**Tuberculosis Image classification using ViT Transformers**

Name – Ranjit singh

Student id – 202112053

**Summary**

This project discusses the process of creating a Visual Transformer (ViT) model for image classification task using 2 datasets. The first was a custom dataset: photos of Japanese models and a young boy and the second one chest X-ray images of healthy individuals and patients with tuberculosis.

At first a ViT model was created from the beginning, resulting in inadequate accuracy between 2 to 10%. Following that a pre-trained ViT-B16 model was utilized, resulting in a noteworthy enhancement in accuracy reaching 70% to 75% on the initial or first dataset. After achieving success, the pre-trained model was used for another dataset of chest X-ray images from Kaggle, achieving an impressive 98% accuracy.

After training and testing the model, it was built into a small application to classify normal lung and TB lung images. The model was saved using the pickle module, which is used for saving and loading machine learning models. The saved model was then deployed on the web using the Flask framework, with HTML and CSS used for the web structure. This application demonstrates the model's practical utility in distinguishing between normal and tuberculosis-affected lungs, making it a valuable tool in medical diagnostics.

This report ends by emphasizing the possible practical uses of ViT models in medical imaging especially in situations where traditional Convolutional Neural Networks (CNNs) face challenges due to the immense size of medical images.

**Introduction**

Visual Transformers (ViT) have emerged as a promising approach for image classification tasks, offering an alternative to traditional Convolutional Neural Networks (CNNs). This report documents the development and evaluation of ViT models for classifying two distinct types of image datasets that are used in this project: Japanese model images versus images of boys and chest X-ray images of normal versus tuberculosis patients.

**Working of Vit Transformers**

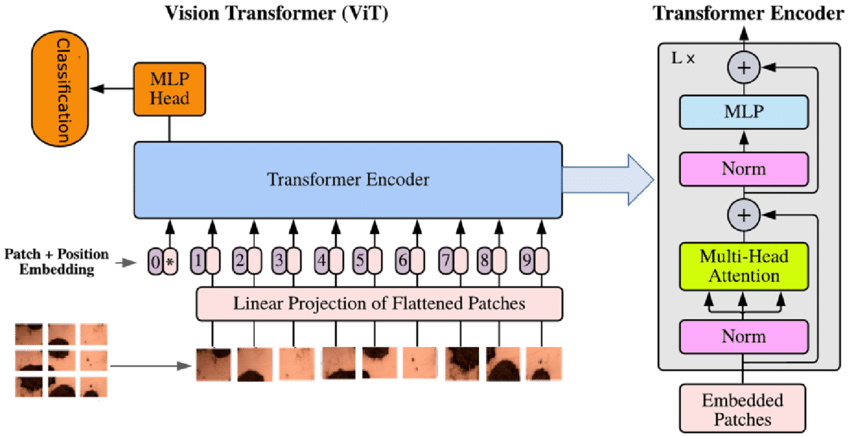


Figure1 : ViT Model architecture

Visual Transformers (ViT) work by breaking down an image into smaller pieces called patches. Each patch is then transformed into a format that the model can understand, like a language. Just as you read words in a sentence, the model "reads" these patches to understand the image.

When you feed images into a ViT model, the image is divided into patches, and each patch is represented as a vector, which is like a set of numbers that describe the patch's features. These vectors are then flattened and combined into a sequence, much like arranging words into a sentence.

The model then processes this sequence of patch vectors through multiple layers of attention mechanisms. These attention mechanisms allow the model to focus on different parts of the image while considering the relationships between them. It's like paying attention to different words in a sentence to understand its meaning.

As the model processes these sequences, it learns to recognize patterns and features in the image data. With enough training, the model can accurately classify images based on these learned patterns.

So, in simple terms, the images are broken down into smaller pieces, transformed into a format the model can understand, and then processed through layers of attention mechanisms to learn and recognize important features, ultimately making predictions about what the image contains.

**Methodology**

The project began with the creation of a ViT model from scratch. However initial results proved to be inadequate with accuracy range varying between 2 to 10%. Recognizing the limitations of the custom-built model, a pre-trained ViT-B16 model was selected as an alternative. After choosing the pretrained model the performance was significantly improved which achieved the accuracies between 70% and 75% on the first dataset.

|  |  |
| --- | --- |
|  |  |

Figure 2: ViT scratch model accuracy on custom image dataset

**Results**

Following the success of the pre-trained ViT-B16 model on the first dataset, it was further applied to a separate dataset containing chest X-ray images. This dataset, downloaded from Kaggle, consisted of images depicting normal and tuberculosis-affected conditions. The pre-trained ViT model demonstrated exceptional performance on this dataset attaining an accuracy of 98%.

|  |  |
| --- | --- |
|  |  |

Figure 3: pertained model accuracy on chest X ray dataset

**Build an application**

After achieving impressive results in image classification, the next step was to transform the trained model into a practical tool for real-world applications. To accomplish this, I decided to develop a small application that could classify normal lung and TB lung images in a user-friendly manner.

Firstly, I saved the trained model using the pickle module, a widely used tool for saving and loading machine learning models. This ensured that the model's learned parameters and architecture could be easily accessed and utilized for inference.

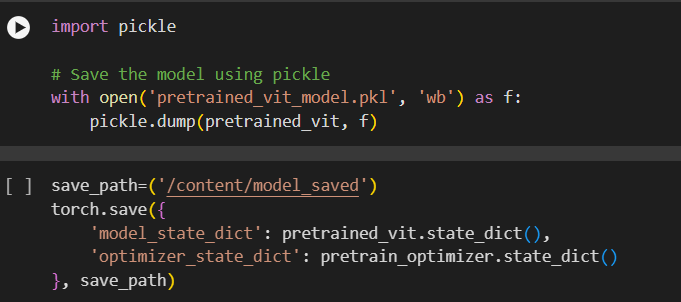
****

Figure 4: Saving Vit Model using pickle Module

Next, I employed the Flask framework, a lightweight and versatile Python web framework, to deploy the saved model on the web. Flask provided a straightforward way to create a web server and expose the model's classification functionality as a web service, accessible via HTTP requests.

To enhance the user experience and provide a visually appealing interface, I utilized HTML and CSS for the web structure. This allowed me to design a user-friendly interface where users could upload the lung images and receive immediate classification results.

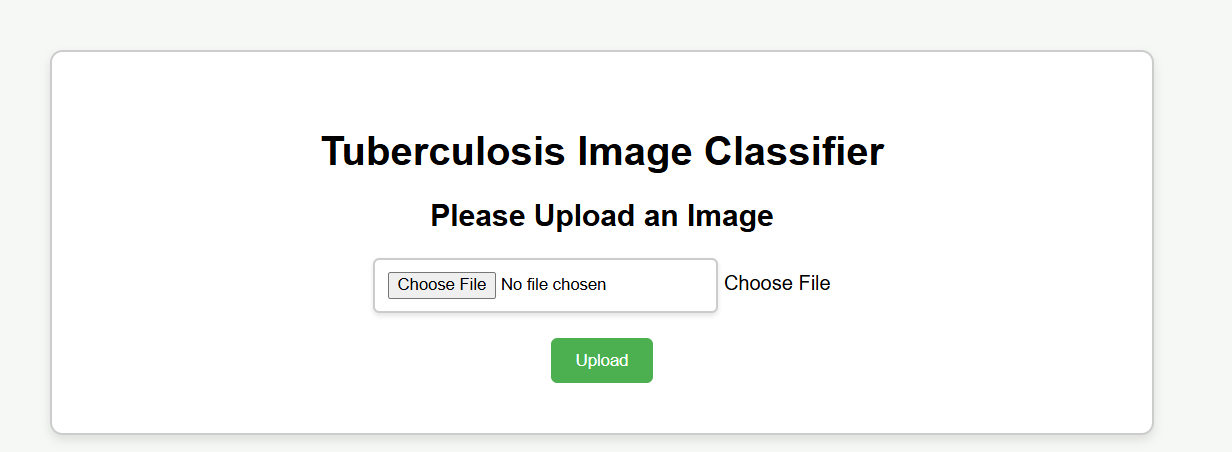


Figure 5: Application Homepage Window

After uploading an image, the application directs the image to the trained model for prediction. The model analyses the image and makes a classification decision based on its learned patterns and features.

Once the prediction is made, the application displays the prediction results in a separate window. This window includes essential information such as the "prediction class" indicating whether the image depicts a normal lung or a TB-affected lung. Additionally, the "Prediction confidence score" is presented, providing users with insight into the model's certainty regarding its classification. Alongside these details, the original image is displayed, allowing users to visually confirm the classification outcome.

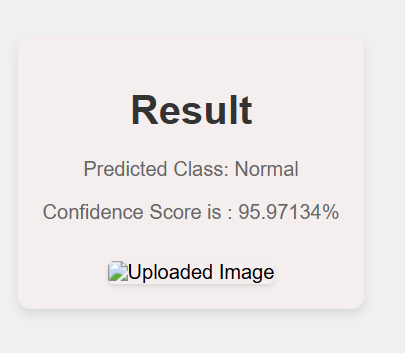


Figure 6: Result Window

**Discussion**

The significant improvement in accuracy achieved by leveraging a pre-trained ViT-B16 model underscores the effectiveness of transfer learning in image classification tasks. The success of the ViT model in classifying both custom images and medical images highlights its versatility and potential applications in various domains.

**Conclusion**

In conclusion, this project showcases the effectiveness of Visual Transformer models, particularly the pre-trained ViT-B16 architecture, in image classification tasks. The promising results obtained on both custom and medical image datasets suggest that ViT models hold immense potential for real-world applications, particularly in medical imaging where conventional CNNs may struggle to capture the entirety of large-sized medical images. Further exploration and refinement of ViT models could lead to even more significant advancements in image classification and analysis.