Vision Transformers (ViTs) for the Classification and Diagnosis of Diabetic Retinopathy

Sarim Farrukh Siddicky

W.H. Morden

Table of Contents:

Abstract:

It is predicted that approximately 2.2 billion people live with some form of vision impairment, and around 103 million people worldwide have some variant of diabetic retinopathy. Out of these 2.2 billion people, a large proportion of whom (90%) live in middle-to-low income countries, where access to eye care/diagnosis is limited.

The purpose of this project is to create an easy, affordable ,and efficient diagnostic tool for diabetic retinopathy.

It was hypothesized that if a Python script is written to take 2-D fundus images of diabetic retinas as input data and placed into the Vision Transformer architecture with class labels corresponding to the type/severity of diabetic retinopathy, the ViT should be able to classify diabetic retinopathy. It should be able to achieve this with a high degree of accuracy as the transformer model will have been trained with a sufficient amount of epochs. This hypothesis was proven correct.

In this project, I download diabetic retinal images from Kaggle, and separate them into five classes, corresponding to the types of Diabetic Retinopathy (DR). These five classes are labeled “no DR”, “mild NPDR”, “moderate NPDR”, “severe NPDR” as well as “PDR”. Then, I use Pytorch, an open source library, to implement the ViT architecture and train/validate it using 30 epochs, resulting in a peak training accuracy of ~95%. I do this all within a Kaggle notebook. Finally, I downloaded the model path after it was done training and implemented it onto a simulation meant to detect the DBRP using user image inputs.

Inquiry Question:

Is it possible to detect diabetic retinopathy (DBPR), along with its variants, using the Vision Transformer (ViT) architecture as well as diabetic retinal fundus camera images as input data?

Introduction:

The eye, a complex and convoluted tool responsible for vision, consists of numerous components, one of which including the retina. The retina is a thin light-sensitive layer of tissue located at the back of the eye, responsible for our perception of light and color. The retina does this by converting incoming photons into an electric or chemical pulse, which is then transmitted into the optic nerve and relayed to the brain.

Nevertheless, since the process of vision is so complex, it is inevitable that complications with it will arise, which is why vision impairment diseases are such a pressing issue. According to the World Health Organization, it is predicted that approximately 2.2 billion people live with some form of vision impairment, and approximately 103 million people worldwide have some variant of diabetic retinopathy. Out of these 2.2 billion people, a large proportion of which (90%) live in middle-to-low income countries, where access to eye care is limited.

Although diabetic retinopathy has no immediate cure, there are many ways to alleviate the determinants which come with it. Regular eye check-ups, maintaining healthy blood sugar levels, and lifestyle modifications can help manage the condition and prevent vision loss. And after your blood sugar stabilizes, your eye's vision will go back to normal again.

One of the main issues regarding diabetic retinopathy diagnosis (and most optic diseases in general) is that it often requires screenings using specialized equipment and trained professionals, which in many developing countries, only a select few can afford. To take part in addressing this dilemma, the scope of this project aims to make an easy, affordable and efficient diagnostic tool for diabetic retinopathy, through the integration of Vision Transformers (ViTs), a new and emerging computer vision model.

In the dataset I am using, a clinician has assessed each image in the set for the presence of diabetic retinopathy, classifying each type of DR on a scale of 0-4. The dataset in question is the Kaggle EyePACS dataset, which provides a substantial amount of training data totaling to 35,126 data points, as depicted in the table below:

DR Scale	Type of DR	Number of Datapoints
0	No DR	25,810
1	Mild NPDR	2,443
2	Moderate NPDR	5292
3	Severe NPDR	873
4	PDR	708

Additionally, I use the ViT-Base which uses 12 layers, a hidden dim of 768, an MLP size of 3072, 12 heads and ~86 million trainable parameters. (see Table 1).

Model	Layers	Hidden size D	MLP size	Heads	Params
ViT-Base	12	768	3072	12	86M
ViT-Large	24	1024	4096	16	307M
ViT-Huge	32	1280	5120	16	632M

Table 1: Details of Vision Transformer model variants.

Hypothesis:

If a Python script is written to take 2-D fundus images of diabetic retinas as input data and placed into the Vision Transformer architecture with class labels corresponding to the type/severity of diabetic retinopathy, the ViT should be able to classify diabetic retinopathy. It should be able to achieve this with a high degree of accuracy (85-90%) as the transformer model will have been trained with enough epochs and data points to recognize the visual patterns that correspond to discrepancies within the retina's blood vessels, this ultimately being diabetic retinopathy.

Background Research:

The eye, a complex and convoluted tool responsible for vision, consists of numerous components, one of which including the retina. The retina is a thin light-sensitive layer of nervous tissue, composed of mainly photoreceptor cells as well as glial (nerve) cells. It is located at the back of the eye, and is mainly responsible for our visual processing of light and color.

In order for light to reach the retina, it must progress through a series of steps. First, the incoming light will enter through the cornea of the eye. The cornea helps refract/bend light to allow the eye to focus on the light, before it passes through an area located at the back of the cornea, called the aqueous humor. After passing through the aqueous humor, some of this light enters an opening within the eye located on the iris, called the pupil. The iris is the pigmented tissue surrounding the pupil, responsible for determining how much light the pupil lets it. Following this step, the resulting light hits the lens, a clear inner part of the iris, which works in tandem with the cornea to pass light through a clear, gelatinous substance called the vitreous. Similar to the aqueous humor, the vitreous doesn't really play a major role in the context of vision, but instead, is responsible for maintaining the overall structure/roundness of our eye. Next, the light from the vitreous will focus onto the retina. Finally, the retina will use photoreceptors and nerve impulses in order to pass visual information to the optic nerve, ultimately being relayed to the brain. This process also encompasses the connection between the eye and brain.

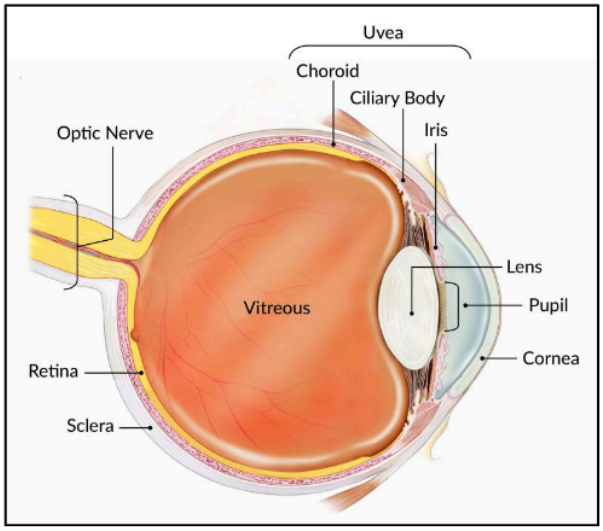

Figure 1[1]: The side view of a healthy human eye, labeling most of the major components

Photoreceptors are a type of specialized nerve cell (neuron) that convert light into neural impulses. They do this by converting incoming photons into an electric or chemical pulse, which is then transmitted into the optic nerve and relayed to the

brain. In the retina, there are two types of photoreceptor cells: cones and rods. These photoreceptors both get their names from their respective shape, and are located in the fovea of the retina.

Rods, in the retina, account for our peripheral and 3-dimensional vision. They can help the eye detect motion, as well as see at night or when there is dim lighting conditions. There are ~120 million rods located within the human retina, and they can be found anywhere along the outer edge of the retina, or the peripheral retina. This area is located outside of the macula.

Moreover, cones provide individuals with a clean and sharp central vision. They are accountable for detecting fine detail as well as color. Furthermore, the retina consists of three distinct types of cones, namely those sensitive to blue, green, and red wavelengths. Each of these cones are responsible for facilitating our perception of the corresponding colors. While the human eye's color vision is primarily based on the three types of cones (red, blue and green), our ability to see a broader spectrum of colors is achieved through blending those primary colors. Within the human eye, there are approximately 6 million cones found mostly concentrated in the central retina, this part of the eye being called the fovea centralis (commonly referred to as the “fovea”). The fovea is a vital part of the retina, housed in the macula. Additionally, the fovea has the highest visual acuity compared to anywhere else in the human eye.

Located at the back of the eye, the retina, a piece of nervous tissue, ultimately helps us perceive light into images using rods and cones. Nevertheless, since the process of vision is so complex, it is inevitable that complications with it will arise, which is why vision impairment diseases are such a pressing issue. According to the World Health Organization, as of August 2023, it is predicted that

approximately 2.2 billion people live with some form of vision impairment, and around 103 million people worldwide have some variant of diabetic retinopathy, one of the worlds' leading optic diseases.

Diabetic retinopathy (DBRP) is a complication of diabetes caused by high blood glucose levels in the body. DBRP occurs when the glucose in the body blocks the blood flow within the blood vessels connecting to the retina. This allows for the retinal vessels to bleed or leak fluids, ultimately leading to vision loss within the eye. DBRP is most prevalent among individuals in the 65-79 year old age group. According to nhs.uk, some of the underlying symptoms of diabetic retinopathy include gradual worsening vision, sudden vision loss, shapes floating in your field of vision, blurred of pathy vision, eye pain or redness and difficulty seeing in the dark. Furthermore, blood sugar/pressure levels, lipid levels (fat, oils, hormones, waxes, etc.), as well as duration of the disease all influence the severity of diabetic retinopathy.

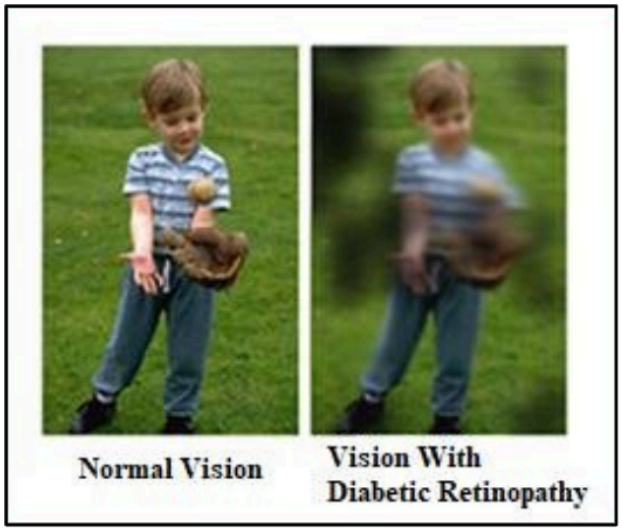

Figure 2[2]: A comparison between healthy vision and vision with diabetic retinopathy.

In order of severity, DBRP progresses through four distinct stages, these being mild nonproliferative diabetic retinopathy (the earliest stage), moderate nonproliferative diabetic retinopathy, severe nonproliferative diabetic retinopathy as well as the proliferative diabetic retinopathy, respectively. These stages are subcategorized into two main classes, these being proliferative (PDR) and nonproliferative (NPDR). Proliferative DBRP is determined through the presence

of new, proliferating, blood vessels within the retina. On the contrary, non-proliferative DBRP occurs when the blood vessels in the retina weaken, or leak blood, causing swelling or the formation of deposits in the retina called exudates.

These stages can be determined through the presence or absence of microaneurysms, hemorrhages, soft/hard exudates, venous beading, newly formed retinal blood vessels, intraretinal microvascular abnormalities (IrMA), etc. Optometrists diagnose these traits through part of a dilated eye exam or fundus camera examination. In fundus camera examinations, ophthalmologists take close up pictures of an eye, allowing them to observe and capture parts such as the retina, optic nerve head (optic disc), sclera, macula, retinal blood vessels, choroid or vitreous.

Retinal hemorrhages are the medical term for bleeding within the retina. In contrast, microaneurysms refer to the tiny areas of swelling in the retina's blood vessels. Exudates in the retina, also known as pus, are fluids that leak blood out of the vessels into neighboring tissues. Venous beading is the act of irregular dilation of the venules (a type of vein) in the retina. IrMA, which are one of the most distinguishable features of diabetic retinopathy, are the abnormalities of the blood vessels in the eye that supply the retina.

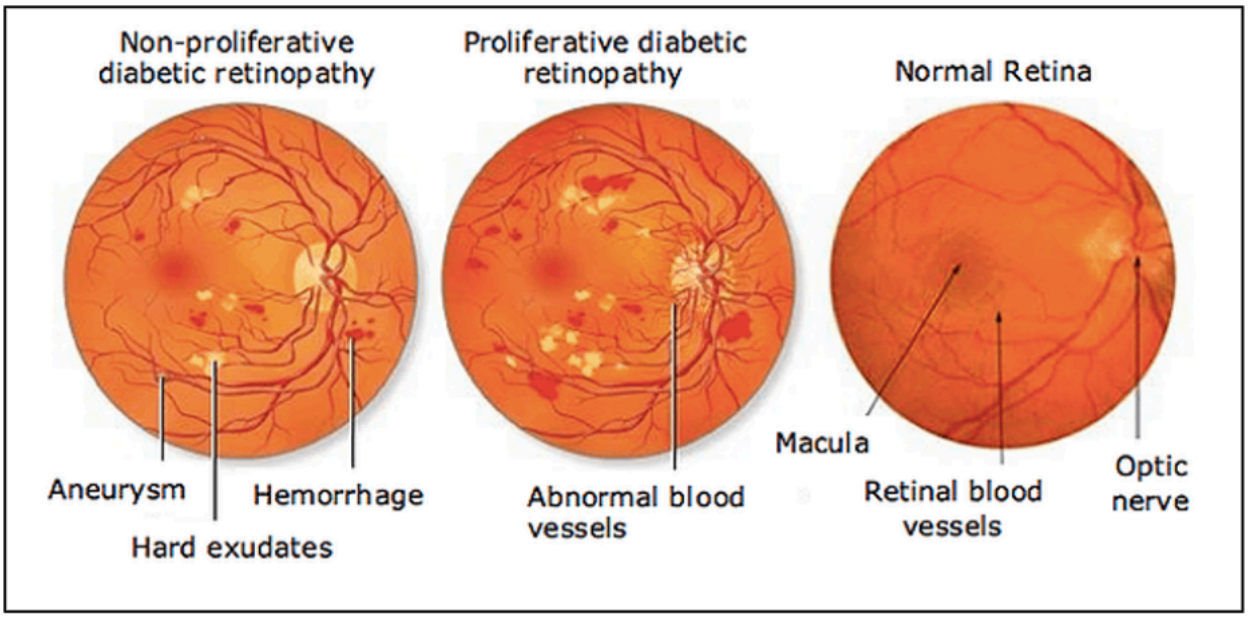

Figure 3[3]: A 2D layout of the diabetic human retina, comparing the diabetic traits found on an eye with NPDR, PDR, and no PDR .

Diabetic retinopathy diagnosis typically involves screenings performed by eye care professionals or ophthalmologists. However, the scope of this project aims to incorporate the application of Vision Transformers (ViTs) for the detection of DBRP, along with its four variants.

Transformers were initially introduced in the 2017 research paper “Attention is All You Need”. After finding widespread use in Natural Language Processing, the Transformer was adapted for computer vision in the 2021 conference research paper “An Image is Worth 16x16 Words”, ultimately yielding Vision Transformers

The Vision Transformer (ViT) has recently emerged as a competitive alternative to CNNs. Prior to the release of ViT, Convolutional Neural Networks (CNNs) were a major staple in the world of computer vision (CV). Now, ViTs have been proven to outperform state of the art CNNs by nearly four times in terms of computational efficiency and accuracy. They find applications in an extensive

amount of CV tasks, including image recognition, object detection, segmentation (classifying parts of an image), etc.

Vision transformers are a type of computer transformer model that uses self-attention mechanisms and computer vision, over patches of an image, in order to process/classify images. This means that the transformer will take in an input image, and output a class prediction using class labels. The reason for which the ViT is so unique in comparison to other CV models is because it does this without any convolutional layers, unlike the CNN. Instead, it uses the attention layers that are prevalent within the natural language processing applications.

In order to understand how the typical Vision Transformer works, it is fundamental that we first learn what self-attention is. In computer vision, self-attention is a mechanism that allows each part of an input image to assess and weigh its significance in relation to other parts (these parts being called patch embeddings). Significance in this case refers to the relevance of each part in relation to the entire input sequence. For instance, consider an image of dimensions 4096 x 2160 pixels. The computational cost of processing this entire image is relatively high, so in order to make this process more efficient, we use the self-attention layers which are prevalent within the Vision Transformer. (See Figure 3 below):

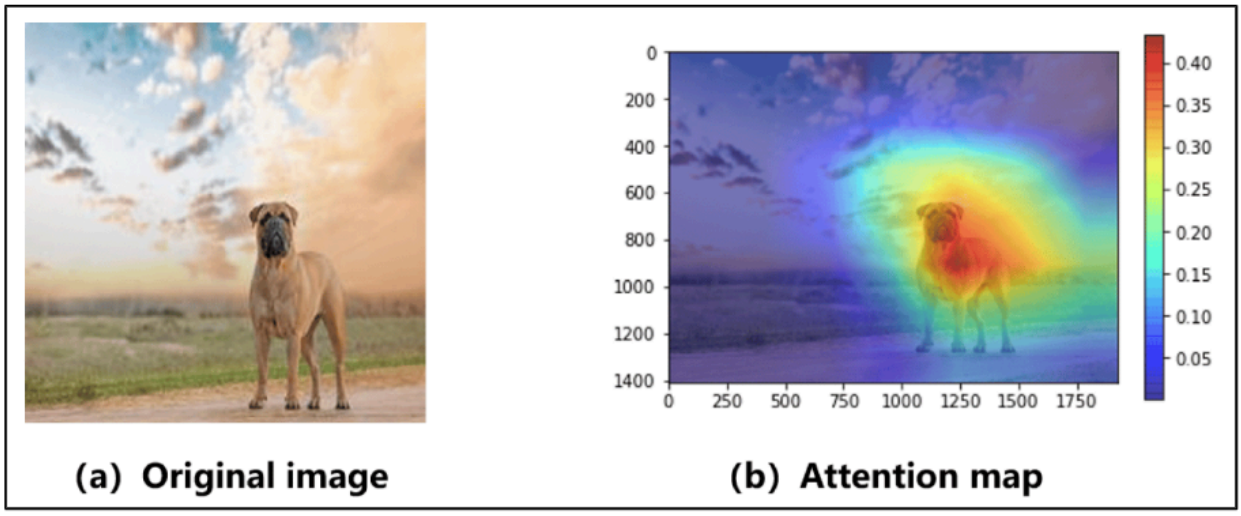

Figure 4[4]: An attention heatmap visualization (b), which highlights the regions of an image of a dog (a) that are most relevant to the output prediction.

Figure 4 perfectly illustrates the overall scope on how self-attention is applied within the Vision Transformer output phase/architecture. Now that we have learned the basics of self-attention, we can now understand how the ViT works. The Vision Transformer architecture comprises several parts, as shown below:

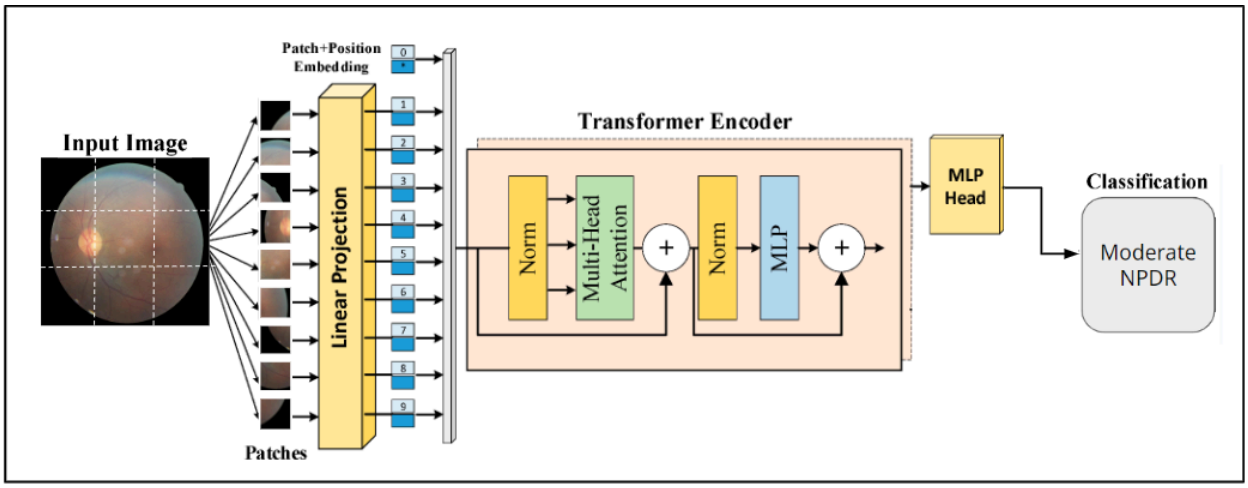

Figure 5: The ViT model overview. The input image is broken into patches, linearly embedded, and then flattened into a linear projection. After that, we add position embeddings and feed the resulting vectors into the Transformer encoder. Finally, we train the model with image labels on a large dataset, and fine-tune the dataset for image classification.

To train a Vision Transformer for image classification, the input images are first fed into the transformer as data. In this architecture, an input image typically consists of 3 color channels, these being red, green and blue. These images contain a width and height as well as a batch dimension, if there is more than one image in the dataset. In the context of computer vision, a batch dimension, more commonly referred to as a 'batch size' is the number of images being processed at once. A higher batch size typically results but using too high of a batch size can result in many unintended errors such as overfitting, where the model generalizes too well.

Following the image being inputted, image patching is then performed. This process breaks down an image into a grid of fixed-sized square images which are processed as a sequence of tokens, called patch embeddings. In order to perform the patching, we must first resize each of the data points/images. This gives each of the images the same dimensions, ensuring uniformity. The smaller the patch size is, the better the model tends to perform (in terms of accuracy). However, since a smaller patch size results in more tokens to process, this will ultimately make your model run slower. It is also important to note that the patch size you are using has to be a multiple of your image dimensions. For example, if you were to resize your images to 144x144, it would not be possible to utilize a patch size of 32, as it is not a divisor of 144. In the model being presented in this paper, a patch size of 16 is utilized for images of dimensions 224x224.

Next, we linearly embed the 2D vectors (patches), flattening them into a linear projection, or a 1D array. This allows the computer to process them as a sequence of tokens, rather than treating the entire image as a single entity. By doing this, the

ViT model will be able to process the information sequentially and capture the spatial relationships between different patches, enabling effective representation learning for the input data. Additionally, since the standard ViT receives a 1D sequence of token embeddings (in the form of numerical vectors) as input data, in order to handle 2D images, we must reshape the image into a sequence of flattened 2D patches.

After flattening the patches, we pass the resulting vectors through the Input Embedding block, so it can be later fed into the Transformer Encoder. The Input Embedding block is a fully connected neural network, in which each of the three input color values in the image are put into nodes/neurons and passed through some linear layers. This block returns position embeddings, which add information about the relative position of the image patches in the sequence.

The Transformer Encoder block is one of the most crucial elements in both the Transformer and ViT architecture. There are multiple layers within the Transformer Encoder made up of several nodes/deep learning neurons, and they each comprise three major processing elements:

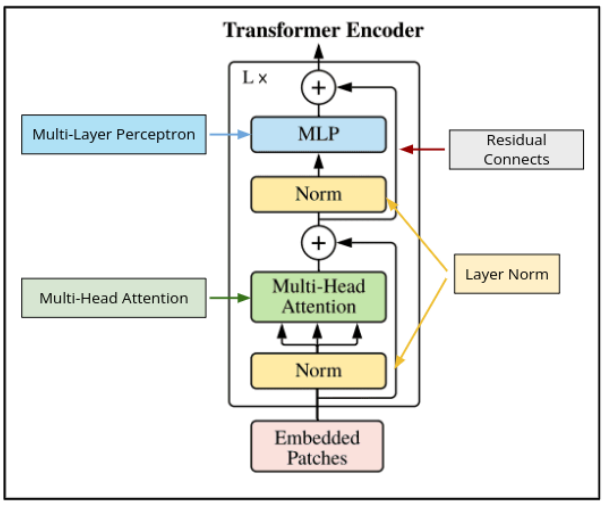

Figure 6[5]: The Transformer Encoder block.

- Layer Normalization (LN)

- Multi-Head Attention Network (MSP)
- Multi-Layer Perceptrons (MLP)

Layer Norm is applied before each layer in the transformer encoder, as shown in Figure 5. The Normalization layer normalizes the input patch embeddings by subtracting the mean of the number of datapoints in each class from each element in the input tensor, and dividing the result by the standard deviation of the input tensor. Normalization ultimately helps stabilize the training process

The Multi-Head Attention Network layer is arguably one of the most vital components of a ViT. In this layer, the main attention mechanisms of the ViT are held. Patches are split into multiple heads, which send a variety of attention scores, represented in matrices (which are later simplified in the output tensor). In this context, attention scores indicate how much each patch in an image should attend to other patches. This also allows multiple attention operations to be performed on the input embeddings. In addition, each head in the MSP computes its own attention scores converting the embedding tokens into keys (k), queries (q) and values (v). These three transformations of the input embeddings essentially help compute the relationships between the input embedding tokens.

In the Multilayer Perceptron, each token representation from the attention mechanism is passed through a neural network, called the Feed-Forward layer.

After the embeddings are passed through the Transformer Encoder block, they are then finally proceeded to formulate a class prediction. This is accounted for by adding an extra learnable “classification token” to the sequence of vectors derived from the patch embeddings (which is represented as embedding 0 in Figure 5.)

Finally, the model is trained on a set amount of epochs as well as fine tuned, where a model weights trained beforehand (or pretrained weights) are further trained to accommodate for our dataset.

Observations:

Losses in machine learning refer to the average error over data (essentially the opposite of accuracy.) During my training phase, I get a peak accuracy of ~95% and a loss of ~0.24. For the validation of the same epoch, I got an accuracy of ~72% and loss of ~1.25. These results were obtained after training the model for 30 epochs. Below is a graph representing the training loss between my testing and validation phase:

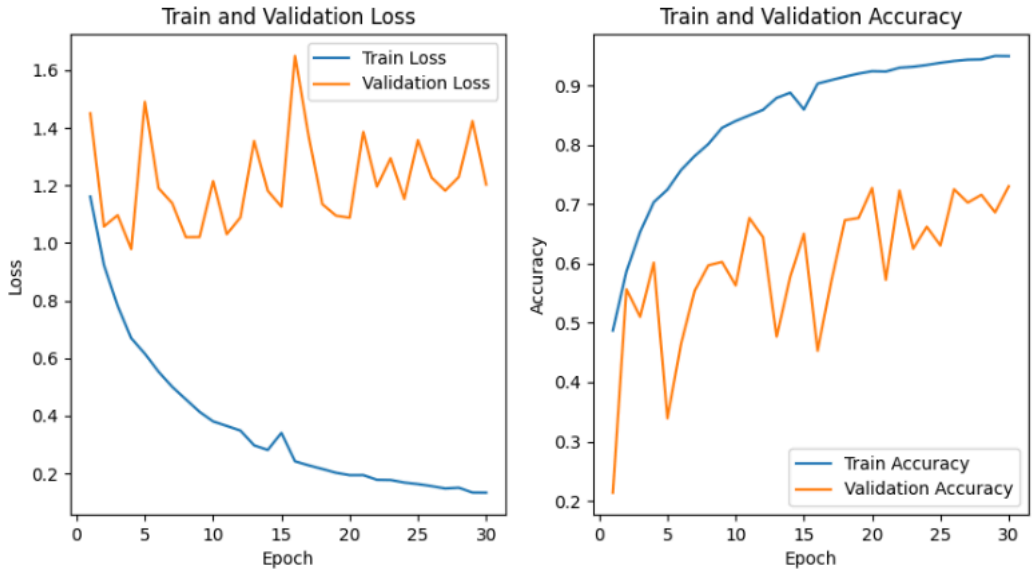

Figure 7: A set of graphs representing the losses and accuracies obtained from each epoch in the model (as per training and validation).

In the accuracy graphs provided, it is evident that the losses differed dramatically from each other. The accuracies on the other hand, had a bit of a similar curve. The main distinguishing feature between these two are that the

validation loss and accuracy fluctuates way more in comparison to the steady train accuracy/loss.

Conclusion:

In regard to my inquiry whether it is possible to use Vision Transformers to diagnose diabetic retinopathy, I have concluded they indeed have the capability to be used as a screening tool for the vision impairment disease.

Materials:

- Computer with Pytorch installed
- Kaggle Notebook
- Images of the diabetic retina separated into five classes:
    - No DR
    - Mild NPDR
    - Moderate NPDR
    - Severe NPDR
    - PDR

Procedure:

Data was Downloaded & Structured

1. Kaggle notebook was created, including a folder for my data
2. Kaggle datasets with diabetic fundus images were downloaded.

3. Downloaded images were divided into five subdirectories, corresponding to the types of PDR.
4. Downloaded images were placed into my training folder.
5. Data was split into an 80-10-10 ratio between my training, validation and testing data.

The Vision Transformer was Created

1. Necessary libraries were downloaded onto the notebook using the ‘pip’ and ‘import’ commands.
2. Vision Transformer architecture was implemented and debugged using Pytorch.

Training, Testing and Validation Phase

1. Model was trained, tested and validated using 28 epochs, using a learning rate of 3e-4.
2. The GUI verification system for testing was created using Tkinter.

References

[1] https://www.nei.nih.gov/learn-about-eye-health/healthy-vision/how-eyes-work

[2]https://www.researchgate.net/figure/A-qualitative-comparison-of-normal-vision-and-vision-affected-by-Diabetic-Retinopathy-fig1-350930649

[3]https://eyedoc.sg/diabetic-eye-conditions/

[4]https://www.mdpi.com/2076-3417/12/8/3846

[5]https://www.researchgate.net/figure/Transformer-Encoder-block-by-36-We-follow-this-paradigm-to-implement-the-Transformer_fig2_372827716

(n.d.). Colaboratory. Retrieved February 22, 2024, from https://colab.research.google.com/drive/1P9TPRWsDdqJC6IvOxjG2_3QlgCt59P0w?usp=sharing#scrollTo=SHZBgsNCdCRk

(n.d.). Wikipedia. Retrieved February 22, 2024, from https://www.sciencedirect.com/topics/computer-science/attention-sco

(2016, November 13). analyticsvidhya. Retrieved February 20, 2024, from https://www.analyticsvidhya.com/blog/2021/03/an-image-is-worth-16x16-words-transformers-for-image-recognition-at-scale-vision-transformers

Aqueous and Vitreous Humor: Anatomy, Function & Location. (2022, December 27). Cleveland Clinic. Retrieved February 21, 2024, from https://my.clevelandclinic.org/health/body/24611-aqueous-humor-vitreous-humor

Asymmetric diabetic retinopathy - PMC. (n.d.). NCBI. Retrieved February 20, 2024, from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8725155/

At a glance: Diabetic Retinopathy. (2023, November 15). National Eye Institute. Retrieved February 20, 2024, from https://www.nei.nih.gov/learn-about-eye-health/eye-conditions-and-diseases/diabetic-retinopathy

At a glance: Diabetic Retinopathy. (2023, November 15). National Eye Institute. Retrieved February 20, 2024, from https://www.nei.nih.gov/learn-about-eye-health/eye-conditions-and-diseases/diabetic-retinopathy

Blindness and vision impairment. (2023, August 10). World Health Organization (WHO). Retrieved February 20, 2024, from https://www.who.int/news-room/fact-sheets/detail/blindness-and-visual-impairment

Boesch, G. (n.d.). *Vision Transformers (ViT) in Image Recognition - 2024 Guide - viso.ai*. Viso Suite. Retrieved February 20, 2024, from https://viso.ai/deep-learning/vision-transformer-vit

Dataset from fundus images for the study of diabetic retinopathy. (2021, June). Science Direct. Retrieved February 22, 2024, from https://www.sciencedirect.com/science/article/pii/S2352340921003528?via%3Dihub

Diabetic retinopathy. (n.d.). NHS. Retrieved February 20, 2024, from https://www.nhs.uk/conditions/diabetic-retinopathy/

Diabetic retinopathy: An update - PMC. (n.d.). NCBI. Retrieved February 20, 2024, from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2636123/

Diabetic retinopathy: An update - PMC. (n.d.). NCBI. Retrieved February 20, 2024, from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2636123/

Diabetic Retinopathy Arranged. (n.d.). Kaggle. Retrieved February 22, 2024, from https://www.kaggle.com/datasets/amanneo/diabetic-retinopathy-resized-arranged?rvi=1

Diabetic retinopathy - Symptoms & causes. (n.d.). Mayo Clinic. Retrieved February 20, 2024, from https://www.mayoclinic.org/diseases-conditions/diabetic-retinopathy/symptoms-causes/syc-20371611

11.8. Transformers for Vision — Dive into Deep Learning 1.0.3 documentation. (n.d.). Dive into Deep Learning. Retrieved February 22, 2024, from https://d2l.ai/chapter_attention-mechanisms-and-transformers/vision-transformer.html

Google Research, Brain Team. (2021, June 3). *An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale*. arxiv. Retrieved February 20, 2024, from https://arxiv.org/pdf/2010.11929.pdf

Google Research, Brain Team. (2022, June 23). *How to train your ViT? Data, Augmentation, and Regularization in Vision Transformers*. arxiv. Retrieved February 20, 2024, from https://arxiv.org/pdf/2106.10270.pdf

How the Eyes Work. (2022, April 20). National Eye Institute. Retrieved February 20, 2024, from https://www.nei.nih.gov/learn-about-eye-health/healthy-vision/how-eyes-work

Kim, J. E., & Hsu, J. (2023, July 15). *Diabetic Retinopathy - EyeWiki*. EyeWiki. Retrieved February 20, 2024, from https://eyewiki.aao.org/Diabetic_Retinopathy

Lloyd, W. C. (2023, May 25). *Microaneurysms in Your Eye from Diabetic Retinopathy*. Healthline. Retrieved February 20, 2024, from https://www.healthline.com/health/diabetes/diabetic-retinopathy-microaneurysms

MA, C. (n.d.). *Anatomy, Head and Neck: Eye Retina - StatPearls*. NCBI. Retrieved February 20, 2024, from https://www.ncbi.nlm.nih.gov/books/NBK542332/

Neuroanatomy, Retina - StatPearls. (2023, August 8). NCBI. Retrieved February 20, 2024, from https://www.ncbi.nlm.nih.gov/books/NBK545310/

Pulfer, B. (2022, February 3). *Vision Transformers from Scratch (PyTorch): A step-by-step guide*. Medium. Retrieved February 22, 2024, from https://medium.com/mlearning-ai/vision-transformers-from-scratch-pytorch-a-step-by-step-guide-96c3313c2e0c

Rastogi, R. (2023, February 8). *Papers Explained 25: Vision Transformers | by Ritvik Rastogi | DAIR.AI*. Medium. Retrieved February 16, 2024, from

https://medium.com/dair-ai/papers-explained-25-vision-transformers-e286ee8bc06b

Retina: Anatomy, Function & Common Conditions. (2022, April 7). Cleveland Clinic. Retrieved February 20, 2024, from https://my.clevelandclinic.org/health/body/22694-retina-eye

Shah, D. (2022, December 15). *Vision Transformer: What It Is & How It Works [2023 Guide]*. V7 Labs. Retrieved February 20, 2024, from https://www.v7labs.com/blog/vision-transformer-guide

Simple Anatomy of the Retina - Webvision. (2005, May 1). NCBI. Retrieved February 20, 2024, from https://www.ncbi.nlm.nih.gov/books/NBK11533/

Vision Transformer Quick Guide - Theory and Code in (almost) 15 min. (2023, July 4). YouTube. Retrieved February 22, 2024, from https://www.youtube.com/watch?v=j3VNqtJUoz0

What is the macula? (n.d.). Macular Society. Retrieved February 20, 2024, from https://www.macularsociety.org/macular-disease/macula

Wydmanski, W. (2022, December 3). *Self attention vs attention in transformers | MLearning.ai*. Medium. Retrieved February 20, 2024, from https://medium.com/mlearning-ai/whats-the-difference-between-self-attention-and-attention-in-transformer-architecture-3780404382f3

Zuppichini, F. (2021, January 1). *Transformers VisionTransformer*. Towards Data Science. Retrieved February 22, 2024, from https://towardsdatascience.com/implementing-visualttransformer-in-pytorch-184f9f16f632