Final Project - Deep Learning Course - 2025-Semester B

Adi Levi Malach (ID. 312230667)

August 16, 2025

1 Introduction

This report details the development and comparison of two deep learning architectures, a Vision Transformer (ViT) and a Convolutional Neural Network (CNN), for the binary classification of chest X-ray images into 'Normal' or 'Pneumonia' classes. Leveraging pre-trained models and a systematic fine-tuning process, we aimed to develop robust classifiers for this challenging and imbalanced medical imaging dataset. This work follows the project guidelines which suggest comparing ViT and CNN architectures, referencing foundational papers such as Dosovitskiy et al. [2020] and literature on training improvements like **?**.

1.1 Dataset

The public "Chest X-Ray Images (Pneumonia)" dataset from Kaggle was used. After performing an 80/20 stratified split of the original training data to create a validation set, the final dataset distribution was as follows:

- **Training Set:** 4,172 images (1,073 Normal, 3,099 Pneumonia)
- **Validation Set:** 1,044 images (268 Normal, 776 Pneumonia)
- **Test Set:** 624 images (234 Normal, 390 Pneumonia)

2 Methodology

Our approach focused on transfer learning to adapt powerful, pre-trained architectures to this specialized task.

2.1 Models Architecture

2.1.1 Vision Transformer (ViT)

The ViT model is based on the ViT-Base/16 architecture from the original paper by Dosovitskiy et al. [2020]. The model, pre-trained on ImageNet-21k, was instantiated using the ‘timm‘ library. For fine-tuning, the original classification head was replaced with a new linear layer for our 2-class problem, and the first three transformer blocks of the encoder were frozen to preserve the learned low-level feature representations.

2.1.2 Convolutional Neural Network (CNN)

The CNN model is based on the ResNet-18 architecture, loaded with pre-trained ImageNet1k weights using ‘torchvision.models‘. The final model utilized a full fine-tuning approach where all layers were trainable. The final fully-connected layer was replaced with a new classification head consisting of a ‘Dropout‘ layer for regularization followed by a ‘Linear‘ layer for our binary task.

2.2 Data Preprocessing & Augmentation

To enhance model generalization, a series of data augmentations were applied to the training images, including ‘RandomResizedCrop‘, ‘RandomHorizontalFlip‘, and ‘RandomRotation‘. Validation and test images were processed using a standard ‘Resize‘ and ‘CenterCrop‘ pipeline to ensure consistency. All images were resized to 224x224 and normalized using ImageNet statistics.

2.3 Training Strategy

- **Loss Function:** Due to the significant class imbalance, a standard ‘CrossEntropyLoss‘ function was used with class weighting to assign a higher penalty to errors made on the minority ’Normal’ class.

- **Optimizer & Scheduler:** The ‘AdamW‘ optimizer was used for both models. A ‘CosineAnnealingLR‘ scheduler with a linear warmup phase was implemented to stabilize training.

- **Hardware & Platform:** All experiments were performed on a single NVIDIA T4 GPU via Google Colab. The final ViT model trained in approximately 8 minutes, while the final CNN model trained in approximately 2.5 minutes.

2.4 Hyperparameters

The key hyperparameters for the best-performing version of each model are summarized in Tables 1 and 2.

3 Experimental Results

The final model was always selected based on the epoch with the lowest validation loss to prevent overfitting and improve generalization.

Table 1: Final ViT Hyperparameters

Parameter	Value
Learning Rate	1e-4
Batch Size	32
Weight Decay	0.3
Max Epochs	12
Warmup Epochs	5

Table 2: Final CNN Hyperparameters

Parameter	Value
Learning Rate	1e-4
Batch Size	64
Weight Decay	0.1
Dropout Rate	0.5
Max Epochs	15
Warmup Epochs	3

3.1 ViT Model Results

The ViT's training was characterized by rapid convergence but unstable validation curves. The model selection strategy successfully navigated this instability to find the optimal checkpoint at epoch 4.

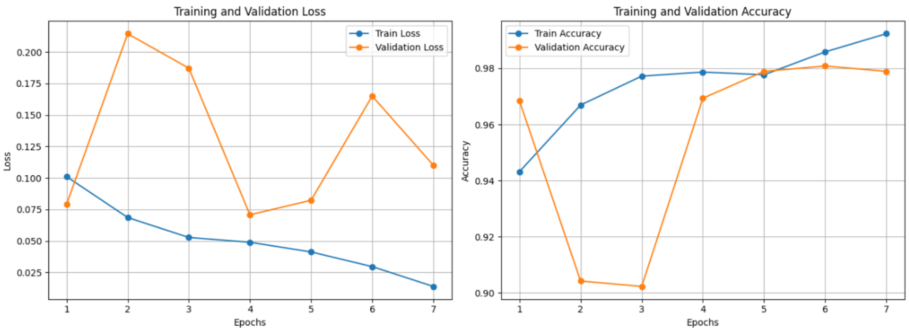

Figure 1: Training & Validation curves for the ViT model.

The final test results demonstrate a strong and well-balanced classifier.

Clinical Interpretation: The final confusion matrix was [TN: 176, FP: 58], [FN: 4, TP: 386]. The model generated only 4 False Negatives, correctly minimizing the most dangerous error of missing a pneumonia case.

Table 3: ViT Per-Class Test Metrics

Class	Precision	Recall	F1-Score
NORMAL (Class 0)	97.78%	75.21%	85.02%
PNEUMONIA (Class 1)	86.94%	98.97%	92.57%

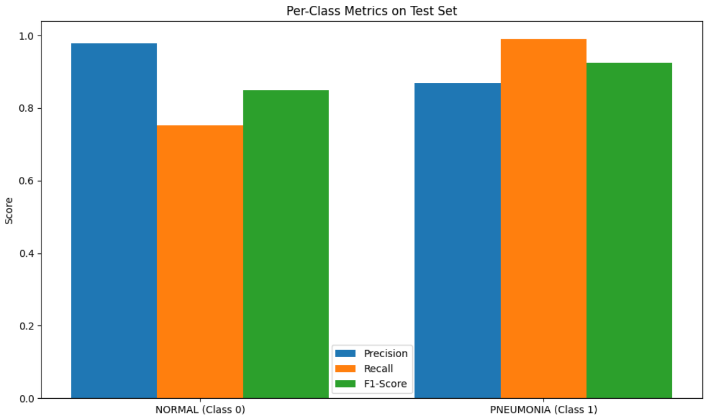

Figure 2: Final per-class metrics for the optimized ViT model on the test set.

3.2 CNN Model Results

After iterative tuning of regularization parameters (higher `weight_decay` and `Dropout`), the CNN's training process became much more stable.

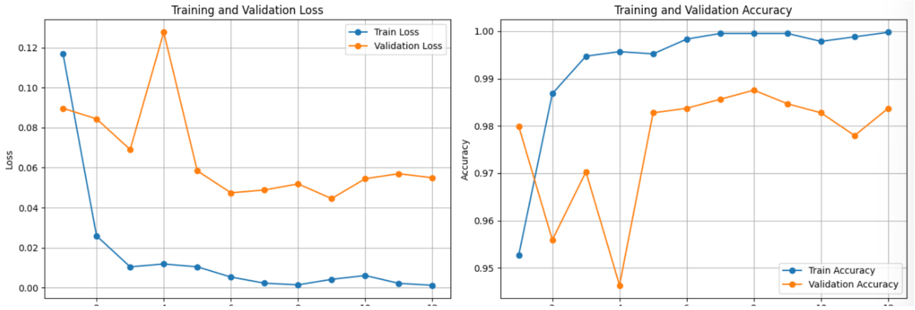

Figure 3: Training & Validation curves for the final CNN model.

The final results for the optimized CNN model were highly competitive.

Clinical Interpretation: The derived confusion matrix was [TN: 171, FP: 63], [FN: 3, TP: 387]. The CNN was even safer in detecting pneumonia, with only 3 False Negatives.

Table 4: CNN Per-Class Test Metrics

Class	**Precision**	**Recall**	**F1-Score**
NORMAL (Class 0)	98.28%	73.08%	83.82%
PNEUMONIA (Class 1)	86.00%	99.23%	92.14%

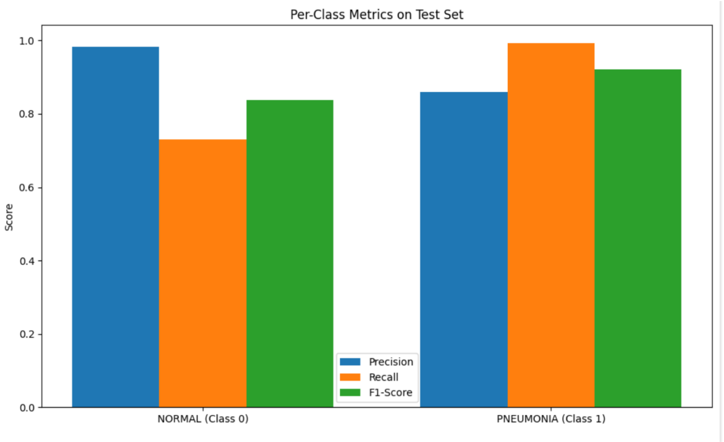

Figure 4: Final per-class metrics for the optimized CNN model on the test set.

4 Discussion: Comparative Analysis

The primary goal of this project was to compare the effectiveness of the ViT and CNN architectures on this task.

Table 5: Final Performance Comparison: ViT vs. CNN

Metric	**ViT Model**	**CNN Model**
Accuracy	**90.06%**	89.42%
Recall (Normal Class)	**75.21%**	73.08%
F1-Score (Normal Class)	**85.02%**	83.82%

The analysis (Table 5) clearly shows that the **Vision Transformer is the better-performing model**. It achieved higher overall accuracy and, most critically, a higher Recall and F1-Score for the underrepresented 'Normal' class. This indicates that ViT produced a more balanced and generalizable classifier for this specific problem. Although the CNN was faster to train and exhibited more stable training curves, the ViT's final predictive performance was ultimately superior.

5 Code Repository

The complete code for this project, implemented in a Jupyter Notebook, is available at the following link:

`https://github.com/your-username/your-repo`

References

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale, 2020.