NTNU
Kunnskap for en bedre verden

IMT4895 - SPECIALISATION IN COLOUR IMAGING

Binary Classification Using Transformers For Skin Lesions

Authors: Alen Bhandari
Supervisor: Sule Yildirim Yayilgan
Co-Supervisor: Sarang Shaikh

December, 2021

Contents

List of Figures

1 Abstract

The success of transformers in the Natural language processing(NLP) domain, it has shift it's focus in the Computer vision task. Due to which numerous research are being carried out along with creation of different models either in hybrid form(combination of transformers and Deep Neural Network) or in the pure form. In this project we are trying to dive deep into the implementation of Vision transformer(ViT),trying to understand it's architecture for the Binary Classification task and finally implementing it. The domain that we choose for this project is Medical domain where we will be utilizing all the three color channels which is different from traditional computer vision task. Finally, we will be evaluating the Transformer model with other state-of-the-art Deep Neural Network like ResNet50 and pre-trained ResNet50.

2 Introduction

Skin cancer has been one of the mist common form of cancer with an annual exceeding of $8 billion. But if diagnosed early then the survival rate can be up to 99%.Due to this reason, the importance of early detection has diverted the researchers to focus on increasing accuracy and diagnostic method. For this project, we have selected the dataset from International Skin Imaging Collaboration (ISIC) challenge for analysing our model.

After the success in Natural Language Processing(NLP) tasks, we have noticed Transformers evolving in the computer vision task. In the nutshell transformers can be implemented in two ways:

1. Either implementing attention incurrence with Convolutional Neural Network(CNN),
2. or, Combine the best features from both CNN and Transformers and build a hybrid form.

Due to it's ability of computational efficiency and scalability we can now build models of bigger size, train models in a large datasets. But despite it's scalable abilities and success in NLP, in terms of computer vision tasks Convolutional architectures still remain dominant. The major factor for this is, despite it being combined with different CNN models, or completely replaced it with attention model, it has not been able to scale efficiently on modern hardware accelerators due to the use of specialized attention patterns. Hence, in a scenario where we have large scale images, traditional Deep Neural Networks(DNN), models similar to ResNet-like architecture are still state of the art.

Therefore, we were interested to investigate the potentiality of transformers in Medical imaging where the focus is on Skin Lesions. Which is why we have decided to continue this project on the Masters thesis level. This project takes the motivation from above statements and act as a preliminary part for the Masters thesis. Technically in this project we are experimenting the performance of classical ResNet50 model, ResNet50 with a pretrained weight(Transfer Learning) and finally, the Transformers with colored image. Further Description is provided in the Methodology Section, Section 5.

2.1 Research Questions

This Project will try to answer the following research questions:

1. Are Transformers a better replacement for the Classification problem compared to classical Deep Neural Networks for color images?
2. How does transformer performs in the small dataset compared to classical Deep Neural Networks(DNN)?
3. Does Segmentation help to improve the accuracy of the transformers?

2.2 Scope

In order to narrow the project down to the course project level, we needed to define some limitations to the overall Scope. We have thus created these limitations:

- The project aims to cover the importance of transformers in a Skin Lesions domain.
- All the models implemented takes colored images(RGB channels) as an input.
- Classification for No-Data Augmentation flow as shown in the proposed system model in section 5 is not covered in this project
- This project focus only on the implementation of transformers as a binary classification along with the evaluation of the model.

2.3 Structure of Paper

This report starts with section 3 - Background, which outlines the introduction of transformers along with brief description of the architecture and it's components . No solutions or conclusions are given, but rather their purpose is to both make the reader aware of them, and to mention them as they are used later in the report.

After, getting the introduction about the Transformers we will look into it's relevant literature in section 4. We will look into the use of transformers as a classification model in various domains, keeping the main focus on the Medical Imaging.

The next section is section 5, Methodology. The goal here is to describe the implementation pipeline. Describing individual steps involved in the pipeline and their contribution in the experiment.

In section **??**, Findings and Results, In this section we will be reporting each and every result from different experiments along with the description of the experimental setup.

In Section 7 - Discussion debates the outcome of the experiment giving insight of the reason behind the outcomes and shortcomings.

Finally, section 8 Concludes the project with a concise summary and the future works.

3 Background

Before we dive into the architecture, we need know the definition of the keywords used: **Attention, Transformers,Vision Transformers(ViT),Patch embedding,Positional encoding,Multi-Layer Perceptron Head(MLP)**

3.1 Transformers:

3.1.1 Attention

According to [2], attention was evolved from the problems related to time-varying data. Due to which it became famous for the task such as sequences. Before transformers Seq2Seq models were used for the tasks such as translation, eg: translating sequence of words from one language to another [14]. But the problem was it performs good only with small sequence of time stamps data. With longer sequences the system eventually pays attention to the last part of the sequence.

Finally with attention, [2] they were able to form a direct connection with each timestamp. And this was originally developed for the computer vision tasks. According to [13], by taking just glimpse of the images it can accumulate information about the shape and classify images accordingly. The same principle was

later used for sequences where the model can look at different part of images at the same time by creating **patch embedding** and learn to "pay attention" to the correct ones.

After the paper "Attention Is All You Need"[21], the way how we treated attention in DNN, or Deep Learning Models (DLM) changed. As described above Transformers implements attention-mechanism. With sufficient data, we can now perform matrix multiplication, layer normalization along with State-of-the-art-machine-translation [2].

Transformers were able to achieve all this just by changing the input representation i.e. sets and tokenization. There are three main steps to consider for pre-processing of the input to the model, which are:

1. Tokenization,
2. Word Embedding, here the semantics of the inputs are captured by keeping semantically close inputs together in the embedding space,
3. Positional encoding, when the input sequences are converted into sets , the information about the order is changed. Due to which positional encoding are added to the word embedding vector before the 1st self-attention layer.

After the input is ready for the model it is fit into the transformer encoder. Further description of the transformer encoder is found in the below section where we explain the architecture in terms of Vision Transformers.

3.2 Vision Transformers(ViT):

ViT divides an image into same size patches, embed them sequentially, and add the positional embedding as an input to the Transformer Encoder.If trained with efficient data, it can outperform the State-of-the-art CNN by about 4 times less computational resources.

Vision transformers have gained increased popularity for image recognition tasks recently, signaling a transition from convolution-based feature extractors (CNNs) to attention-based models (ViTs)[9].After the success of beating the state of the art for NLP the transformers are gaining quite popularity in the image domain.The transformers shows quite promising results by outperforming the CNNs on Standard vision task such as IMAGENET classification [1], along with the object detection [2] and semantic segmentation [3].

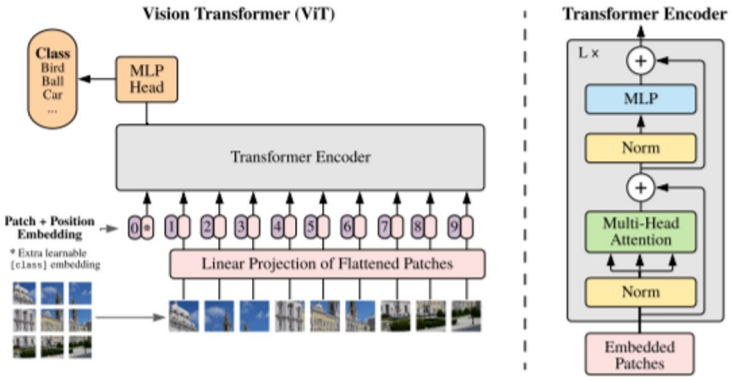

Figure 1: Figure Vision Transformer Architecture [9] on the left side of the image and on the right side transformer encoder

Transformers basically lacks inductive biases of CNNS, which is basically a set of assumptions that the model uses to predict the output. In a nutshell Vision Transformers works as explained in the following steps:

1. First the images are converted into patches,
2. the patches are then flattened,
3. from the flattened patched they produce lower dimension linear embedding,
4. after that the positional embedding is added,
5. then the input sequence is feed into the transformer encoder,
6. the model is then pre-trained in a huge dataset with image labels,
7. it is then finetuned on required dataset for the image classification.

Converting images into sequence of patches: For splitting the images into fixed-size patches it is first split into fixed size patches. Here, the 2D image of size H * W is split into N patches where N = H * W / P $\hat{2}$. For example,if the image with size of 48 by 48 and the patch size is 16 by 16, then there will be 9 patches for the images.

Positional Encoding: Before the positional encoding happens the image is flattened into 1D path embedding by concatenating all the pixel channel into patch and then linearly projecting it into the desired input dimension. After that, the learnable positional encoding are added to each individual path allowing the model to learn about the structure of the image.

3.2.1 Transformer Encoders:

As shown in figure 1, the Transformer encoders three main components:

1. **Multi-Head Self Attention Layer(MSP):** It is a repeated version of scaled dot product, where we apply Attention onto several copies of the same vector which have been filtered/transformed in meaningful ways, then aggregating the results.
2. **Layer Norm (LN):** It is used to improve the training time along with generalizing the performance. Also, it is applied before every block.
3. **Residual Connections (RC):** It basically allows the gradient flow through the network directly without the non-linear activation. Also, it is applied after every block.
4. **Multi-Layer Perceptron(MLP):** It is the module at the output layer which provides the desired class prediction. Also, we can pre-train the model in the large dataset of an images and then the final MLP head can be fine tuned to a specific task via the process called transfer learning.

4 Relevant Literature

According to [23] In a computer vision domain, the image information is stored as array of pixels where the pixel arrays are processed by convolutions. Despite having successful models(Convolution), there are some major challenges:

1. Convolutions processes the image patches regardless of importance,
2. Every image has their own high-level features, so the feature used for one image cannot be applied to another,
3. Finally, Convolutions lack the information of long-range interaction i.e. they struggle to relate spatially-distant concepts.

To address the above challenges ViT was introduced [9]. In [23], they create their own Visual Transformer as a Image Classification Model by inheriting the backbones from the ResNet [10]. They conducted the experiment where they compare VT-ResNets (Proposed model) with the default vanilla ResNet keeping the setup same. With their one of the variation of VT-ResNet they were able to get accuracy of **75.0 %** on the validation dataset(Imagenet) and **80.8 %** on the Training set compared to the same variation of ResNet with accuracy of **73.3%** on validation dataset and **73.9%** on training.

Due to the success of Vision Transformer(ViT), researchers are constantly trying to improve the model performance by either modifying/changing the components of the transformers or by combining it with the convolution models. In the process of doing so a new model was introduced CrossViT [5]. In the paper they try to figure out how to learn multi-scale feature representation in transformer models specifically for image classification task. They propose a dual branch transformer model that combines the image patches of different size to produce more robust image feature. They were able to get an accuracy of **82.8%** with their CrossViT-18 model compared to DeiT [20] baseline on ImageNet1k which was **81.8%**

5 Methodology

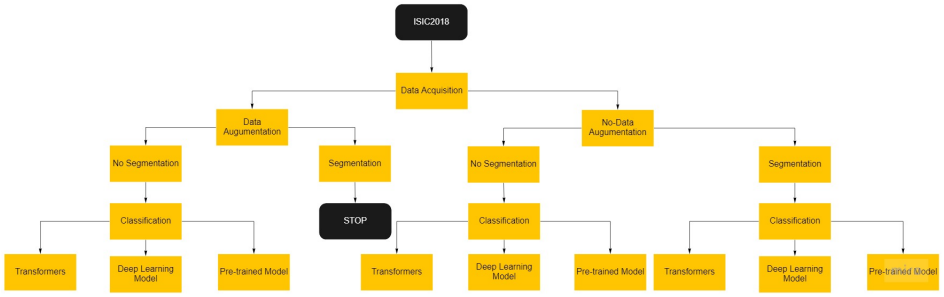

Figure 2: Proposed System Pipeline

The Proposed system Pipeline is as shown in the figure above, but it is sub-divided into smaller pipeline based on the experimental setup. Each sub divided pipeline has its own purpose which collectively is responsible for answering the research questions mentioned in the section 2.In a simplified form the Proposed system consist of the following components:

1. **Data Acquisition**
2. **Data Augmentation**
3. **Segmentation**
4. **Classification**
5. **Evaluation**

5.1 Data Acquisition:

For this experiment we decided to select the database provided International Skin Imaging Collaboration(ISIC) 2018 challenge [6]. The challenge contains 3 main parts based on which they provided the database. But for our experiment we are only taking the database from **Part 1: Lesion Segmentation** and **part 3: Lesion Disease Classification**.

Due to the reason that all the experiments were performed in the local machine, we had to reduce the number of images in the database along with reducing the size of the image. For our experiment we have selected only the Melanomas and Non-Melanomas(for a binary classification) which has 1400 training images and 193 testing images. For training the segmentation model we had limited the images to 743.

```
Database
   |-- Classification
           |--Segmentation Classification
               |--ISIC2018
                   |--Test
                       |-- images
                   |--Train
                       |--Melanoma
                           |-- images
                       |-- Non_melanoma
                           |-- images
                   |--Validation
                       |--Melanoma
                           |-- images
                       |-- Non_melanoma
                           |-- images
           |-- Non_segmentation_classification
               |--Similar Structure to Segmentation Classification
   |-- Segmentation
       |--ISIC2018_Task1-2_Training_Input
           |--mask
       |--ISIC2018_Task1-2_Validation_Input
           |--mask
       |--ISIC2018_Task1_Training_GroundTruth
           |--mask
       |--ISIC2018_Task1_Validation_GroundTruth
           |--mask
```

Figure 3: Folder Organization

The Folder organization for the experiment is shown in the figure above which is similar to the next experiment performed for the IMT4305 Image Processing And Analysis course.

5.2 Data Augmentation:

Data Augmentation is the process of increasing the amount of data by adding slight modification on the already existing data or creating newly synthetic data from the existing data [7]. Data augmentation process can be added in the earlier phase of the pipeline as pre-processing step or in the later phase as a post processing step. In this experiment to keep the setup intact the data augmentation step was performed in the pre-processing phase i.e. before segmentation but we found out that changing the shape and size of the image destroys the quality of the segmentation mask. Due to which we decided for the pipeline that doesnot have segmentation to implement Data Augmentation as a pre-processing step and for the other pipeline after segmentation which is post-processing step. The techniques that we implemented for the Data Augmentation are as follows:

1. Scaling.
2. Rotation.
3. Height and Width Shift.
4. Horizontal Flip.
5. Zoom.

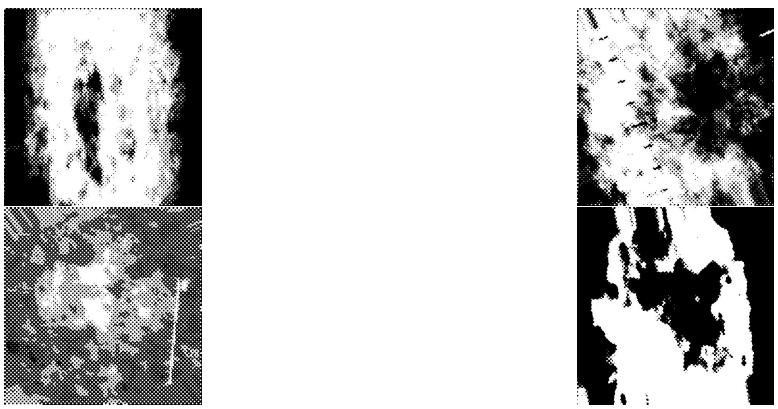

Figure 4: Mask After the Image Augmentation

5.3 Segmentation:

The pipeline splits into two parts in this phase i.e. Segmentation and No-Segmentation. The reason for doing that is we wanted to test the classification model with and without segmentation specifically Vision Transformer classification model. In any image processing pipeline Segmentation is considered as a vital step due to it's ability of extracting area of interest, object detection which are further used for recognition. And since we are using medical images as our database of interest, so instead of feeding the whole images if we could feed the classification model only the affected area of skin from the images by segmenting then the model should be able to identify the cancer more effectively. It is similar to classification problem but here the classification is done on the level of image pixels [4].

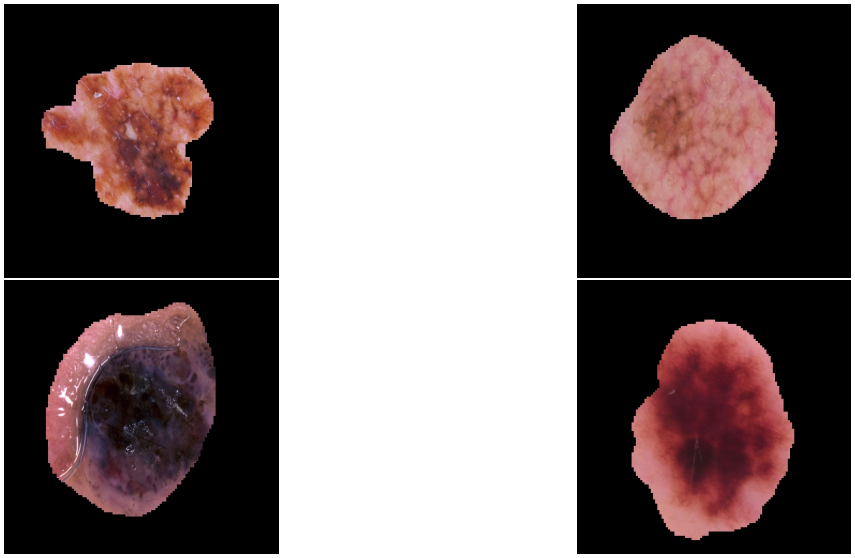

Figure 5: Extracted Images after implementing the segmentation mask

For this experiment we decided to use the U-Net model as our segmentation model. Since, it was specifically developed for biomedical image segmentation [22]. It was created by modifying Fully Convolutional Network for semantic segmentation which was originated in 2015[17]. The model basically consist of contracting path where the input images are downsampled and an expansive path where the input images are upsampled. They are constructed in such a way that they make a U-shaped architecture. In the structure,the upsampling part contains a large number of feature channels through which the context information is passed towards higher resolution layers[22].

5.4 Classification:

Classification is the process of labeling and systematic arrangements of group of pixels or vectors within an image according to the established criteria[18]. They are basically divided into two parts:

1. **Supervised Learning:** Here, we already know which category/labels the image belongs to and we feed both the training images along with the respective labels to the machine learning models for training.
2. **Unsupervised Learning:** Here, we train the model without providing the ground truth labels. The model tries to figure out the categories or the labels by itself.

In our experiment we are using the Supervised learning method where we have separated the images into Melanoma and Non-Melanoma category for every training and testing dataset as shown in 3. The models takes the images from the particular folder along with their labels. For the classification task we have selected ResNet50 model as our deep learning model, the same ResNet50 model but with the pre-trained

weight and finally Vision transformer. As we can see in the proposed system diagram 2, both the pipelines from Segmentation and Non-Segmentation end up in the Classification. The purpose of the classification module is to:

1. test the performance of the Vision Transformer as a classification model,
2. test the best performing model with segmentation among the Deep Learning Model in pure form, Deep Learning model with pre-trained weights, and Vision Transformer,
3. test the best performing model without segmentation among the Deep Learning Model in pure form, Deep Learning model with pre-trained weights, and Vision Transformer

5.5 Evaluation:

During the training phase of the model, we need to keep track of how the model is preforming. Also to determine the best performing model for the particular sample data we need evaluation metrics. During the experiment we have implemented multiple metrics to determine the particular model performance and to determine the best performing model. The Evaluation metrics used were:

5.5.1 F1 Score

F1 score also called F-score is the measure of model's accuracy. It is mainly used for the evaluation of binary classification models that classify the dataset into 'positive'(Melanoma) or 'negative'(Non-Melanoma). It is the harmonic mean between precision and recall. The acceptable range of F1 score is [0,1]. It gives the information about how well the classifier was able to predict correctly along with how robust the model is. F1 score basically tries to find the balance between precision and recall [15].

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

Figure 6: F1 Score Formula

Hence, we know that F1 score is an average of precision and recall, so it gives equal importance to precision and recall [12]:

1. the model under performance will have high F1 score if both the value of precision and recall are high which is 1.0 or 100%,
2. the model under performance will have low F1 score if both the value of precision and recall are low or any one of them is zero which is 0 or 0%,
3. the model under performance will have medium F1 score if one of the value of precision and recall is low and the other is high.

5.5.2 Recall

Recall is the measure of correct Positive prediction done by the model. Thus, in our problem domain for all the images in the dataset, recall gives us the information of how much we correctly identified the images having Melanomas out of all Melanomas that could have been identified. It also gives us the measure of how accurately our model is able to identify the relevant data [16]. In case of binary classification problem having two classes, recall is calculated as the number of true positives divided by total number of true positives and false negatives [3].

$$Recall = \frac{\#\ of\ True\ Positives}{\#\ of\ True\ Positives + \#\ of\ False\ Negatives}$$

Figure 7: Recall Formula

6 Findings and Results

The experiment were carried out for different pipeline as described in the section 5. For both the pipelines i.e with segmentation and without segmentation the model configuration for all the models were kept similar.

In case of ViT, we limit the image size to [128 * 128 * 3] and patch size to 4 (image dimension must be divisible by patch size [9]) with MLP dimension 128. We are using 'gelu' as the activation function which is a smoother version of RELU as explained in [11] with 2 dense layers(for binary classification). The batch size was kept at value 64 along with the maximum epochs of 10 with learning rate of 0.001.

In case of ResNet50, we have limited the image size to [224, 224, 3] with optimizer as Adam having learning rate 0.001 and loss function categorical_crossentropy. The batch size for this model was kept 16 with maximum epoch of 10. We have also implemented early stop i.e. callbacks so that the model can stop itself if it's not progressing anymore with patience level 3. Patience level basically means that the model will wait for 3 epochs before stopping the training if the loss is not decreasing.

In case of ResNet with pre-trained weight, we imported the model from official Tensorflow website [19] which was pretrained in Imagenet [8] having 1000 labels. All the model configuration such as optimizer, loss function, batch size, epoch and callbacks were kept similar to ResNet50 model.

Based on the above experimental setup, the qualitative analysis is displayed in the form of plots and the quantitative analysis is presented in the form of tables.

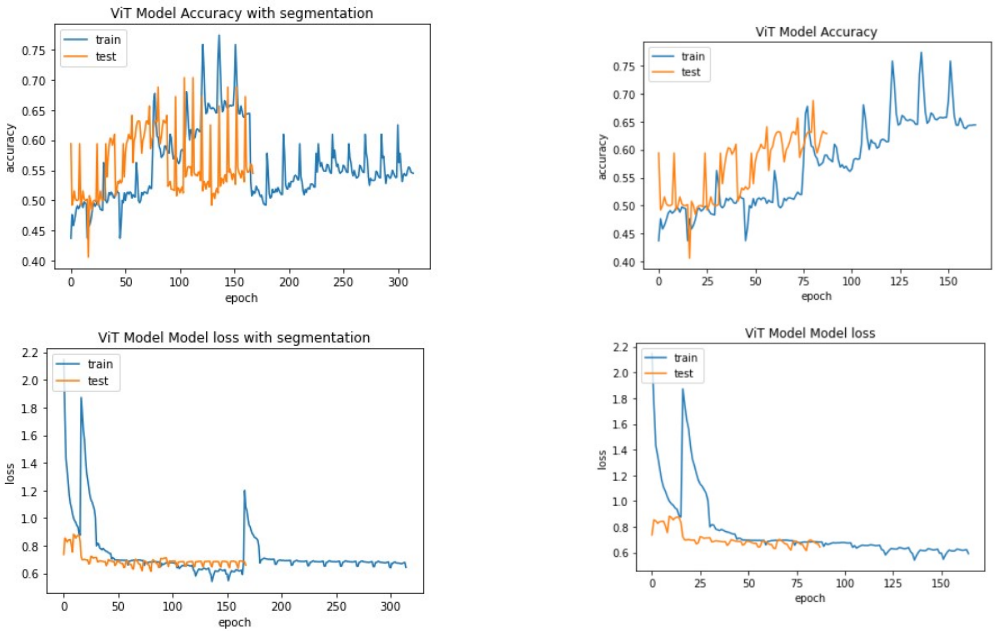

Figure 8: The plot on the left refers to ViT model performance on Segmented images while the plot on the right refers to ViT model performance without segmentation

In the above plot we can see that the ViT model performs better in the setup without segmentation with smoother curve for the model loss in train dataset. But for the test dataset the model has better accuracy

curve in the setup without segmentation. Based on the quantitative analysis, the ViT model accuracy was recorded as 54.50% for train dataset and 54.40% for the test dataset with F1score of 49.80% for the experiment with segmentation. For the experiment without segmentation the accuracy was recorded as 64.39% for the train dataset and 62.85% for the test dataset with F1score of 49.83% in the rest dataset.

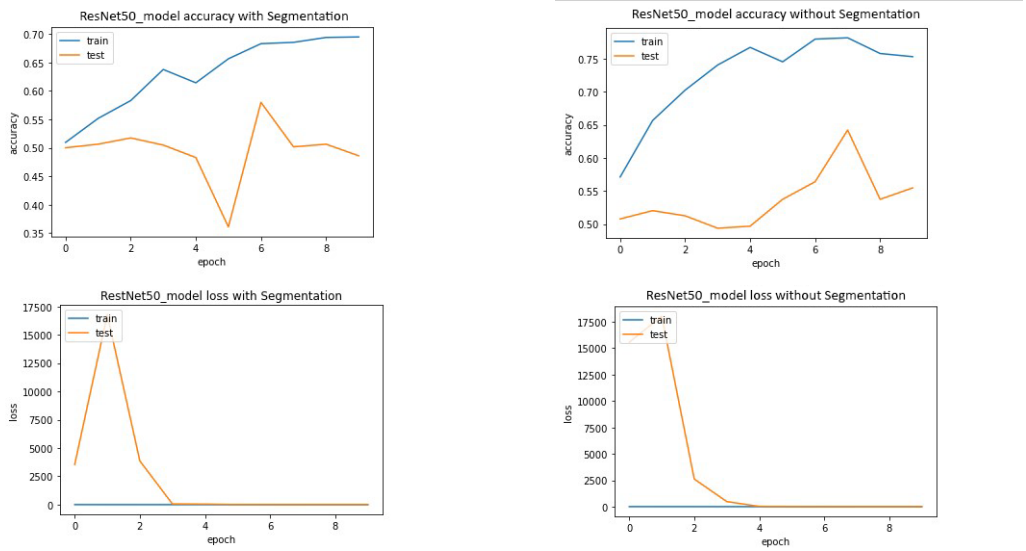

Figure 9: The plot on the left refers to ResNet50 model performance on Segmented images while the plot on the right refers to ResNet50 model performance without segmentation

In the above plot we can see that the ResNet50 model performs better in the setup without segmentation with smoother curve for the model accuracy and loss in train dataset. Based on the quantitative analysis, the ResNet50 model accuracy was recorded as 69.40% for train dataset and 48.59% for the test dataset with F1score of 50.62% for the experiment with segmentation. For the experiment without segmentation the accuracy was recorded as 75.31% for the train dataset and 52.71% for the test dataset with F1score of 53.70% in the rest dataset.

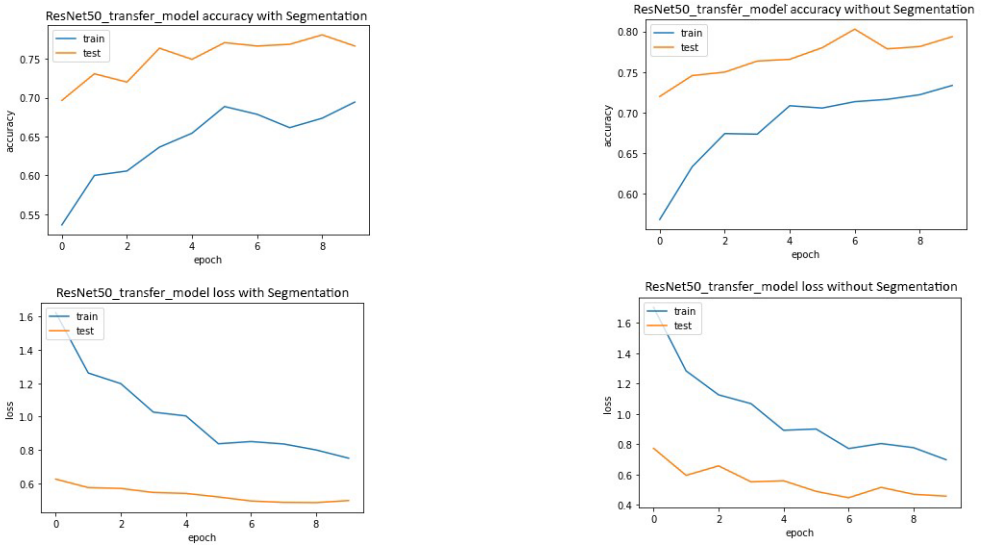

Figure 10: The plot on the left refers to pretrained ResNet50 model performance on Segmented images while the plot on the right refers to pretrained ResNet50 model performance without segmentation

In the above plot we can see that the pre-trained ResNet50 model performs better in the setup without segmentation with smoother curve for the model accuracy and loss for both train and test dataset. Based on the quantitative analysis, the pre-trained ResNet50 model accuracy was recorded as 69.42% for train

dataset and 76.64% for the test dataset with F1score of 78.07% for the experiment with segmentation. For the experiment without segmentation the accuracy was recorded as 75.33% for the train dataset and 79.35% for the test dataset with F1score of 78.14% in the rest dataset.

	Without Segmentation					
Models	Accuracy		Recall		F1Score	
	Train	Test	Train	Test	Train	Test
ResNet50	75.31%	52.71%	75.31%	55.40%	75.57%	53.70%
Transfer Learning(ResNet50)	73.33%	79.35%	73.35%	79.35%	72.21%	78.14%
ViT	64.39%	62.85%	55.26%	80.40%	56.45%	49.83%

Figure 11: Table showing different Evaluation Metrics for all the model for the experiment setup with segmentation

	With Segmentation					
Models	Accuracy		Recall		F1Score	
	Train	Test	Train	Test	Train	Test
ResNet50	69.40%	48.59%	69.46%	48.59%	69.30%	50.62%
Transfer Learning(ResNet50)	69.42%	76.64%	69.42%	76.64%	67.35%	78.07%
ViT	54.50%	54.40%	55.66%	84.80%	56.40%	49.80%

Figure 12: Table showing different Evaluation Metrics for all the model for the experiment setup without segmentation

7 Discussion

As mentioned in the section 2, this project serve as a preliminary work for the master thesis. In this project we are trying to see the feasibility of transformers in the Computer vision task specially in the Medical Image domain. Due to which we have formulated specific research question that we want to answer within this project. Throughout this section, we will be explaining all the findings from the result section 6 while answering the research questions.

7.1 RQ1:Are Transformers a better replacement for the Classification problem compared to classical Deep Neural Networks for color images?

Based on the result that we got from the two experimental setup we can say that classical Deep Neural Networks still perform better than the transformers. But if we analyse the result in more details going deep into the metrics then we might be able to get more insight about feasibility of the transformer.

As, we can see in the table in the figure 11 and 12 ViT has the highest Recall in both the experiment for the test dataset. As we know recall gives the measure of how many actual positive the model was able to predict out of all the positives. So, it means the ViT model implemented in the experiment was actually able to correctly predict maximum number of Melanoma case compared to other models. Also, the domain we are experimenting is the medical domain which is really a sensitive field in the computer vision. So, it is better to get more true Melanoma detected rather false Non-Melanoma which is really bad.

Hence, we can conclude based on the experimental setup that for the computer vision task still classical Deep Neural Networks are better as they have good accuracy and F1 Score but for the Medical domain transformers performs really well as they have high recall values.

7.2 RQ2:How does transformer performs in the small dataset compared to classical Deep Neural Networks(DNN)?

Based on the findings in the paper [23] where they combine transformers with ResNet and train the model in a large dataset such as Imagenet. They reported accuracy of **75.0%** on the validation dataset and **80.8%** on the Training dataset which when compared to our model performance which gave an accuracy of **62.85%** on

the validation dataset and **64.39%** on the training dataset. We can conclude that ViT needs a large amount of dataset for being able to perform better. Also, we can conclude that for small dataset the Pre-trained ResNet performs the best on both the experimental setup.

7.3 RQ3:Does Segmentation help to improve the accuracy of the transformers?

Transformers are based on the attention mechanism where the model gives attention to every pixels in the initial run but after each step of patch encoding and patch embedding the model focus only on the relevant part of image i.e. model give attention only to those pixels that are relevant which is basically the functionality of segmentation. So, the research question were formulated to test if the segmentation makes any difference in the performance of the transformers(ViT).

Based on the result from the experiment the model showed better performance without segmentation with accuracy of 62.85% compared to accuracy of 54.40% with segmentation. But if we look at the F1score of the model then on both experiment its similar i.e 49.83% without segmentation and 49.80% with segmentation. The reason for that is the datasets are completely balanced in our experiment due to which the f1 score does not give any significance information in case of binary classification [1].

8 Conclusion

The main objective of this project was to test the feasibility of the transformers for color images in Medical domain. We realized that transformers are better suited for colored images, is it better for Medical Domain ? Based on our experimental setup, yes, but we cannot generalize it because we haven't experimented for the other image domains. Also, ViT despite having low accuracy and low f1 score compared to other 2 models(ResNet50 and Pre-Trained ResNet50) it has high recall among them which means Transformers are able to detecting more true positives.

9 Future Work

To build further on the proposed system and generalizing the result, we still need to perform some more experiments and we still need in-depth research:

1. Execute the same experimental setup with a large dataset from ISIC challenge,
2. Also, execute the same experimental setup in the standard computer vision dataset such as ImageNet, CIFAR-10. So, that we can generalize the findings and infer some valuable insights,
3. The metrics selected in this experiment were limited, so we need more metrics for evaluating the models more precisely,
4. Also, for transformers we are using R,G,B channels but how does ViT performs in other image channels? We need to perform another experiment before inferring any insights.

Bibliography

[1] shark8me (https://stats.stackexchange.com/users/36717/shark8me). *How to interpret F-measure values?* Cross Validated. URL:https://stats.stackexchange.com/q/238550 (version: 2021-08-11). eprint: https://stats.stackexchange.com/q/238550. URL: https://stats.stackexchange.com/q/238550.

[2] Nikolas Adaloglou. *How transformers work in Deep Learning and NLP: An intuitive introduction*. Dec. 2020. URL: https://theaisummer.com/transformer/.

[3] Jason Brownlee. *How to calculate precision, recall, and F-measure for imbalanced classification*. Aug. 2020. URL: https://machinelearningmastery.com/precision-recall-and-f-measure-for-imbalanced-classification/.

[4] Weiling Cai, Songcan Chen and Daoqiang Zhang. 'Fast and robust fuzzy c-means clustering algorithms incorporating local information for image segmentation'. In: *Pattern Recognition* 40.3 (2007), pp. 825–838. DOI: 10.1016/j.patcog.2006.07.011.

[5] Chun-Fu Chen, Quanfu Fan and Rameswar Panda. 'Crossvit: Cross-attention multi-scale vision transformer for image classification'. In: *arXiv preprint arXiv:2103.14899* (2021).

[6] Noel Codella et al. 'Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (isic)'. In: *arXiv preprint arXiv:1902.03368* (2019).

[7] *Data augmentation*. Sept. 2021. URL: https://en.wikipedia.org/wiki/Data_augmentation.

[8] Jia Deng et al. 'Imagenet: A large-scale hierarchical image database'. In: *2009 IEEE conference on computer vision and pattern recognition*. Ieee. 2009, pp. 248–255.

[9] Alexey Dosovitskiy et al. 'An image is worth 16x16 words: Transformers for image recognition at scale'. In: *arXiv preprint arXiv:2010.11929* (2020).

[10] Kaiming He et al. 'Deep residual learning for image recognition'. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 770–778.

[11] Dan Hendrycks and Kevin Gimpel. 'Gaussian error linear units (gelus)'. In: *arXiv preprint arXiv:1606.08415* (2016).

[12] Joos Korstanje. *The F1 score*. Aug. 2021. URL: https://towardsdatascience.com/the-f1-score-bec2bbc38aa6.

[13] Hugo Larochelle and Geoffrey E Hinton. 'Learning to combine foveal glimpses with a third-order Boltzmann machine'. In: *Advances in neural information processing systems* 23 (2010), pp. 1243–1251.

[14] Maxime. *What is a Transformer?* Mar. 2020. URL: https://medium.com/inside-machine-learning/what-is-a-transformer-d07dd1fbec04.

[15] Aditya Mishra. *Metrics to evaluate your machine learning algorithm*. May 2020. URL: https://towardsdatascience.com/metrics-to-evaluate-your-machine-learning-algorithm-f10ba6e38234.

[16] *Precision vs recall: Precision and recall machine learning*. Mar. 2021. URL: https://www.analyticsvidhya.com/blog/2020/09/precision-recall-machine-learning/.

[17] Olaf Ronneberger, Philipp Fischer and Thomas Brox. 'U-Net: Convolutional Networks for Biomedical Image Segmentation'. In: *Lecture Notes in Computer Science* (2015), pp. 234–241. DOI: 10.1007/978-3-319-24574-4_28.

[18] M Shinozuka and B Mansouri. '4 - Synthetic aperture radar and remote sensing technologies for structural health monitoring of civil infrastructure systems'. In: *Structural Health Monitoring of Civil Infrastructure Systems*. Ed. by Vistasp M. Karbhari and Farhad Ansari. Woodhead Publishing Series in Civil and Structural Engineering. Woodhead Publishing, 2009, pp. 113–151. ISBN: 978-1-84569-392-3. DOI: https://doi.org/10.1533/9781845696825.1.114. URL: https://www.sciencedirect.com/science/article/pii/B9781845693923500049.

[19] *Tf.keras.applications.resnet50.ResNet50 nbsp;: nbsp; Tensorflow core v2.7.0*. URL: https://www.tensorflow.org/api_docs/python/tf/keras/applications/resnet50/ResNet50.

[20] Hugo Touvron et al. 'Training data-efficient image transformers & distillation through attention'. In: *International Conference on Machine Learning*. PMLR. 2021, pp. 10347–10357.

[21] Ashish Vaswani et al. ‘Attention is all you need’. In: *Advances in neural information processing systems*. 2017, pp. 5998–6008.

[22] Wikipedia. *U-Net*. [Online; accessed 1-December-2021]. 2021. URL: https://en.wikipedia.org/wiki/U-Net.

[23] Bichen Wu et al. ‘Visual transformers: Token-based image representation and processing for computer vision’. In: *arXiv preprint arXiv:2006.03677* (2020).