A Comparison of Pre-Trained vs. Randomly Initialized ResNet18 Model Performance in Pneumonia Detection

Annalisa Vitulli

Department of Computer Science

University of Central Florida

Annalisa.Vitulli@ucf.edu

1. Abstract

The ResNet-18 architecture was adapted for binary classification of pneumonia x-rays, and two models—one trained from scratch and one with pre-trained weights—were tested using the newly developed network. Their results were reported using validation and accuracy loss curves and analyzed for the best performance.

2. Method

The adapted ResNet-18 model is proposed for binary classification of x-ray images. Two versions of the model are implemented to compare their performance for pneumonia detection: Model 1, which was initialized with random weights and trained from scratch, and Model 2, which was pre-trained and fine-tuned.

2.1. Augmentation of the ResNet-18 Architecture

Binary Classification Layer. The standard fully connected output layer with 1000 neurons in ResNet-18 was replaced with a custom output layer for binary classification. This binary classification layer contained:

- **Dropout layers:** Two dropout layers were added ($dropout\_prob = 0.5$ per layer).
- **Linear layers:** Two linear layers were implemented. The first linear layer reduced output features to half of the input size, and the second linear layer reduced output features to a single value for binary classification.
- **ReLU:** The ReLU activation function was added after the first linear layer.

2.2. Data Augmentation

The following data augmentation techniques were applied to the training data. Images were randomly cropped to between 80-120% of the original image's area and then resized to 224x224 pixels to match ResNet-18's input size. ColorJitter was applied with a brightness of 3 and a contrast of .4 to simulate varying x-ray image qualities, and both RandomAffine (10° rotation, 10° sheer) and RandomHorizontalFlip were applied. Lastly, the data was normalized to ImageNet standards and converted to a tensor. Validation and test data were solely resized to 224 x 224, normalized to ImageNet standards, and converted to a tensor.

2.3. Implementation Details

Dataset. The models were trained and tested on the Chest X-Ray Images (Pneumonia) [1] dataset, which contains a combination of 5,863 anterior and posterior chest x-rays—1,583 normal and 4,273 pneumonia—of pediatric patients between the ages of one to five years old. The pneumonia cases included both bacterial and viral pneumonia which present different visual characteristics (i.e., a localized lobar pattern vs. a dispersed bilateral lung presentation, respectively).

Training. Model 1 and Model 2 both utilized the same previously discussed modified ResNet-18 architecture. However, Model 1 was initialized with random weights during the training process while Model 2 loaded pre-trained weights from ImageNet.

Hyperparameters. Model 1 was trained for 30 epochs while Model 2 was fine-tuned for only 15 epochs. The following hyperparameters were used for both Model 1 and Model 2:

- **Batch Size:** 32.
- **Learning Rate (LR):** A LR of 0.003.
- **Regularization:** A weight decay of 0.001.
- **Optimizer:** The Adam optimizer.
- **Scheduler:** StepLR with a step size of 7 and gamma of 0.1 to reduce the LR periodically.
- **Loss Function:** Binary Cross Entropy with Logits and a positive class weight to account for the dataset's class imbalance using the following calculation:

$$pos\_weight = \frac{\#\ pneumonia\ cases}{\#\ normal\ cases} \quad (1)$$

3. Results

Metrics. Training and validation loss and training and validation accuracy were tracked for both models, and their values are displayed via loss curves in Figs.

1-4. The overall testing accuracy as well as the testing accuracy for each class is documented for Model 1 and Model 2.

3.1. Training and Validation Loss

Model 1's training loss decreases steadily over the 30 epochs, starting at 0.712 and ending at 0.296. However, as shown in Fig. 1, validation loss is sporadic and does not exhibit any noticeable trend. In Fig. 2, Model 1's training accuracy improves steadily, ending at 91.6% while the validation accuracy remains low at 68.75% resulting in a 22.85% difference between training and validation accuracy. For Model 2, Fig. 3 shows a fairly steady decrease in training loss over the 15 epochs, starting at 0.643 and ending at 0.465. However, while initially erratic through epoch 6, validation loss begins to trend upward at epoch 7, beginning at 0.503 and ending at 0.593. Although the training accuracy only reached 83.47% by epoch 15, Model 2's validation accuracy also reliably achieved 75% producing only an 8.47% difference between training and validation accuracy as shown in Fig. 4.

Model 1's significant difference in training and validation accuracy along with the high validation loss suggests overfitting. Since Model 1 was training from scratch, the unbalanced dataset (i.e., significantly greater pneumonia images in comparison to normal images) could be the cause.

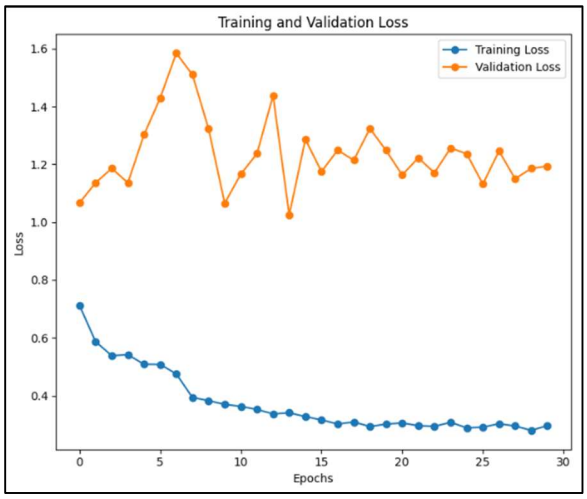

Figure 1. Model 1's Training and Validation Loss Curve

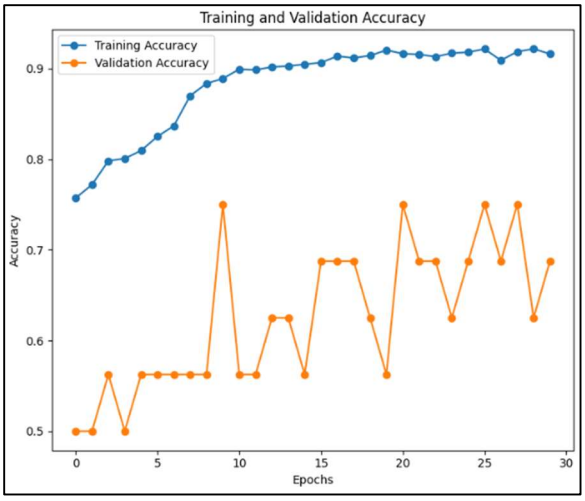

Figure 2. Model 1's Training and Validation Accuracy Curve

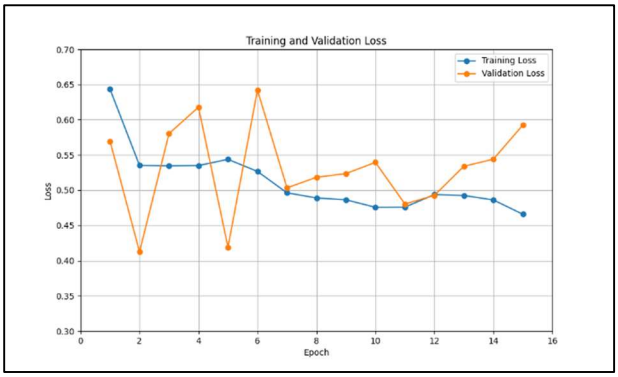

Figure 3. Model 2's Training and Validation Loss Curve

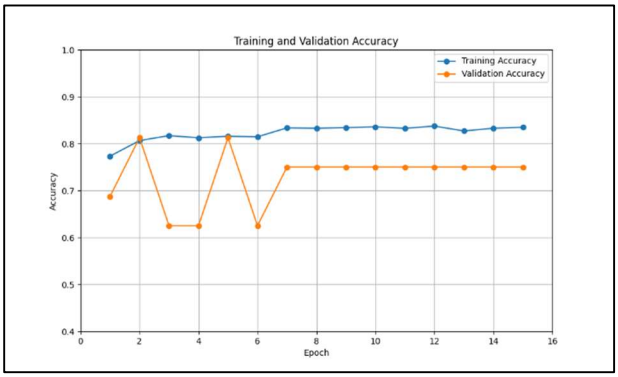

Figure 4. Model 2's Training and Validation Accuracy Curve

3.2. Classification Accuracy

Overall Accuracy. The overall test accuracy for Model 1 and Model 2 was 89.26% and 86.06%, respectively. Both models exhibited extremely similar testing loss values at 0.407 for Model 1 and 0.405 for Model 2.

Normal Class Accuracy. Model 1 identified "Normal" x-rays at a significantly higher accuracy than Model 2 (i.e., 75.64% vs. 65.38%). This difference in "Normal" class performance was largely responsibly for Model 1's slightly improved overall test accuracy. Fig. 5 displays a confusion matrix of Model 2's testing results, and although only 6 false negatives occurred, there were 153 false positives. Image 1 and Image 2 display false positive images from Model 1 and Model 2, accordingly.

Pneumonia Class Accuracy. Model 2 surpassed Model 1 marginally in pneumonia detection with a class accuracy of 98.46% vs. 97.44%. False negative results from Model 1 and Model 2 are shown in Image 3 and Image 4.

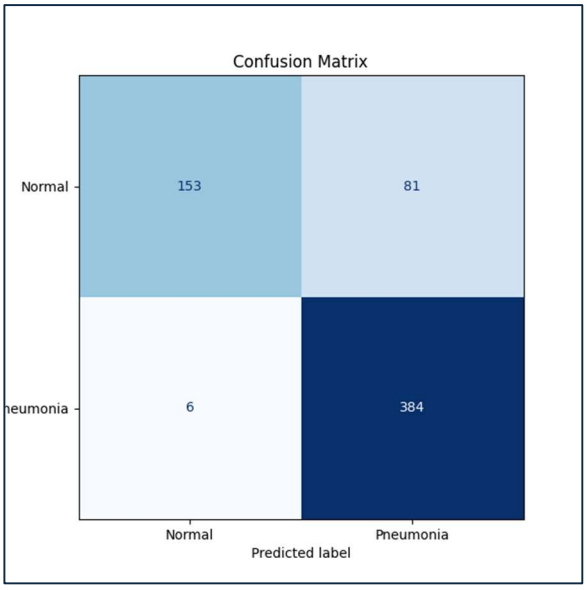

Figure 5. Model 2 – Confusion Matrix for Pneumonia Detection

Figure 1. Model 1 – Misclassified Image (False Positive)

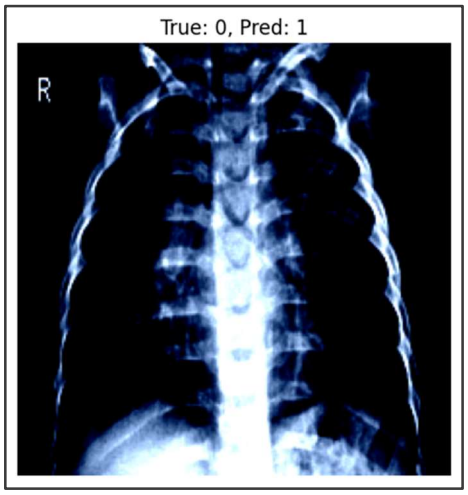

Figure 2. Model 2 – Misclassified Image (False Positive)

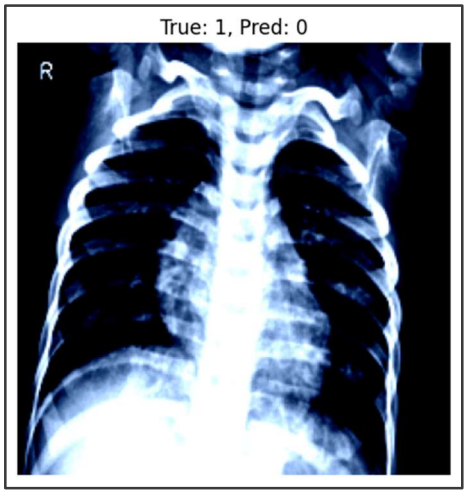

Figure 3. Model 1 – Misclassified Image (False Negative)

Figure 4. Model 2 – Misclassified Image (False Positive)

4. Conclusions and Future Improvements

Although dropout and a class weight parameter were both implemented to counteract overfitting, both Model 1 and Model 2 struggled with this issue. Overall, Model 1 exhibited better performance on the testing data as shown by the classification accuracies, but the model still produced an abnormal number of false positives, which could result from the relatively small dataset with limited "Normal" x-ray images to train on.

In the future, to correct the class imbalance issue, SMOTE could be implemented to augment normal images. Additionally, the value that pneumonia cases are being penalized at in the pos_weight equation could be reduced. However, since Model 1 already struggled to identify pneumonia cases in comparison to Model 2, reducing this penalty is not preferable since this could increase false negatives.

References

[1] Daniel S. Kermany, et al. Identifying Medical Diagnoses and Treatable Diseases by Image-Based Deep Learning, Cell, Volume 172, Issue 5, 1122 – 1131.e9.