rate of 0.5 to prevent overfitting. The final layer, a single-unit with sigmoid activation, was designed for binary classification. These modifications were aimed at fine-tuning ResNet152V2 for our specific classification needs.

Deep Learning

We developed SZANet, a custom convolutional neural network for chest X-ray image analysis (size 224x224x3, representing RGB channels). Its architecture features three convolutional blocks: the first with 16 filters, the second with 32, and the third with 64 filters, all 3x3 in size. Each block includes 'valid' padding, batch normalization, ReLU activation, max pooling, and dropout layers (20% for the first two blocks, 40% for the third) to reduce overfitting. The network also contains a fully connected head with a flattened layer, a 64-unit dense layer with ReLU activation, and a 50% dropout rate. The final output layer is a single dense unit with a sigmoid function, ideal for binary classification tasks such as pneumonia detection.

Compilation and Training

Our model, designed for binary classification of X-ray images, used Binary Crossentropy as its loss function and the Adam optimizer with a learning rate 5e-5, benefiting from its adaptive learning rate capabilities. We measured performance using Binary Accuracy. During training, we implemented two key callbacks: Early Stopping, which halts training when validation loss ceases to improve and Reduce Learning Rate on Plateau, lowering the learning rate when validation loss stops improving. These were vital for training efficiency and preventing overfitting. The model was then trained on the dataset using these settings, aiming to enhance its accuracy in classifying chest X-ray images

Model Evaluation

Our model training employed an 80-20 Test-Train split and K-fold cross-validation. The 80-20 split divided the data into 80% for training and 20% for validation, offering continuous performance feedback and checking for overfitting. In contrast, K-fold cross-validation, with a k-value of 5, rotated training and testing across different data subsets, enhancing generalization. This approach allowed for comparing the model's performance under different data scenarios, providing insights into overfitting, generalization, and consistent accuracy.

Model Compression by Pruning

As a final step, we performed the process of model pruning. In pruning, we randomly set a fraction of model weights to be zero, and this fraction is called model sparsity. Then, we re-evaluate the model’s performance on the same validation data. By iterating over these fractions, we can evaluate the model on different sparsity values and see the model performance change. Trivially, we expect that as the model becomes more sparse, its performance decreases as crucial weights are set to zero. Pruning helps the model become lighter, makes inference faster, and helps check overfitting. In our implementation, we only set weights to zero and let the model architecture remain the same across every iteration. Finally, we evaluate the model on test data to get an accuracy.

Additionally, retraining the model once weights have been trimmed away is important. This helps the model adjust to accommodate fewer weights and presents a more efficient way to represent learned parameters.

Model Evaluation and Analysis

To assess our model's performance, we employed key metrics such as precision (the rate of correct positive predictions to total predicted positives), recall (sensitivity, measuring the correct identification of actual positives), and the F1 Score (a harmonized metric combining precision and recall for overall accuracy). Additionally, we used a confusion matrix to show the sensitivity, specificity and calculated model accuracy as the proportion of true results among all cases. Furthermore, ROC-AUC analysis was used, involving the ROC curve (plotting true positive rate against false positive rate) and AUC (measuring the model's class differentiation capability), with higher AUC indicating better performance.

Results

Our custom neural network, SZANet, achieved an ROC_AUC of 0.956 on the validation set. Employing transfer learning on the same dataset resulted in a commendable ROC_AUC of 0.947. Examining the confusion matrices (Fig 4), where 1 represents the Pneumonia class, and 0 represents the healthy class, the pre-trained model revealed 177 true positives (TP), 57 false positives (FP), 17 false negatives (FN), and 373 true negatives (TN). SZANet displayed 132 TP, 102 FP, 6 FN, and 384 TN, comparable to the pretrained model results. This observation is further supported by comparing other model validation metrics between SZANet and the pre-trained model (Fig 5).

While training SZANet, we noticed that the training accuracy was considerably higher than the test accuracy. We suspected that the model was overfitting. Despite trying out different hyperparameter combinations, we failed to observe any increase in test accuracy. To fix this, we decided to prune the model. This would make the model sparser and improve runtime. A summary of different pruning percentage values is included in Table 1 (Appendix).

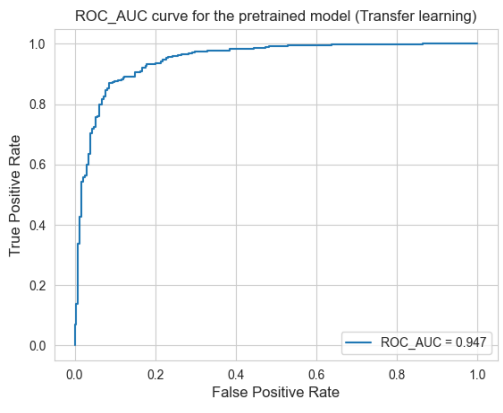

Figure 3a.

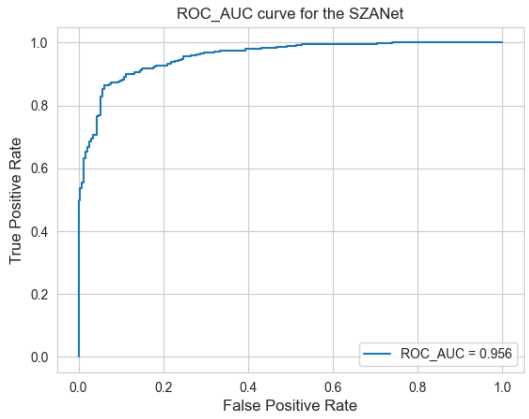

Figure 3b.

Figure 3: (a) ROC-AUC curve for ResNet152v2 (Transfer learning) without cross-validation (b) ROC-AUC curve for SZANet without cross-validation

Based on the results of pruning the model (Table 1, Appendix), it's evident that a 40% sparsity level maintains accuracy close to what our original model achieved. More pruning resulted in a major drop in accuracy, so that's why we have chosen 40% sparsity as the optimal value. To balance between the sparsity and accuracy, we calculate a 'score', that is, an average of model sparsity and accuracy; this helps us choose between the trade-off that comes from trying to control both accuracy and sparsity. The score is maximum at 40% sparsity at a value of 0.62, reiterating our choice for 40% sparsity.

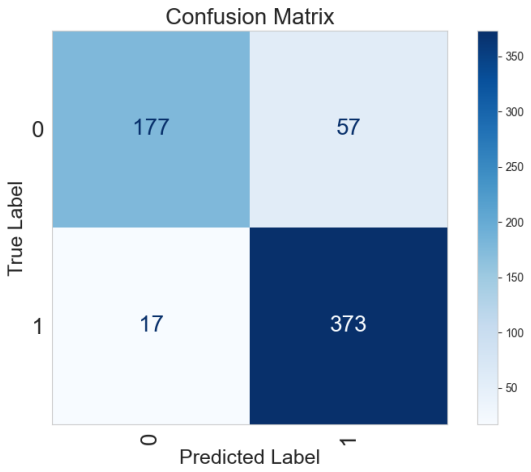

Figure 4a.

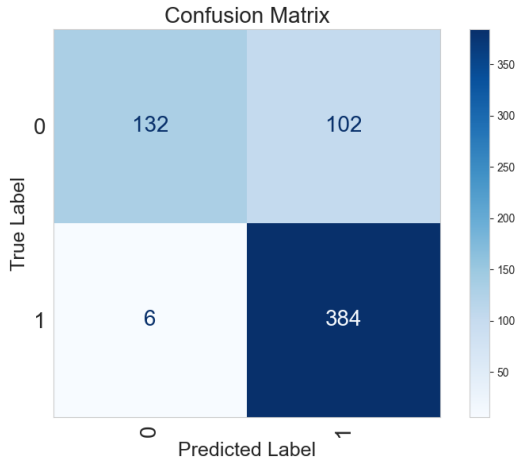

Figure 4b.

Figure 4: (a) Confusion matrix for ResNet152v2 (Transfer learning) without cross-validation (b) Confusion matrix for SZANet without cross-validation

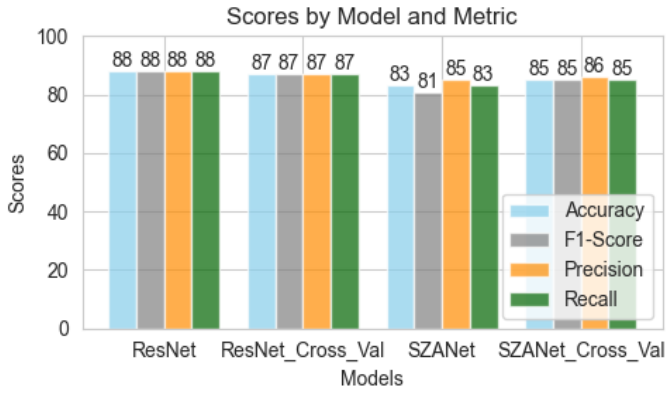

Figure 5: Comparison of metrics across all configurations of ResNet152v2 and SZANet

Conclusion

In conclusion, our results show that transfer learning is a comparable approach to classifying X-ray images for pneumonia classification to the bespoke, pruned SZANet. This shows that pre-trained models do not lose out on accuracy. This was in line with our expectations since ResNet 152V2 is seen to perform well on several biomedical image classification tasks. For both strategies of model evaluation, K-Fold cross-validation and test-train split, our pre-trained model and custom model agreed in accuracy, precision, F1, and recall.

References

[1] Oates, A., K. Halliday, A.C. Offiah, C. Landes, N. Stoodley, A. Jeanes, K. Johnson, et al. "Shortage of Paediatric Radiologists Acting as an Expert Witness: Position Statement from the British Society of Paediatric Radiology (BSPR) National Working Group on Imaging in Suspected Physical Abuse (SPA)." *Clinical Radiology* 74, no. 7 (2019): 496–502. https://doi.org/10.1016/j.crad.2019.04.016.

[2] Kermany, Daniel S., Michael Goldbaum, Wenjia Cai, Carolina C.S. Valentim, Huiying Liang, Sally L. Baxter, Alex McKeown, et al. "Identifying Medical Diagnoses and Treatable Diseases by Image-Based Deep Learning." *Cell* 172, no. 5 (2018). https://doi.org/10.1016/j.cell.2018.02.010.

[3] Kundu R, Das R, Geem ZW, Han G-T, Sarkar R (2021) Pneumonia detection in chest X-ray images using an ensemble of deep learning models. PLoS ONE 16(9): e0256630. https://doi.org/10.1371/journal.pone.0256630

[4] Kermany, D., Zhang, K., & Goldbaum, M. (2018). Labeled Optical Coherence Tomography (OCT) and Chest X-Ray Images for Classification [Data set]. Mendeley Data, V2. https://doi.org/10.17632/rscbjbr9sj.2

APPENDIX

Pruning Percentage	Sparsity	Accuracy	Score (Sparsity+Accuracy/2)
30	0.3	0.88	0.59
40	0.4	0.86	0.62
50	0.5	0.63	0.56
70	0.7	0.375	0.54
80	0.8	0.375	0.59

Table A1: Model pruning percentage and accuracy. The values of sparsity, accuracy, and score with successive pruning.

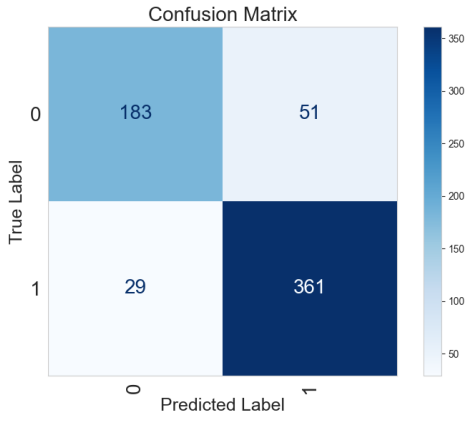

Figure A1a.

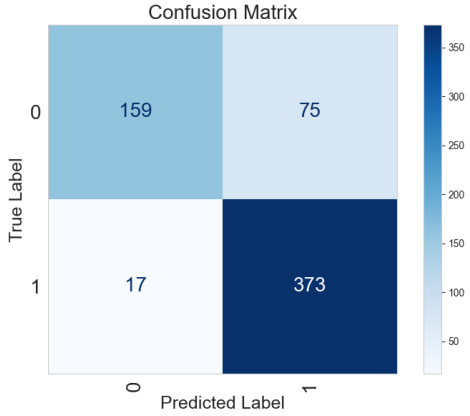

Figure A1b.

Figure A1: (a) Confusion matrix for ResNet152v2 (Transfer learning) with cross-validation (b) Confusion matrix for SZANet with cross-validation

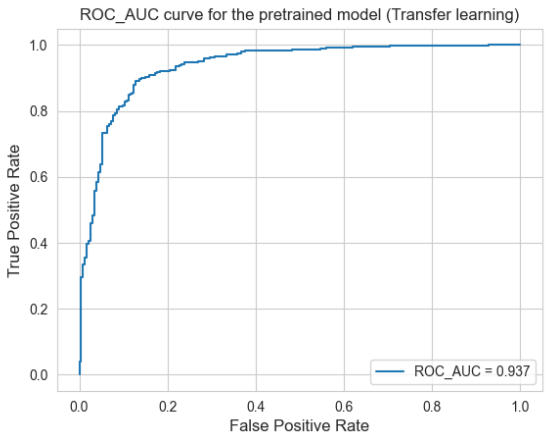

Figure A2a.

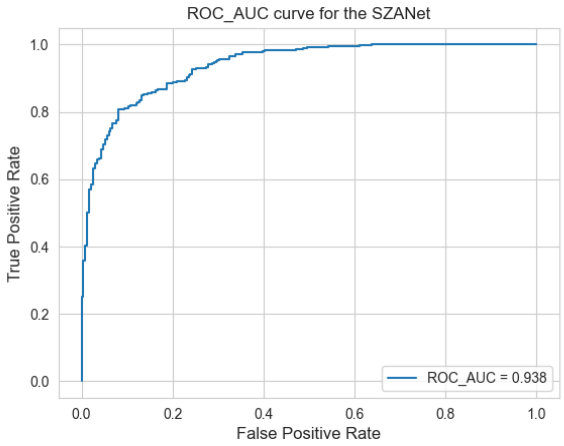

Figure A2b.

Figure A2: (a) ROC_AUC for ResNet152v2 (Transfer learning) with cross-validation (b) ROC_AUC for SZANet with cross-validation