Pneumonia Detection in Chest X-rays using Vision Transformers

Josh Burgess
School of Engineering and Computer Science Victoria University of Wellington
jgburgess173@gmail.com

Abstract

This project compares Vision Transformers (ViTs) against Convolutional Neural Networks (CNNs) for pneumonia detection in chest X-ray images. Using a dataset of 5,863 chest X-ray images from Kaggle, we will fine-tune pre-trained ViT and ResNet-50 models to evaluate their relative performance on medical image classification. The research aims to determine whether ViT's attention mechanisms provide an advantage over CNNs for identifying pneumonia patterns in X-ray images. Evaluation will include accuracy, precision, recall and F1-score metrics. A particular focus will be on the implications of false positives and negatives in medical diagnosis.

Keywords*—vision transformers, convolutional neural networks, pneumonia detection, medical imaging, chest X-rays*

I. Introduction

Pneumonia is one of the leading causes of death globally [1], with chest X-ray interpretation showing 22-26% disagreement rates between radiologists [2]. This inconsistency creates diagnostic delays and workflow bottlenecks, highlighting the need for automated diagnostic assistance. While Convolutional Neural Networks (CNNs) are the current standard for medical imaging, Vision Transformers (ViTs) offer a fundamentally different approach that may be better suited for pneumonia detection.

CNNs analyse images by applying kernels to small local regions, progressively combining these local features into increasingly complex patterns through multiple layers. ViTs employ self-attention mechanisms to capture global relationships between image patches from the outset.

This project looks into whether Vision Transformers can outperform traditional CNNs for pneumonia detection in chest X-ray images by comparing pre-trained ViT-Base and ResNet-50 models fine-tuned on medical data.

II. Related Work

Deep CNNs have established themselves as the standard for pneumonia detection in chest X-ray images with ensemble approaches achieving accuracies of 97.2% and recall rates exceeding 99% using transfer learning and attention mechanisms [3]. Recent studies have begun exploring ViTs as a potential alternative with Singh et al. achieving 97.61% accuracy and demonstrating ViT's ability to capture global context and spatial relationships [4].

III. Methodology

A. Dataset

We will use the Kaggle chest X-ray pneumonia dataset containing 5,863 front view images with binary classification labels, normal vs pneumonia [5]. This established dataset provides sufficient variety in pneumonia presentations and normal cases to enable robust training model and evaluation. Images will be pre-processed to standard dimensions and normalised for consistency across both models.

B. Data Preprocessing

Images will be pre-processed from their original variable dimensions to 224×224 pixels using aspect ratio preserving resize followed by border padding to maintain proportions while meeting model input requirements. This approach preserves the complete lung fields and chest anatomy, enabling ViT's to use their global attention mechanism across the whole image.

Standard medical imaging augmentations will be applied during training including random rotation (±10°), brightness adjustment (±20°), and contrast enhancement to improve model generalisation and prevent overfitting [6]. These augmentations simulate natural variations in X-ray conditions while preserving clinically relevant conditions.

C. Model Selection

For ViT we will use ViT-Base pre-trained on ImageNet-21k, it will serve as our transformer baseline. This model processes 224×224 images as sequences of 16×16 with 12 transformer layers and 768 hidden dimensions.

For CNN we will use ResNet-50 pre-trained on ImageNet which will represent the CNN baseline. This model uses residual connections and has demonstrated strong performance in medical imaging applications.

Both models represent the most widely used and studied architecture in their respective categories, providing appropriate range for this project.

D. Training Strategy

Both models will be fine-tuned using their pre-trained weights, with the final classification layer altered for binary pneumonia detection. This approach uses learned visual representations to adapt to medical imagery.

We will begin with a subset of 500 images to validate the training pipeline approach before scaling to the full dataset. This ensures robust methodology and enables early identification of potential issues.

Standard medical imaging augmentations and contrast adjustment will be applied to improve generalisation and prevent overfitting.

E. Evaluation Framework

Models will be assessed using accuracy, precision, recall and F1-score. Particular emphasis will be placed on recall to minimise false negatives, as missing pneumonia cases has more severe clinical consequences than false positives.

ViT attention maps and CNN activation maps will be generated to visualise model decision making processes and evaluate clinical relevance of the learned features.

Multiple training runs will be conducted to ensure reproducible results, with proper train, validation and test splits to prevent data leakages.

IV. Expected Results

A. Hypothesis

We hypothesise that ViTs will demonstrate higher performance for pneumonia detection, especially for cases with distributed pathological patterns. The global attention mechanism should enable ViTs to better understand spatial relationships between distant lung regions compared to CNN's hierarchal local to global processing.

B. Performance Predictions

ViTs are expected to achieve higher F1-scores, especially for complex pneumonia presentations where consolidations appear across multiple lung zones. The improvement may be slight but clinically significant for reducing missed diagnoses.

The advantage should be most noticeable for multifocal and diffuse pneumonia patterns, where spatial relationships between scattered infections are crucial for accurate diagnosis. For simple lobar consolidations, performance differences may be minimal.

C. Interpretability Analysis

ViT attention maps should highlight clinically relevant regions and demonstrate coherent focus on pathological areas. The attention patterns may show how the model integrates information across different lung regions.

CNN activation maps are likely to show more localised feature detection, while ViT attention should demonstrate broader spatial integration, potentially aligning better with radiological patterns.

D. Clinical Relevance

If successful this project will demonstrate that architectural choices significantly impact medical AI performance. The findings could inform future medical imaging system design and contribute to more reliable automated diagnostic tools.

V. Statement

This project will use Python as the primary programming language with PyTorch for deep learning implementation. The Hugging Face Transformers library will be used for ViT models, while ResNet-50 will be implemented through Pytorch's torchvision package. Additional libraries include scikit-learn for evaluation metrics, matplotlib and seaborn for visualisation, and NumPy for data manipulation. Training will be conducted on Victoria University's ECS GPU resources. The Kaggle chest X-ray pneumonia dataset will be used under its public license for academic research purposes.

References

[1] World Health Organization, "The top 10 causes of death," WHO Fact Sheet, 2020. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/the-top-10-causes-of-death

[2] M. A. Elemraid, M. Muller, D. A. Spencer, S. P. Rushton, R. Gorton, M. F. Thomas, K. M. Eastham, F. Hampton, A. R. Gennery, and J. E. Clark, "Accuracy of the Interpretation of Chest Radiographs for the Diagnosis of Paediatric Pneumonia," PLoS One, vol. 9, no. 8, pp. e106051, Aug. 2014. https://pmc.ncbi.nlm.nih.gov/articles/PMC4141860/

[3] Q. An, W. Chen, and W. Sha, "A Deep Convolutional Neural Network for Pneumonia Detection in X-ray Images with Attention Ensemble," Diagnostics, vol. 14, no. 4, pp. 390, Feb. 2024. https://www.mdpi.com/2075-4418/14/4/390

[4] A. Singh et al., "Efficient pneumonia detection using Vision Transformers on chest X-rays," Sci Rep, vol. 14, no. 2024, Jan. 2024. https://www.nature.com/articles/s41598-024-52703-2

[5] Paul Moony "Chest X-ray images (Pneumonia)," Kaggle, 2018. https://www.kaggle.com/datasets/paultimothymooney/chest-xray-pneumonia/data

[6] A. A. Nasser and M. A. Akhloufi, "A Review of Recent Advances in Deep Learning Models for Chest Disease Detection Using Radiography," Diagnostics (Basel) , vol. 13, no. 1, pp. 159, Jan. 2023. https://pmc.ncbi.nlm.nih.gov/articles/PMC9818166/