Design and Implementation of a Vision Transformer for Pneumonia Detection

T Vikash Patra
Department of Computing Technologies, School of computing
SRM Institute of Science and Technology
Kattankulathur, Chennai
tt7597@srmist.edu.in

NSS Pavan Kumar
Department of Computing Technologies, School of computing
SRM Institute of Science and Technology
Kattankulathur, Chennai
nn5610@srmist.edu.in

Vetriselvi D
Assistant Professor
Department of Computing Technologies, School of computing
SRM Institute of Science and Technology
Kattankulathur, Chennai
vetrised@srmist.edu.in

1.Abstract

Pneumonia remains one of the most burdensome diseases within the population, as well as one of the most frequent causes of death, thus necessitating prompt diagnosis to reduce fatalities and improve the patients' prognosis. In conventional diagnostic techniques, such as a physical examination followed by a chest x ray, interpretation of these images is usually manual and is therefore prone to errors and quite slow, particularly in places where radiologists are few or unaffordable. In recent years, a significant focus has been placed on deep learning-based methods, notably Convolutional Neural Networks (CNN) have shown promising results in the automated identification of pneumonia in medical images. Nonetheless, a limitation of CNN models is that their local receptive fields prevent them from effectively capturing long-range dependencies within the image. This paper presents an innovative use of Vision Transformer (ViT), which was originally a model for natural image classification, for detecting pneumonia using chest X-ray images. Vision transformer (ViT) architecture primarily utilizes the self-attention mechanism, that helps in context-aware representation of features of the image, given that in medical image analysis most of the time the abnormal patterns are often spread throughout the image, quite far from each other. For this purpose, in this work, we trained a ViT model on public cough chest X-ray database.

2.Introduction

Infections of the lungs due to bacteria, viruses, and fungi, known as pneumonia, are still one of the greatest problems that the world faces due to the associated morbidity and mortalityAccording to World Health Organization (WHO) statistics, pneumonia is among the top causes of death in children under five, responsible for 14% of fatalities, highlighting the importance of early and accurate diagnosis. Chest X-ray films are the primary imaging method for diagnosing pneumonia; however, interpreting these films can be complex and time-consuming, often depending on the expertise of experienced radiologists.. In many low-resource settings of the world, where there are very few if not any radiologists, this further translates to wastage of time in diagnosis and subsequently high mortality rates. For this reason, there is a pressing demand for assistive diagnostic tools that are quick and dependable, particularly in regions with little health care provision.

We recommend the application of the Vision Transformer, a model designed for image classification purposes that utilizes a transformer based architecture. A self-attention network such as the Vision Transformer does not have layers in the framework that define distance and position between objects in an image, as is the case with Convolutional Neural Networks, instead images are presented as a whole. It is this trait that gives an upper hand to ViT since it can process an entire image even though it can also divide the image into sections referred to as patches which are treated as words in a language task. This is suitable for image-based processes that require a greater level of context as well as additional details. This work seeks to evaluate Vision Transformer and some of the modern techniques based on convolutional neural networks for chest X-ray images in detecting pneumonia, and provide comparisons of the results obtained. We set out to demonstrate this by employing the ViT self-attention mechanism and proving that it is capable of modelling complex structures of the medical images and thus improving performance of the task compared to traditional CNN model. Based on the results, we believe that Vision Transformers are very valuable in aiding the diagnostic processes in systems that classify medical images automatically.

3.Problem Statement

Pneumonia remains a significant global health concern, highlighting the importance of prompt and accurate diagnosis. The traditional method of interpreting chest

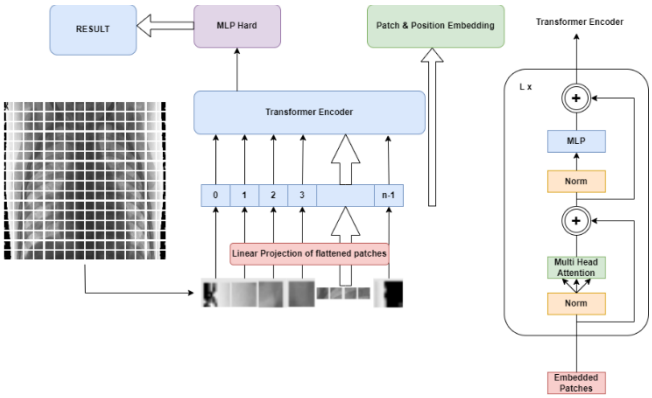
X-rays relies heavily on the expertise of radiologists, resulting in delays in diagnosis, particularly in less developed regions. In response, some researchers have turned to Convolutional Neural Networks (CNNs) to automate pneumonia detection. However, limitations in these networks, such as restricted local receptive fields, may hinder their ability to capture global contextual information, potentially impacting classification accuracy. To tackle this problem, this research seeks to improve the accuracy of pneumonia identification from chest X-ray images through the implementation of Vision Transformer (ViT) technology. By utilizing a self-attention mechanism, ViT can proficiently capture long-range relationships and intricate details, thereby enhancing overall diagnostic efficacy.

4.Existing Model

Recent advancements in deep learning methodologies and computational capabilities have resulted in the development of numerous frameworks dedicated to the analysis of pneumonia from chest radiographs. Convolutional Neural Networks (CNNs) have emerged as pivotal components in the realm of medical image classification, showcasing superior performance. Notable models like VGG16, ResNet, and Densenet have been successfully applied to pneumonia detection. These models excel due to their utilization of convolutional layers for extracting hierarchical features, enabling them to identify spatially arranged features within the data hierarchy. Additionally, models employing transfer learning have proven effective in pneumonia detection by repurposing deep learning architectures originally trained on tasks such as ImageNet to operate on pneumonia-specific datasets.

5.Vision Transformers

The Vision Transformer (ViT) is an innovative architectural design that adapts the Transformer model, traditionally designed for Natural Language Processing (NLP), to the task of image classification. In contrast to normal Convolutional Neural Networks (CNNs) which mainly solve the local feature extraction problem via convolutions, ViT self-attention enables focusing on every position of the image in relation to the other positions. This aspect renders it useful in applications which entail comprehension of context, for instance, in medical imaging. Accompanying the attention mechanism is a feed-forward neural network whose role is to enhance the representations obtained from attention. To improve the stability of training, layer normalization as well as residual connections are applied. The output of a special [CLS] token, which combines information from all patches, is then sent to a classification head that assigns the final class labels. The merits of ViT consist of its power to model long-range dependencies, amenability to larger scale training with datasets, and compatibility with varying input size which makes it powerful for sophisticated tasks. In the area of medical imaging, ViT has been able to carry out diagnosis by detecting pneumonia in chest X rays and this is because owing to its contextual understanding it can see changes that are difficult to spot in normal breast images which helps the doctors in diagnosing faster and accurately. All in all, the Vision Transformer represents a radical transformation in the way image classification is executed as it seeks to counter the use of CNNs and opens opportunities for further exploration and application in many different fields, particularly the diagnosis of diseases from medical images.



Architecture diagram of ViT

Patch Extraction: First, the input image is divided into non-overlapping patches, usually measuring 16×16 pixels. Each patch is then flattened into a one-dimensional vector. A linear transformation is applied to these vectors to produce a lower-dimensional representation, transforming the 2D image data into a series of patch embeddings. This approach allows the model to perceive each patch as a distinct token, similar to words in a sentence.

Positional Encoding: To preserve the spatial context of the patches, positional encodings are incorporated into each patch embedding. These encodings deliver vital information about the location of each patch within the original image, which is important for the model to comprehend the arrangement of features.

Transformer Encoder Layers: The backbone of the ViT consists of multiple layers of transformer encoders. Each encoder consists of two main components: a multi-head self-attention mechanism and a feedforward neural network. The self-attention mechanism allows the model to evaluate the relevance of different patches relative to each other, effectively capturing long-term dependencies. This capability is in contrast to CNNs, which may miss such global relationships. The output of the self-attention layer is then processed by a feed-forward network, followed by normalization and residual connections to improve training stability.

Classification Mechanism: A classification head receives the output of the last Transformer layer that corresponds to a unique [CLS] token that represents the full image. This head uses the combined data from the patches to produce the final

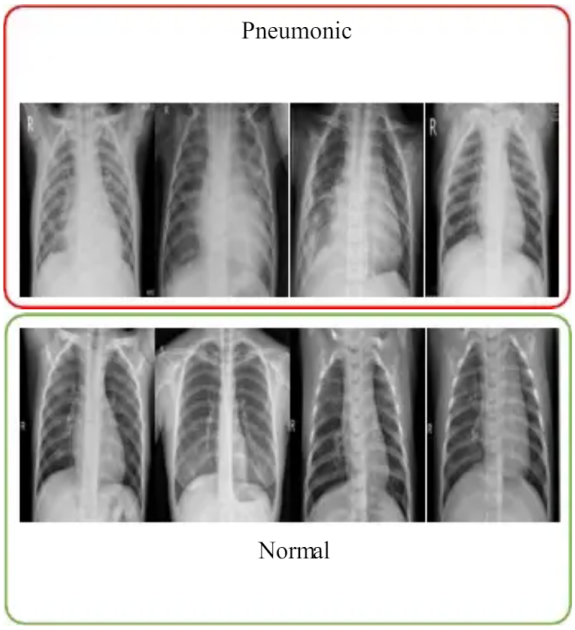
class predictions; it is usually a straightforward feed-forward network.

6.Literature Survey

Through the use of chest X-ray scanning, deep learning techniques have been extensively published throughout the course of the previous ten years. These techniques have been used to automatically diagnose lung abnormalities, including infections. Deep learning models are superior to other medical imaging approaches when it comes to identifying pneumonia from chest X-rays, as demonstrated by a study that is particularly important and was conducted by Rajpurkar and colleagues. In particular, these advancements in the field are important because they assist to tackle the problem of accurately identifying pneumonia, which is an illness that is often mistaken with other lung-related disorders. This is especially true in youngsters. The work used seven pretrained Convolutional Neural Network (CNN) models to improve the accuracy of pneumonia diagnosis in paediatric patients. Moreover, it showed how well CNN models distinguished X-ray pictures of normal, viral, and bacterial pneumonia while obtaining remarkable sensitivity and Area Under the Curve (AUC) values. Particularly in vulnerable paediatric groups, the writers underlined the vital requirement of accurate and quick diagnosis of pneumonia. By suggesting a simplified deep learning approach utilising MobileNet, they were able to attain admirable accuracy, recall, and F-score metrics. With its low computing demand and short training length, the model's efficiency marks a major progress in enabling early identification and treatment of pneumonia. The work presented a unique three-tier optimisation strategy to solve the problem of limited annotated computed tomography (CT) scan availability for pneumonia diagnosis. This method efficiently increases the accuracy of pneumonia identification by means of CT data from one domain to improve the performance of deep learning models in another area lacking labelled scans. Using CT scans, a new deep learning algorithm called Pneumonia-Plus produced consistent identification of several kinds of pneumonia with impressive diagnostic performance reflected by AUC values. This strategy demonstrated its potential in reducing misdiagnosis and enhancing clinical decision-making, as it surpassed two of three radiologists in the classification of bacterial and viral pneumonia. The research emphasised the necessity of intelligent systems to address the limitations of traditional human-assisted methods, such as cost and accessibility to expertise, and introduced a scalable and interpretable deep convolutional neural network (DCNN) designed to automate pneumonia diagnosis from chest X-ray images.The proposed DCNN model has superior performance in feature extraction and image categorisation into normal and pneumonia categories compared to existing state-of-the-art approaches, as evidenced by comprehensive evaluations across many performance metrics. The study significantly aided in the identification of complex pulmonary diseases through chest X-ray images and provided a comprehensive framework for assessing lung ailments, including pneumonia and COVID-19. This framework encompassed image enhancement, precise region of interest extraction, resilient feature extraction, and classification employing sophisticated methodologies such as ensemble classifiers, Artificial Neural Networks (ANN), Support Vector Machines (SVM), K-Nearest Neighbours (KNN), and a custom deep learning architecture utilising Recurrent Neural Networks (RNN) with Long Short-Term Memory (LSTM).

7. Dataset

This study utilised a dataset of 5,856 X-ray pictures for the training, testing, and validation of the models. 1,583 images depict Chest X-Rays of patients without Pneumonia, while 4,273 images represent Chest X-Rays of patients with Pneumonia. All chest X-ray images were acquired during the daily regular examinations of patients. For additional information on image quality, please refer. Chest X-ray images may exhibit blurriness, making pneumonia detection challenging for the human eye. Simple phlegm may be misinterpreted as pneumonic pus, as both conditions can induce haziness in X-ray images. The test and training datasets are categorised into two groups: Pneumonia and routine chest X-rays. The model was trained and evaluated using a dataset comprising 90% pneumonic chest X-ray images and 84.7% normal chest X-ray images in the training set, while the testing set included 10% pneumonic chest X-ray images and 15.3% normal chest X-ray images.



Showing Two Categories of images in Data-set

8.Methods:

The proposed methodology for pneumonia detection using Vision Transformers (ViT) consists of a systematic pipeline, including dataset selection, preprocessing, model architecture, training, and performance evaluation. The following sections provide a comprehensive breakdown of the process.

7.1 Data Preprocessing

Model training requires certain preparation, especially preprocessing the images. The following techniques are performed on the acquired raw data, performed on images depicting different subjects:

Image Resizing: In the case of the ViT where the models require a fixed-length vector input, all images are resized to a standard dimension that is likely to be 224 × 224 in most cases. This helps maintain the quality of the images and also helps in the use of the ViT model that has been developed.

Patch splitting: Now after resizing those images, a resized image will be broken down into smaller fragments known as patches measuring square 16 × 16 pixels with no overlapping. For an image measuring UIImage of 224×224 results in 196 images. Each patch will then undergo a transformation process which flattens the patch into a single vector ready for processing by the Vision Transformer.

Image Normalization: Correction to the pixel dimension high and low ends during the estimation process of pictures is interpolated and image data normalized to either nonnegative values – ranging between 0 and 1 or centered image data with unit standard deviation. This normalization process helps in accelerating the chances of convergence during training of the model.

Data Augmentation: In combating overfitting especially in small datasets, age techniques are in built. Such as random horizontal flips, little rotations, increases in brightness, contrast. These techniques use the concept of enlarging the dataset which enhances the performance of the model over changes present in the data.

Data Segmentation for Training and Testing: The available data is partitioned into segments designated solely for training, exclusively for testing, and contains a validation component. Typically, 70% of the data is allocated for training, 15% for validation, and 15% for testing. Stratified sampling ensured that the ratio of pneumonia cases to normal patients was preserved throughout all categories.

7.2 Vision Transformers

The ViT model is rather an exception when compared to other neural networks, especially the Convolutional Neural Networks. Below is a thorough explanation of its standpoints:

Patch Embedding: Since ViT is unable to perform 2D image processing the same way as CNN does, the image is rather segmented into smaller sections known as patches, which are flattened and projected into a fixed-size vector. Every patch is treated like a “token” similar to how words are used in NLP tasks. For a 224×224 image having 16×16 patches, this gives rise to 196 tokens, which vary in dimension by 768.

Positional Encoding: The Raster scan pattern of the linear array engenders a variation of the spatial focal attention model. This means that all image sections would be processed equally without the biases embedded in position relative to image centering that a net

Transformer Encoder Layers: The model’s backbone is all composed of several layers of Transformer encoders. Each of those comprises:

Multi-Head Self-Attention: This is useful for determining how much each patch (or piece of data) is relevant to other patches and for capturing the dependencies between different areas of the image extending beyond the boundaries. This is important especially with medical images where the critical aspects may be scattered within the image.

The output from the self-attention phase is directed into a two-layer Feed-Forward Network (FFN) employing a non-linear activation function (ReLU). This procedure occurs subsequent to the self-attention phase. The FFN promotes additional refinement of the representations derived from the self-attention layer.

Layer Normalization: In order to speed up and stabilize training, layer normalization is made sure to be applied to every input before the self-attention and FFN layers. This is necessary to prevent excessive gradient flow and help in convergence.

Residual connections: Skip - or residuum - connectors are present at the entrance to the layer, as well as at its exit, to avoid problems with the disappearing gradient, thus contributing to the training of networks with a larger number of layers.

Classification Head: 'Special' token [CLS] is added to the beginning of a sequence with patch embeddings. After that, the complete image is represented through the last hidden state of [CLS] token after going through all available layers of the Transformer. This token is then forwarded to the dense output layer which generates probabilities for the output classes (pneumonia versus normal cases).

7.3 Training Process

There are various ways of constructing Vision Transformer model.

Loss Function: In this case, in this binary classification cross-entropy is employed as the loss function. Thus, cross-entropy helps in measuring how different the probabilities predicted (output) are from the actual labels.

Optimization: Adam optimizer is mostly preferred in this task as it can take the best of two worlds momentum and RMSProp. A learning rate scheduler is used to manipulate the learning rate when training, that is, a higher learning rate is used at first and then lowered after some time.

Transfer Learning: Vision Transformer model is transformed by utilizing a pre-trained Vision Transformer on ImageNet which is an example of a much larger dataset. Since this model has seen many similar images, it can perform pneumonia detection task very well. Moreover, Fine-tuning helps to modify the pre-learned weights on the chest X-ray imaging features.

Batch Size and Epochs: The training is done with the help of small mini-batches and the standard size of batch is either 32 or 64. The number of epochs (full pass through the dataset) is variable and there is early stopping to take care of overfitting if validation loss does not reduce or increases with time.

Regularization: To take care of overfitting, regularization methods like dropout, for instance, are implemented. Dropout is a robust method that zeroes out a proportion of the weights of the network during training to promote better generalization of the model.

7.4 Evaluation Metrics

Statistical techniques are frequently employed in data analysis to investigate classification challenges, such as those examined in this paper. A primary instrument employed for this objective is the confusion matrix, a tabular depiction that outlines the occurrences of true positives, true negatives, false positives, and false negatives generated from the predicted results during the assessment phase. These numerical values serve as the foundation for deriving various statistical parameters that evaluate the precision of the predictions made. In this particular study, a range of statistical metrics are employed, including:

1. The true positive rate, referred to as recall, sensitivity, or TPR, is computed as TPR=TP/(TP+FN) and serves as a measure of the proportion of positive instances that were correctly predicted. Given the stringent requirements of medical systems for a high recall to minimize false negatives, this metric plays a critical role in medical diagnostics.

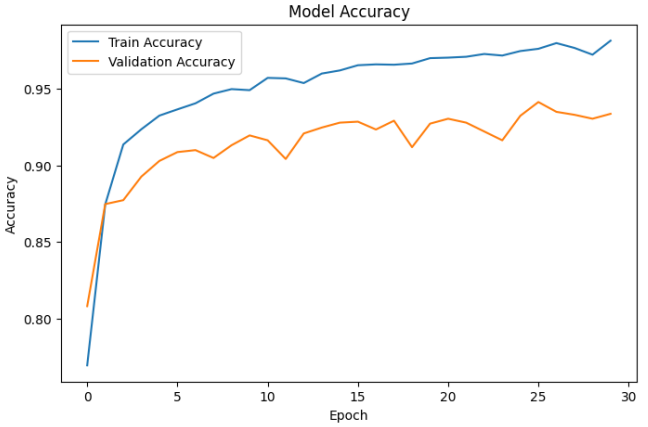
2. Precision, or positive predictive value (PPV), refers to the proportion of real positives among the instances identified as positive by the model. The equation for precision is PPV = TP / (TP + FP).

3. The F1 score, a statistic defined as the harmonic mean of True Positive Rate (TPR) and Positive Predictive Value (PPV), is expressed as: F1 = 2 × TPR × PPV / (TPR + PPV). This metric assigns equal importance to erroneous positives and false negatives, rendering it a significant measure of model efficacy.

4. Accuracy (ACC) quantifies the proportion of correct predictions, calculated as (TP+TN)/(TP+TN+FP+FN). This score offers a comprehensive evaluation of the model's proficiency in making accurate selections across all categories.

8. Result

Recall	97.59%
Precision	98.64%
F1 score	98.10%
Accuracy	98.12%



9. Conclusion

The results of the project clearly demonstrate that the Vision Transformer (ViT) is a strong contender in the field of image classification, especially in distinguishing medical X-rays. Through the utilization of a dataset containing binary-labeled chest X-ray images, our ViT model achieved an impressive accuracy of 98.12%. Our study indicates the promising advantages of employing ViT and confirms its efficacy as a viable replacement for Convolutional Neural Networks (CNN). The noteworthy discoveries made by ViT have garnered attention, paving the way for further

investigation by the computer vision community in the coming years.

10. References

[1] Chandan Chakraborty, Asok K. Maiti, Mallika Pal, Madhumala Ghosh, and Dev Kumar Das (2013) utilized light microscopic images and machine learning to automatically conduct tests for malaria parasites, as detailed in Micron 45 (2013).

[2] George Thoma, Stefan Jaeger, Richard Maude, Mahdieh Poostchi, and Kamorat Silamut (2018) conducted research on malaria detection through the utilization of machine learning and image analysis, as highlighted in Research on Translation (2018).

[3] Adriano G. Duse, David M. Rubin, Charles J. Pritchard, and Nicholas E. Ross (2006) developed an automated image processing method for the detection and categorization of thin blood smear malaria, as discussed in Computing and Medical and Biological Engineering 44, 5 (2006).

[4] Stefan Carlsson, Josephine Sullivan, Hossein Azizpour, and Ali Sharif Razavian (2014) emphasized the effectiveness of Convolutional Neural Networks (CNN) as a reliable tool for recognition, as presented in the Proceedings of the IEEE Workshops on Pattern Recognition and Computer Vision. In their groundbreaking 2012 study entitled "Deep Convolutional Neural Networks for ImageNet Categorization,"

[5] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton explore the use of deep convolutional neural networks for this specific purpose. The study was published in Developments in Neural Information Processing Systems and has significantly contributed to advancements in image classification technology.

[6] Andrew Zisserman and Karen Simonyan published a research paper in 2014 titled "Convolutional networks with extreme depth for large-scale picture recognition" in the arXiv preprint arXiv:1409.1556.

[7] According to F. Chollet's research conducted in 2016, Xception utilizes separable convolutions to enhance the performance of deep learning algorithms. This study was published in the arXiv Preprint arXiv1610 2357.

[8]Kaiming He, Shaoqing Ren, Jian Sun, and Xiangyu Zhang authored a paper in 2016 on image recognition through the use of deep residual learning, which was presented at the IEEE Conference on Computer Vision and Pattern Recognition Proceedings.