SOLVING CHEST X-RAY IMAGE CLASSIFICATION USING A VARIANT OF VISION TRANSFORMER

XINYI PANG [PXINYI@SEAS.UPENN.EDU], XILING ZOU [XILING@SEAS.UPENN.EDU],

ABSTRACT. Transformers have been successfully applied to solve computer vision problems in a variety of fields, including medical imaging. In this project, we aim to implement a variant of vision transformers (ViT), the Input Enhanced ViT (IEViT), to classify Chest X-ray images, which are critical in the diagnosis of many diseases. Comparing to the vanila ViT that has F1-score of 90%, the F1-score of IEViT we implemented increased to 92%.

1. INTRODUCTION

Medical imaging has largely improved the process of diagnosis and treatment of numerous medical conditions in both children and adult by using ionizing radiation to generate images of the body. Inside the family of medical imaging, chest X-ray (CXR) remains the most commonly performed radiological exam in the world, as the first imaging study acquired and remains central to screening, diagnosis, and management of a broad range of conditions. Comparing to other methods, it is cost-effective and has low radiation dose, and has a reasonable sensitivity to a wide variety of pathologies.

It is difficult and challenging to interpret CXR because of the overlapping of different organs and tissues being projecting onto the same image along the same direction, which makes it very hard for abnormality detection in some particular locations, detection of small or subtle abnormalities, or accurately distinguishing between different pathological patterns. Because of this, the analysis of CXR images are different due to the difference in the measurements between observers. Therefore, given the complication of the CXR image, the high volume acquired, the difficulty of interpretation, and the value in diagnosis and treatment, it is crucial to build a system of algorithms that can automatically detect and interpret abnormalities in an accurate, effective, consistent way.

Various approaches in machine Learning has been used for automated CXR analysis that show the exciting opportunity of matching or exceeding the performance of medical experts. Transformer architecture has been a state-of-art model solution for a variety of Natural Language Processing tasks since the paper “Attention Is All You Need.” was published in 2017. In 2021, the first transformer model, Vision Transformer (ViT), pre-trained on ImageNet was proposed. Our goal for this project is to implement a VIT based network to solve CXR image classification problems.

1.1. **Contributions.** We implemented a customized ViT by adding convolutional blocks to the vanilla ViT. We experimented the proposed architecture on a binary classification task of detecting pneumonia via CXR images. We achieved 92% test accuracy, 0.92 weighted macro F1-score, and 0.99 recall in identifying the disease. In comparison to vanilla ViT, it exhibits a 2.2% increase in F1-score and a 13.8% improvement in recall. This method can be used to enhance the effectiveness of radiograph-based illness diagnosis

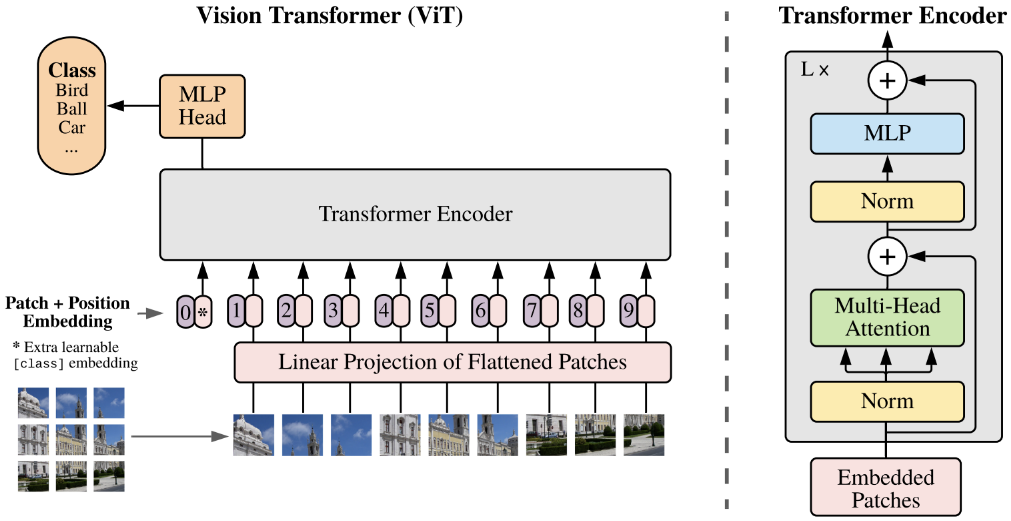
2. BACKGROUND

Vision Transformer (ViT)

Vision Transformer (ViT)[1] is a state–of-the-art model on image classification that adapts the architecture of BERT for language modeling with some modification. The detailed architecture is shown in the figure below. The input image is splited into fixed size patches and each patch is pass to be processed to a linear embedding. The embeddings are then feed to many Transformer Encoder layers to extract the features of each patch. In the end, for classification task, a dense layer is used to predict the label of the input image using the extracted features.

3. RELATED WORK

One research[4] on Chest X-ray imaging implemented Transfer Learning with ResNet to solve complex CXR image classification problems. The data they used is the CoronaHack-Chest X-Ray-Dataset on Kaggle. The layer freeze technique adopted in this paper is that using the previously trained model as a fixed feature extractor, and then training



the last layer as the classifier. Their F1-scored on the CoronaHack-Chest X-Ray-Dataset are 0.9304, 0.929, 0.9428, 0.9424, 0.9447 for ResNet18_v1, ResNet34_v1, ResNet50_v1, ResNet101_v1, and ResNet152_v1 respectively.

Comparing with ResNet152_v1, VIT has more parameters and self-attention layers to enhance the ability to extract features from the images. We hypothesized that our custom model that combined the concept of ResNet and VIT could have a better performance than ResNet itself if using the same data.

One research[5] looks into the viability of employing a deep learning-based decision-tree classifier to detect COVID-19 in CXR pictures. To achieve a more detailed classification, the proposed classifier consists of three binary decision trees, each trained by a deep learning model with a convolution neural network. The first decision tree determines if the CXR images are normal or aberrant. The second tree identifies aberrant photos with TB symptoms, while the third does the same for COVID-19. The first and second choice trees have accuracies of 98 and 80%, respectively, while the third decision tree has an average accuracy of 95%. Our binary classification model has the potential to be employed as each level of a decision-tree based classifier for more complicated, multiclass categorization.

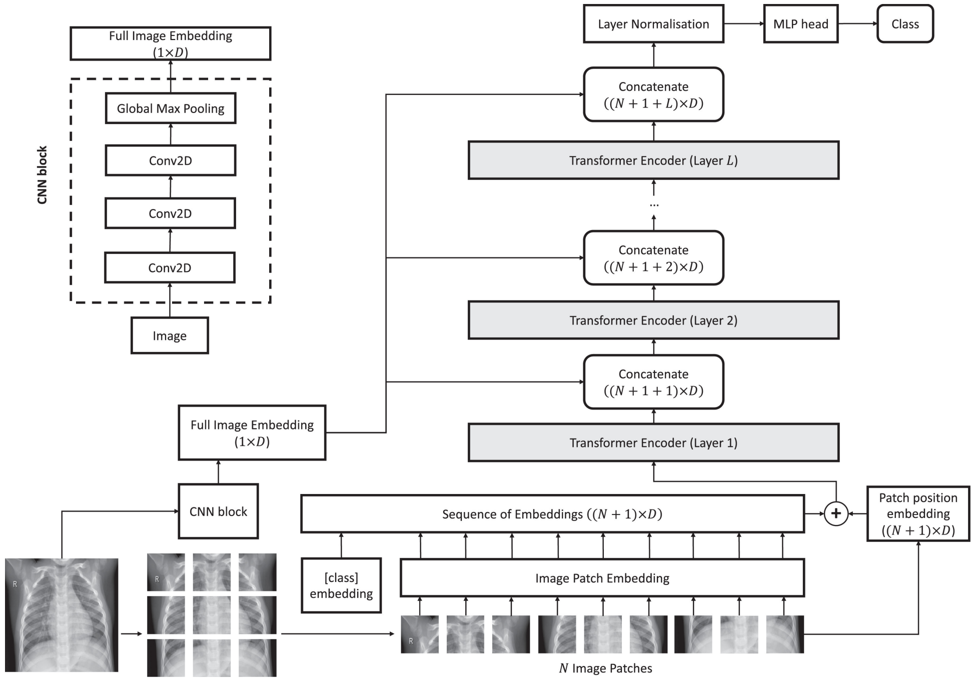
The other research[2] proposed an improved IVGG13 model, a modified VGG16 model for classification pneumonia X-rays images using the same dataset as we did, the Chest X-Ray Images (Pneumonia) from Kaggle. The proposed model was a reduced depth VGG16 to avoid underfiting/overfitting during training. The proposed IVGG13 model required less training time and resources compared with the other considered CNNs and had f1-score of 84.6%. Comparing to our model, this proposed IVGG13 model indeed had less training time, but just like mentioned before, we hypothesized that our custom model could capture more features using self-attention mechanism and achieve a better performance.

4. Approach

In this research, we implement the Input Enhanced ViT (IEViT) proposed by Gabriel Iluebe Okolo and Stamos Katsigiannis and Naeem Ramzan[3]. As mentioned in that paper, we first fit a vanila version of ViT as a baseline to examine its performance on our CXR data. Then, we implement IEViT to see whether the performance was improved.

4.1. Model.

The IEViT model combined the concept of ViT and Resnet by adding the representation of the original input to the output of each transformer encoder block. Comparing to the vanila version of ViT, an vector representation of the whole image is calculated for the original input using a CNN block in parallel with the ViT network. Then, this embedding vector is concatenated iteratively to the output of each Transformer encoder layer. In the end, we add a dropout layer before the final dense layer to avoid overfitting. The CNN block consists of 3 convolution layers and a max pooling layer. The ViT model used in this research is vit_b16 provided by vit-keras.



4.2. **Training Process.**

Both the vanila ViT and the ViT architecture in IEViT were initialized with the pre-trained weights. The Vanilla ViT was first trained for 1 epoch using all the weights. Then, all the layers were frozen except the last dense layer and was fine-tuning for another 10 epochs. The IEViT was frist trained for 3 epoch using all the weights. Then all the layers except the CNN block and the last two layers were frozen and were fine-tuning for another 10 epochs.

5. Experimental Results

5.1. **Data.** The dataset used in this work are Chest X-Ray Images (Pneumonia) posted on Kaggle. It is divided into three folders: train, test, and val, with subfolders for each image category (Pneumonia/Normal). There are 5,863 JPEG X-Ray images and two categories (Pneumonia/Normal). All chest radiographs were initially reviewed for quality control before being analyzed, with all low quality or unreadable scans being removed. There are a total of 5216 images in the training set, with 1341 labeled as Normal and 3875 as Pneumonia. We utilized 16 photos to tune the learning rate and validate the early stopping setting, and 624 images to evaluate the model.

5.2. **Preprocessing.** Data augmentation has been shown to be an excellent method for picture classification, and it is most commonly employed in deep learning approaches to expand the amount of training data and help reduce over-fitting. To achieve better generalization, we performed data augmentation using the ImageDataGenerator provided by Keras. We chose a brightness shift value range of 0.5 to 1.5, as well as a width and height shift range of 0.1. The shear, rotation, and zoom ranges are all set at 0.1.Random vertical and horizontal flips were used. We only augmented training and validation data; the test data findings only pertain to the original images. Following this procedure, batches of augmented images were created in real-time during each training procedure.

5.3. **Evaluation Matrix.** Classification performance was measured for both the vanilla and modified ViTs using the following metrics: accuracy, precision, recall, and F1-score, which is the harmonic mean of precision and recall. The macro and weighted average of the last three are also calculated.

$$F1\ score = 2 * \frac{Precision * Recall}{Precision + Recall}$$

In the case of F1-score, recall, and precision, the metrics were computed for each class and the average across classes (macro average), as well as a weighted average, was presented. F1-score was used as the primary performance metric because we have more pneumonia images than normal images and classification accuracy can be skewed in cases of

	Precision	Recall	F1-score	Number of Smaples
Nomral	0.82	0.94	0.87	234
Pneumonia	0.96	0.87	0.91	390

TABLE 1. Classification Performance of Vanilla ViT

	Precision	Recall	F1-score	Number of Smaples
Nomral	0.97	0.88	0.88	234
Pneumonia	0.89	0.99	0.94	390

TABLE 2. Classification Performance of Customized ViT

Metrics	Precision		Recall		F1-score		Training Accuracy	Test Accuracy
	macro	weighted	macro	weighted	macro	weighted		
Vanilla ViT	0.89	0.91	0.91	0.90	0.89	0.90	0.9632	0.90
Customized ViT	0.93	0.92	0.90	0.92	0.91	0.92	0.9711	0.92

TABLE 3. Performance Comparison between Vanilla ViT and Customized ViT

unbalanced data sets. We would also favor recall over precision because we don't want to omit any patients by making false negatives.

5.4. **Result.** The classification performance by vanilla ViT and our customized ViT are shown in 1 and 2. The comparison between the two are presented in 3. It is evident that our customized ViT outperformed the vanilla one in all F1-score and accuracy measures. Both the macro and weighted F1-scores have improved by 2.2%. We have also increased test accuracy by 2.2%.

6. DISCUSSION

Our experiment demonstrates that adding the CNN block to the ViT architecture and concatenating its output to the output of each Transformer encoder layer enhanced binary classification performance. The average improvement in F1-score across the dataset over the original ViT is 2.2%. It's also worth mentioning that the recall for pneumonia has increased by 13.8% compared to the vanilla ViT, reaching 99%, indicating that we've greatly reduced false negative cases. It is critical in medical diagnostics to increase recall since we aim to accurately diagnose every patient who gets the disease. In comparison, a few false positives are more tolerable.

The improved classification performance comes at the expense of extra convolutional layers to compute, as well as an increase in the amount of the input to the Transformer encoder layers and hence more parameters to train.We'd like to consider the tradeoff between improved accuracy and greater computational cost while adding convolutional layers in various practical applications.

If we have additional time, we would like to fit the model using the datasets mentioned in the IEViT paper to compare the performance of our implementation and theirs. In that paper, we noticed that the accuracy of the model on all the datasets they used was over 95% and some was even close to 100%. By doing this, we could further explore whether the difference in accuracy is caused by the difference in data or in the way of implementation.

In addition, we would like to test the customized ViT in multiclass classification scenarios, such as further distinguishing CXR of bacterial pneumonia from viral pneumonia and other diseases that can be detected using X-ray pictures.

REFERENCES

[1] 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.

[2] Zhi-Peng Jiang, Yi-Yang Liu, Zhen-En Shao, and Ko-Wei Huang. An improved vgg16 model for pneumonia image classification. *Applied Sciences*, 11(23), 2021.

[3] Gabriel Iluebe Okolo, Stamos Katsigiannis, and Naeem Ramzan. Ievit: An enhanced vision transformer architecture for chest x-ray image classification. *Computer Methods and Programs in Biomedicine*, 226:107141, 2022.

[4] Sadia Showkat and Shaima Qureshi. Efficacy of transfer learning-based resnet models in chest x-ray image classification for detecting covid-19 pneumonia. *Chemometrics and Intelligent Laboratory Systems*, 224:104534, 2022.

[5] Geng H Yoo SH. Deep learning-based decision-tree classifier for covid-19 diagnosis from chest x-ray imaging. *Front Med (Lausanne)*, 7(427), 2020.