# DIABETIC RETINOPATHY GRADING USING VISION TRANSFORMERS

*Thesis Submitted in partial fulfillment of the requirements of the degree of*

**Master of Technology**

**In**

**Computer Science**
**(Specialization in Software Engineering)**

*by*

**ANUSREE M NAIR**
**48021003**

Under the guidance of

**Dr. Bijoy Antony Jose**

**DEPARTMENT OF COMPUTER SCIENCE**
**COCHIN UNIVERSITY OF SCIENCE AND TECHNOLOGY**

**APRIL 2023**

DEPARTMENT OF COMPUTER SCIENCE
COCHIN UNIVERSITY OF SCIENCE AND TECHNOLOGY
ERNAKULAM, KOCHI-682022

# CERTIFICATE

This is to certify that the dissertation work entitled **DIABETIC RETINOPATHY GRADING USING VISION TRANSFORMERS** is a bonafide record of work carried out by ANUSREE M NAIR (48021003) submitted to the Department of Computer Science in partial fulfillment of the requirements for the award of the degree of Master of Technology in Software Engineering at Cochin University of Science and Technology during the academic year 2023.

Project Guide

**Dr. Bijoy A Jose**
Associate Professor
Dept. of Computer Science
CUSAT

**Dr. Philip Samuel**
Professor and HoD
Dept. of Computer Science
CUSAT

DATE : 25 April 2023

Office seal

# DECLARATION

I declare that the work presented in this dissertation titled **DIABETIC RETINOPATHY GRADING USING VISION TRANSFORMERS** represents my ideas in my own words and where others' ideas or words have been included, I have adequately cited and referenced the original sources. I also declare that I have adhered to all principles of academic honesty and integrity and have not misrepresented or fabricated or falsified any idea/data/fact/source in my submission. I understand that any violation of the above will be cause for disciplinary action by the University and can also evoke penal action from the sources which have thus not been properly cited or from whom proper permission has not been taken when needed.

Signature:

Name: ANUSREE M NAIR    Date: 25 April 2023
Reg No: 48021003    Place: Ernakulam

# ACKNOWLEDGEMENT

I take this opportunity to thank the supreme being, the source of all knowledge whose blessings are my guiding light in any venture we take up. I express my gratitude in short words to Dr.Bijoy A Jose, my project guide, who guided me and constantly helped with their inputs and suggestion, without which I could not have pursued this project theway it is working today. And gratitude to Dr.Philip Samuel, head of the Department of Computer science, for all kinds of encouragement extended to us.

Anusree M Nair

# ABSTRACT

Diabetic Retinopathy(DR) is a disease caused by blood leakage into the retinal tissues due to prolonged diabetic condition which eventually leads to permanent vision loss. It is the leading risk factor for poor vision in patients aged 25 to 74. Early DR detection helps avoid vision loss. Manual DR grading using fundus images is time-consuming and requires expert ophthalmologists. Computer-aided diagnosis using deep learning models has shown promising results in recent years, with convolutional neural networks (CNNs) being the most commonly used architecture. However, vision transformers have emerged as a viable alternative, offering the potential for improved accuracy and efficiency in DR grading. To address this problem, Vision Transformer (ViT) based DR severity classification method is proposed. In this approach, the fundus images are initially divided into non-overlapping patches to retain location information. Then, the flattened patches are converted into sequences before going through a linear and positional embedding process. The generated sequence is then fed into several multi-head attention layers, which produce the recognition output. It is tested on Kaggle(APTOS 2019) database. The results obtained are better than the convolutional neural networks and state-of-the-art. Finally, we discuss and outline some of the challenges and future directions in this field. Our aims is to provide a comprehensive overview of the use of vision transformers for DR grading and to highlight the potential of this approach for improving the diagnosis and management of DR.

**Keywords: Diabetic Retinopathy, VisionTransformers(ViT), Fundus Images**

# Contents

# List of Figures

# Abbreviations

| | |
|---|---|
| CNN | Convolutional Neural Network |
| DR | Diabetic Retinopathy |
| MSA | Multi-headed Self Attention |
| ViT | Vision Transformers |
| MLP | Multi Layer Perceptron |
| PDR | Proliferative Diabetic Retinopathy |

# Chapter 1

# Introduction

Diabetic retinopathy is a medical condition in which damage occurs to retina of eye due to diabetic mellitus. It is caused by prolonged high blood glucose damaging the small blood vessels of the retina. It’s the most recurrent cause of irreversible blindness and is highly likely when diabetes is poorly controlled. The primary cause of DR development and its consequences are avoiding the precautionary measure for blood sugar control and a healthy lifestyle. Its main symptom is that it produces mutilation of blood vessels of retina.

The DR has two major types: the Non-Proliferative Diabetic Retinopathy (NPDR) and Proliferative Diabetic Retinopathy (PDR). The DR in the early stages is called NPDR which is further divided into Mild, Moderate, and Severe stages. The mild stage has one micro-aneurysm (MA), a small circular red dot at the end of blood vessels. In the Moderate stage the MAs rapture into deeper layers and form a flame-shaped hemorrhage in the retina. The severe stage contains more than 20 intra-retinal hemorrhages in each of the four quadrants, having definite venous bleeding with prominent intra-retinal microvascular abnormalities. PDR is the advanced stage of DR which leads to neovascularization, a natural formation of new blood vessels in the form of functional microvascular networks that grow on the inside surface of the retina. Vision transformers, a type of neural network architecture, have recently shown promising results in computer vision tasks such as image classification and object detection. They can be used for DR grading by analyzing retinal images and predicting the severity of the disease.

In the context of DR grading, the vision transformer model is trained on a large dataset of retinal images that have been annotated with the corresponding severity of DR. The model is then used to predict the DR severity of new retinal images, which can help doctors and healthcare professionals in the diagnosis and management of DR.

Compared to traditional methods of DR grading, which rely on manual inspection by trained professionals, the use of vision transformers can potentially improve the accuracy and efficiency of DR grading. However, further research and validation are needed to fully evaluate the performance of vision transformers for DR grading and to ensure their safe and effective use in clinical practice.

## 1.1 Motivation:

Diabetic Retinopathy (DR) is an ophthalmic disease that damages retinal blood vessels. DR causes impaired vision and may even lead to blindness if it is not diagnosed in early stages. Manual diagnosis is error-proned and tedious. The number of patients are increasing day by day and hence its high time to find a solution. The motivation for using vision transformers for diabetic retinopathy grading comes from the need to improve the accuracy and efficiency of this critical process. Diabetic retinopathy is a common complication of diabetes that can lead to vision loss and blindness if left untreated. Early detection and grading of the disease are crucial for effective treatment and management.

Traditional methods of DR grading rely on manual inspection by trained professionals, which can be time-consuming, costly, and prone to inter-observer variability. Computer-aided diagnosis using deep learning models has shown promising results in recent years, with convolutional neural networks (CNNs) being the most commonly used architecture. However, vision transformers have emerged as a viable alternative, offering the potential for improved accuracy and efficiency in DR grading.

The motivation for using vision transformers is also driven by the increasing availability of large annotated datasets of retinal images, as well as the availability of powerful computing resources that can handle the high computational demands of training and testing these models. The use of vision transformers for DR grading can potentially provide a faster and more accurate diagnosis, allowing for earlier treatment and management of the disease.

## 1.2 Problem Statement:

Diabetic Retinopathy(DR) is a serious disease that may eventually lead to permanent blindness ruining one's life. Hence a solution to this issue is a must to detect the disease at an early stage so that treatment does not get delayed. Therefore the aim is to detect DR and classify different stages of diabetic retinopathy using vision transformers. In recent years, machine learning techniques such as deep learning have shown promising results in the detection and diagnosis of DR. Vision transformers, a type of deep learning model that uses self-attention mechanisms to process input data, have shown particularly good performance in image classification tasks.

The aim of using vision transformers in the detection and classification of DR is to accurately identify the different stages of the disease using retinal images. This can be achieved by training a vision transformer model on a large dataset of retinal images, along with corresponding labels indicating the presence and severity of DR.

The trained model can then be used to predict the presence and severity of DR in new retinal images, allowing for early detection and treatment of the disease. This can help prevent further damage to the retina and preserve vision for people with diabetes.

In summary, the use of vision transformers in the detection and classification of DR has the potential to improve the accuracy and speed of diagnosis, leading to better outcomes for people with diabetes who are at risk of developing this serious condition.

## 1.3 Dataset Description:

Dataset was downloaded, trained and tested with the publicly available databases Kaggle.com(APTOS 2019).

# Chapter 2
# Literature Survey

## 2.1 Weakly supervised localisation of diabetic retinopathy lessions in retinal fundus images:-

Convolutional neural networks (CNNs) have shown remarkable results in image classification and detection, including the medical image domain. However, medical experts have reservations about relying solely on the non-linear multilayer structure of CNNs to make decisions. Recently, researchers have developed methods to help users understand the regions in an image that CNNs use to make classification decisions. These methods can build trust in CNN predictions, but their efficacy with medical image data is still uncertain, as decision-making in medical images often involves several scattered lesion areas. This study uses the DiaretDB1 dataset to demonstrate that CNNs can detect various lesion areas in retina images with high accuracy for the diagnosis of diabetic retinopathy, even outperforming supervised methods. The proposed CNN model, called Referable Diabetic Retinopathy (RDR), achieves excellent performance on binary classification of different disease stages. Detecting stage-1 features and lesions can be challenging because they may vary significantly or have few sample images. As DR progresses through five stages, early detection of any disease is crucial to prevent it from reaching stage 4, the most dangerous and incurable stage (PDR).

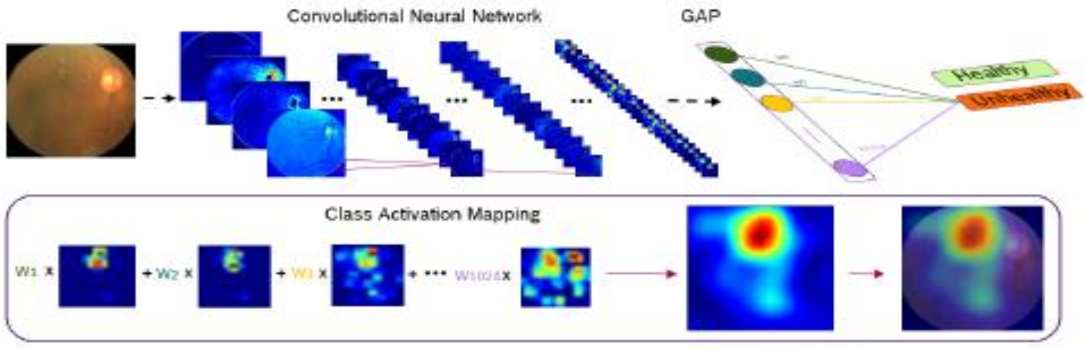


Figure 2.1: CNN setup for generating CAMs[1]

## 2.2 Lesion detection and Grading of Diabetic Retinopathy via Two stages Deep Convolutional Neural Networks:-

An automatic diabetic retinopathy (DR) analysis algorithm based on two-stages deep convolutional neural networks (DCNN) is used. Compared to existing DCNN-based DR detection methods, the proposed algorithm have the following advantages: (1) Our method can point out the location and type of lesions in the fundus images, as well as giving the severity grades of DR. Moreover, since retina lesions and DR severity appear with different scales in fundus images, the integration of both local and global networks learn more complete and specific features for DR analysis. (2) By introducing imbalanced weighting map, more attentions will be given to lesion patches for DR grading, which significantly improve the performance of the proposed algorithm. In this study, we label 12; 206 lesion patches and re-annotate the DR grades of 23; 595 fundus images from Kaggle competition dataset. Under the guidance of clinical ophthalmologists, the experimental results show that our local lesion detection net achieve comparable performance with trained human observers, and the proposed imbalanced weighted scheme also be proved to significantly improve the capability of our DCNN based DR grading algorithm.

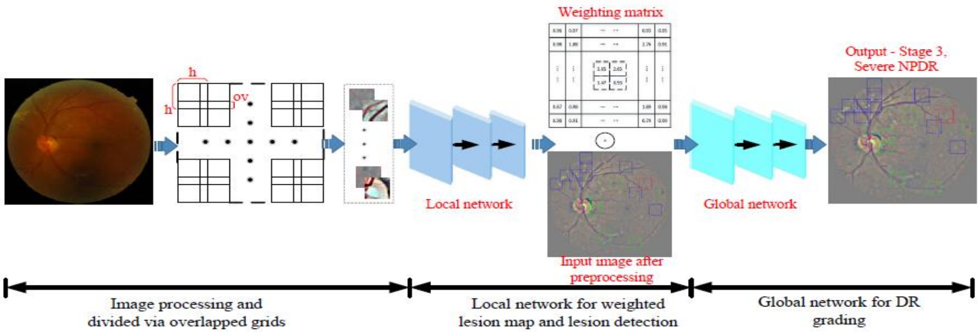

Weighting matrix
h
h
ov
Output - Stage 3, Severe NPDR
Local network
Global network
Input image after preprocessing
Image processing and divided via overlapped grids
Local network for weighted lesion map and lesion detection
Global network for DR grading

Figure 2.2: Main workflow of mentioned algorithm[2]

## 2.3 A deep learning ensemble approach for diabetic retinopathy detection:-

Diabetic Retinopathy (DR) is an ophthalmic disease that damages retinal blood vessels. DR causes impaired vision and may even lead to blindness if it is not diagnosed in early stages. DR has five stages or classes, namely normal, mild, moderate, severe and PDR (Proliferative Diabetic Retinopathy). Normally, highly trained experts examine the colored fundus images to diagnose this fatal disease. This manual diagnosis of this condition (by clinicians) is tedious and error-prone. Therefore, various computer vision-based techniques were proposed to automatically detect DR and its different stages from retina images. However, these methods were unable to encode the underlying complicated features and can only classify DR's different stages with very low accuracy particularly, for the early stages. In this research, the publicly available Kaggle dataset of retina images are used to train an ensemble of five deep Convolution Neural Network (CNN) models (Resnet50, Inceptionv3, Xception, Dense121, Dense169) to encode the rich features and improve the classification for different stages of DR. The experimental results show that the proposed model detects all the stages of DR unlike the current methods and performs better compared to state-of-the-art methods on the same Kaggle dataset.

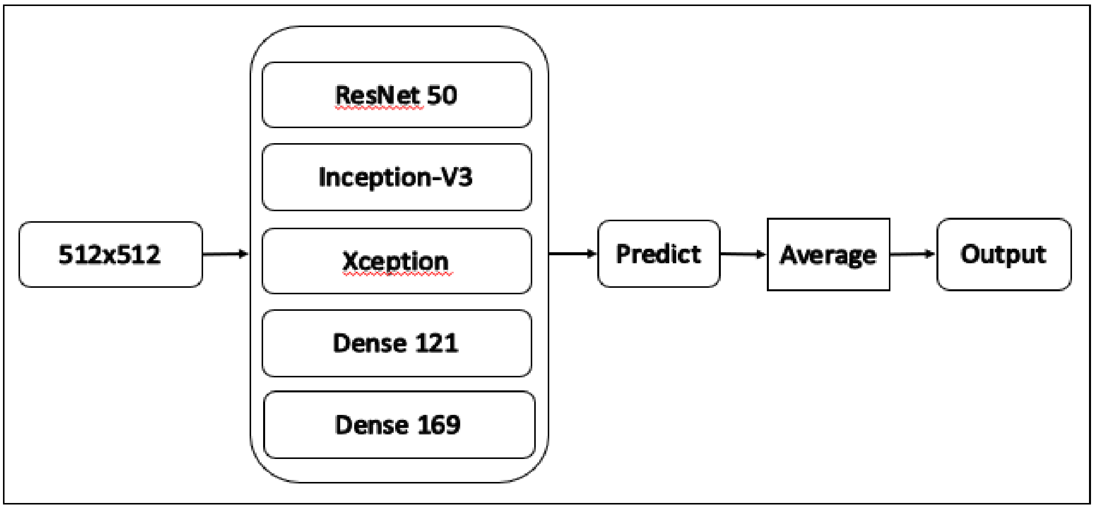


Figure 2.3:Ensemble Method Architecture[8]

## 2.4 A Deep Learning Approach for Diabetic Retinopathy detection using Transfer Learning:-

Diabetic Retinopathy is a primary complication of diabetes which more often than not, affects both eyes and anyone with type-1 or type-2 diabetes can develop it. A Diabetic patient should undergo eye tests periodically as the pace of development of this condition is slow. A Dataset of fundus Photographs of retina is considered. Thus, there is a notable value in automatically categorizing the fundus Photographs. Therefore, to get a consolidated and objective medical diagnosis, a transfer learning based approach for Diabetic Retinopathy categorization is adopted. The resizing of the dataset is performed, which converts the varied images into 224x224 format. The images are augmented using AUGMIX and pooled using GeM. Then, pre-trained models are used, namely SEResNeXt32x4d and EfficientNetb3. The pretraining of the aforementioned neural networks has been done on the ImageNet dataset. Then, the Diabetic Retinopathy images are migrated to these models. Based on the dataset already available, the output is ultimately split up into 5 levels according to the seriousness of the degree of DR. The experimental results show that the training accuracy of this method can reach as high as 0.91. Hence, the retina images of the Diabetic and Healthy patients can be easily classified using this methodology, consequently reducing the number of reviews by medical professionals.

# Chapter 3

# Proposed System

## 3.1 Dataset

The model is trained using kaggle dataset. The data was present in the form of images of different stages of Diabetic retinopathy. The dataset consist of high resolution retinal images of different stages of diabetic retinopathy including normal, mils, moderate, severe and proliferative. The images were captured using various camera models and imaging techniques, which provide a diverse dataset.

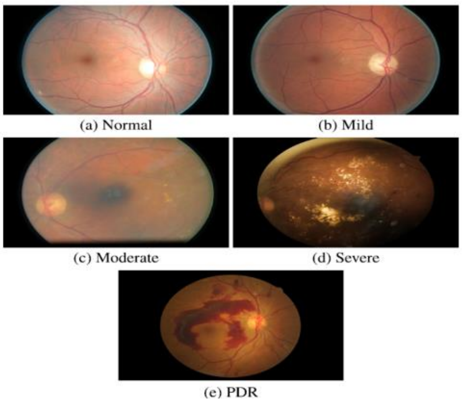


Figure 3.1: Different Stages of DR

## 3.2 Data Preprocessing

The data pre-processing was performed to prepare the dataset ready for training. The dataset provided a diverse range of retinal images with different sizes, resolutions, and imaging techniques. To prepare the dataset for training, several pre-processing steps were performed, including data augmentation, resizing, and removal of unwanted data.

## 3.3 Methodology:

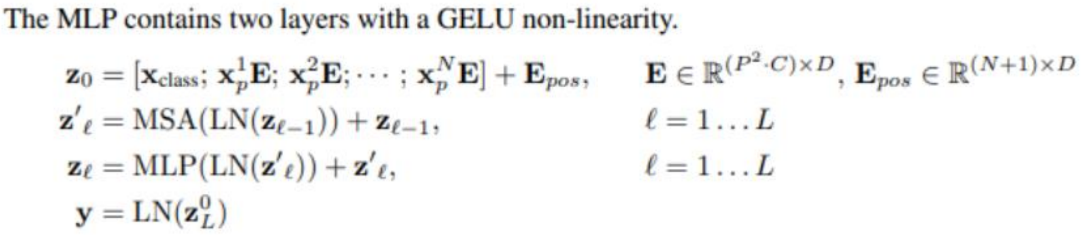
In the field of computer vision, convolutional architectures have been the most widely used approach for image analysis. However, recent advancements in natural language processing (NLP) have led to the exploration of combining CNN-like architectures with self-attention, and some models have even replaced convolutions entirely. These models, while theoretically efficient, have not yet been scaled effectively on modern hardware accelerators due to the use of specialized attention patterns. Hence, in large-scale image recognition, classic ResNet-like architectures are still considered state-of-the-art. Inspired by the success of the Transformer model in NLP, conducted experiments to apply a standard Transformer directly to images, with minimal modifications. To achieve this, divided an image into patches and provided the sequence of linear embeddings of these patches as input to the Transformer. Similar to NLP, treated image patches as tokens (words) and trained the model for image classification in a supervised manner.

## 3.4 Vision Transformers:

The architecture used here is Vision Transformer. In the domain of natural language processing, Transformers are recognized as state-of-the-art models, which opposing to typical convolutional neural networks (CNNs) do not rely on convolution layers. Instead, Transformers employ multi-head attention mechanisms as the main building block to capture long-range contextual relations between image pixels. Recently, CNNs dominated the deep learning solutions for diabetic retinopathy grade recognition. However, spurred by the advantages of Transformers, a Transformer-based method that is appropriate for recognizing the grade of diabetic retinopathy is adopted here.

Transformer model relies on self-attention mechanism raised a big challenge to use it for computer vision tasks. The self-attention mechanism is the reason why Transformer-based models can differentiate the semantic meaning of a word used in different contexts. For example, a BERT-a type of vision transformer model can distinguish the meaning of the word *'park'* in sentences *'They park their car in the basement'* and *'She* *walks her dog in a park'* due to the virtue of this self-attention mechanism. However, there is one problem with self-attention: it's a computationally expensive operation as it requires each token to attend every other token in a sequence. Using self-attention mechanism on image data, then each pixel in an image would need to able compared to every other pixel. The problem is, if it increase the pixel value by one, then the computational cost would increase quadratically. This is simply not feasible if we have an image with a reasonably large resolution.

In order to overcome this problem, ViT introduces the concept of splitting the input image into patches. Each patch has a dimension of 16 x 16 pixels. Let's say that we have an image with the dimension of 48 x 48 pixels, then the patches of our image are formed by Reshaping the input image that has a size of height x width x channel into a sequence of flattened 2D image patches with a size of no.of patches x (patch_size^2.channel) . Then, flattened patches are projected into a basic linear layer to get the embedding of each patch. Flattening process is basically the vector embedding of each patch. This is similar to token embeddings in many Transformer-based language models. A trainable linear projection is used to map the flattened patches to a D-dimensional vector space, referred to as patch embeddings. A learnable embedding is then pre-pended to the sequence of patch embeddings, which serves as the image representation output by the Transformer encoder. During both pre-training and fine-tuning, a classification head is attached to this representation. Position embeddings are added to the patch embeddings to retain positional information, using standard 1D learnable position embeddings. The Transformer encoder consists of alternating layers of multi-headed self-attention and MLP blocks, with layer normalization applied before every block and residual connections after every block.



The MLP contains two layers with a GELU non-linearity.

$$\begin{aligned}
\mathbf{z}_0 &= [\mathbf{x}_{\text{class}};\, \mathbf{x}_p^1\mathbf{E};\, \mathbf{x}_p^2\mathbf{E}; \cdots;\, \mathbf{x}_p^N\mathbf{E}] + \mathbf{E}_{pos}, && \mathbf{E} \in \mathbb{R}^{(P^2\cdot C)\times D},\, \mathbf{E}_{pos} \in \mathbb{R}^{(N+1)\times D} \\
\mathbf{z'}_\ell &= \mathrm{MSA}(\mathrm{LN}(\mathbf{z}_{\ell-1})) + \mathbf{z}_{\ell-1}, && \ell = 1\ldots L \\
\mathbf{z}_\ell &= \mathrm{MLP}(\mathrm{LN}(\mathbf{z'}_\ell)) + \mathbf{z'}_\ell, && \ell = 1\ldots L \\
\mathbf{y} &= \mathrm{LN}(\mathbf{z}_L^0)
\end{aligned}$$

# 3.5 ViT Architecture:

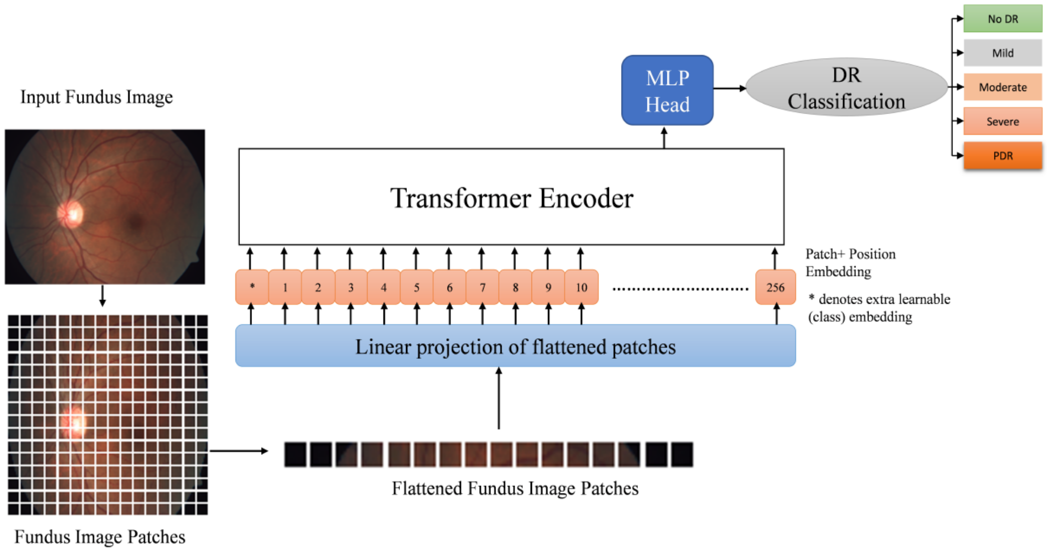


Figure 3.2: ViT-DR Architecture

## 3.5.1 Patch Formation:

Patchifying is the process of dividing an image into smaller non-overlapping patches, each of which is then processed independently. In the context of Vision Transformers (ViT), patchifying refers to the initial step of splitting the input image into a grid of fixed-size patches, which are then treated as ”tokens” and fed into the Transformer network. The reason for patchifying in ViT is that the original Transformer architecture, which was developed for natural language processing, operates on sequential input, such as sentences or paragraphs. In contrast, images

are two-dimensional, and their pixels have a spatial relationship with each other, which is lost when flattening the image into a sequence of tokens. By patchifying the image, we preserve some of this spatial information, allowing the network to learn richer representations of the image. In practice, patchifying involves splitting the input image into patches of fixed size, such as 16x16 or 32x32 pixels. Here 768x768 is pachified by 64x64. Each patch is then linearly projected into a lower-dimensional feature space, typically using a convolutional layer, to obtain a fixedsize embedding vector. These embeddings are then fed into the Transformer encoder, which processes them sequentially, attending to different patches at different layers of the network. Patchifying has been shown to be an effective way of applying the Transformer architecture to image classification tasks, achieving state-of-the-art results on several benchmarks. However, it also has some limitations, such as the inability to handle variable-size inputs and the loss of some spatial information. These limitations have been addressed in follow-up works, such as the DeiT (Data-efficient Image Transformers) and T2T-ViT (Tokens-to-Token ViT) models, which introduce additional techniques to improve the performance of Vision Transformers.

### 3.5.2 Positional Encodings/Embeddings:

In Vision Transformers (ViT), position embeddings are added to the sequence of image patches to preserve their positional information. This allows the model to learn richer representations of the image by attending to different patches at different layers of the Transformer network. In addition to the position embeddings, a learnable class embedding is also attached to the sequence, which is updated by self-attention and used to predict the class of the input image. This class embedding is positioned according to the location of the image patch in the sequence. Finally, the classification task is performed by adding a multi-layer perceptron (MLP) head on top of the Transformer network, which takes the output of the class embedding as input and produces the final class prediction. The MLP head is placed at the position of the class embedding in the sequence, and is trained in an end-to-end manner with the rest of the network. To facilitate the learning process of Transformers, it’s beneficial to include inductive biases in the input of the model. Position embeddings are one example of such biases, which are added to patch embeddings in computer vision tasks to retain the positional information of features. In computer vision, position embeddings can represent either the position of a feature in a 1-dimensional flattened sequence or a 2-dimensional position of a feature, using X and Y embeddings. A sequence of patches typically works better with 1-dimensional embeddings. Another type of position embeddings is relative position embeddings, which define the relative distance between all possible pairs of features.

This helps the model to learn the spatial relationships between different features and improves its performance on tasks such as object detection and segmentation.

### 3.5.3 Transformer Encoder:

The Transformer Encoder Layers in Vision Transformer (ViT) are similar in concept to those in the original Transformer architecture used in natural language processing, but adapted for use in computer vision tasks. In ViT, the input to each Transformer Encoder layer is a patch-based representation of the image, where each patch is flattened and projected to a lower dimension. Like in the original Transformer architecture, each Transformer Encoder layer in ViT consists of two sub-layers: the Multi-Head Self-Attention (MHSA) layer and the Feed-Forward (FF) layer. The MHSA layer in ViT computes the self-attention mechanism across the flattened patches. Specifically, it takes in the patch embeddings as input and computes the query, key, and value matrices, similar to the original Transformer architecture. However, unlike in the original architecture, the patch embeddings are first processed through a positional encoding layer to allow the model to reason about the spatial relationships between the patches. The selfattention scores are then used to weight the value matrix to obtain the output of the MHSA layer. The FF layer in ViT applies a simple feed-forward neural network to each patch independently, similar to the original Transformer architecture. However, in ViT, an additional layer normalization step is added after the feed-forward network to improve training stability. The output of the FF layer is then added to the output of the MHSA layer and passed to the next Transformer Encoder layer. Overall, the Transformer Encoder Layers in ViT allow the model to capture spatial dependencies between patches in an image by leveraging the self-attention mechanism. This allows the model to process the image in a patch-based manner, while also learning meaningful representations that capture the global structure of the image. This has proven to be a successful approach for various computer vision tasks, such as image classification, object detection, and segmentation.

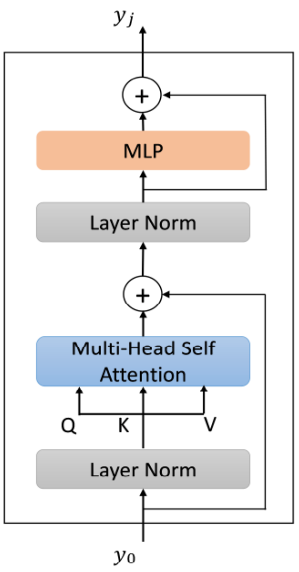


Figure 3.3: Transformer Encoder

### 3.5.4 Multilayer Percetron(MLP):

A Multilayer Perceptron (MLP) is a type of neural network that consists of multiple layers of fully connected neurons. In the context of the Vision Transformer (ViT), an MLP is typically added on top of the Transformer Encoder layers to perform the final classification task. After the image patches have been processed through the Transformer Encoder layers, the resulting embeddings are concatenated and fed through an MLP. The MLP typically consists of several fully connected layers with non-linear activation functions, such as ReLU. The output of the MLP is then fed through a final softmax layer to produce the class probabilities. The purpose of the MLP in ViT is to allow the model to perform the final classification task by learning non-linear relationships between the extracted features and the target classes. The MLP takes the high-level representations learned by the Transformer Encoder layers and transforms them into a form suitable for the classification task. The number of neurons and layers in the MLP can vary depending on the complexity of the classification task and the size of the ViT model. A larger MLP with more neurons and layers can capture more complex relationships between the features and the classes, but also requires more computation and training time. Overall, the addition of an MLP on top of the Transformer Encoder layers in ViT allows the model to perform the final classification task by learning non-linear relationships between the extracted features and the target classes, making it a powerful tool for various computer vision tasks.

### 3.5.5 GeLu Activation Function:

Gelu (Gaussian Error Linear Units) is an activation function that was introduced in a 2016 paper by Dan Hendrycks and Kevin Gimpel. It’s a smooth approximation of the rectified linear unit (ReLU) activation function, which has been widely used in deep learning.

The Gelu activation function is defined as follows:

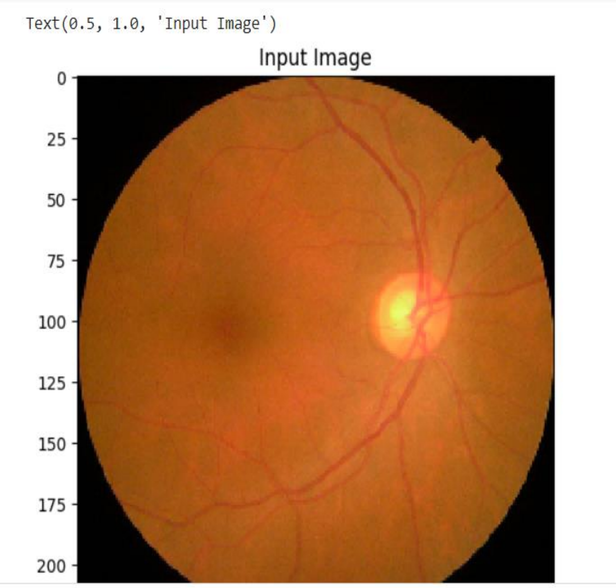
**gelu(x) = 0.5 * x * (1 + tanh(sqrt(2/pi) * (x + 0.044715 * x3 )))**

where x is the input to the function. The main advantage of Gelu over ReLU is that it’s a smooth function that doesn’t have the ”dead neuron” problem. This means that there are no regions where the derivative is zero, which can cause the gradients to vanish during backpropagation and slow down the learning process. Moreover, Gelu has been found to perform slightly better than ReLU on some benchmarks, including the ImageNet classification task. This may be due to the fact that Gelu has a nonzero mean and a non-unit variance, which can help to reduce the internal covariate shift and improve the generalization ability of the model. However, Gelu is computationally more expensive than ReLU, as it involves the calculation of the hyperbolic tangent function and the square root function. Therefore, it may not be suitable for some applications where speed is critical.

# Chapter 4

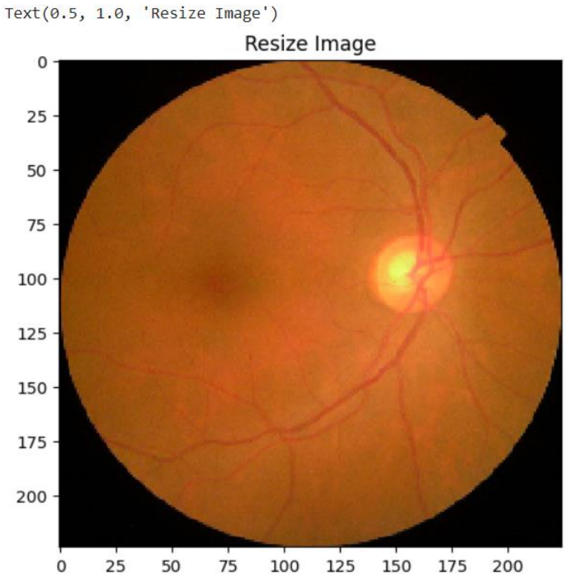
# Results

## 4.1 Input Image:



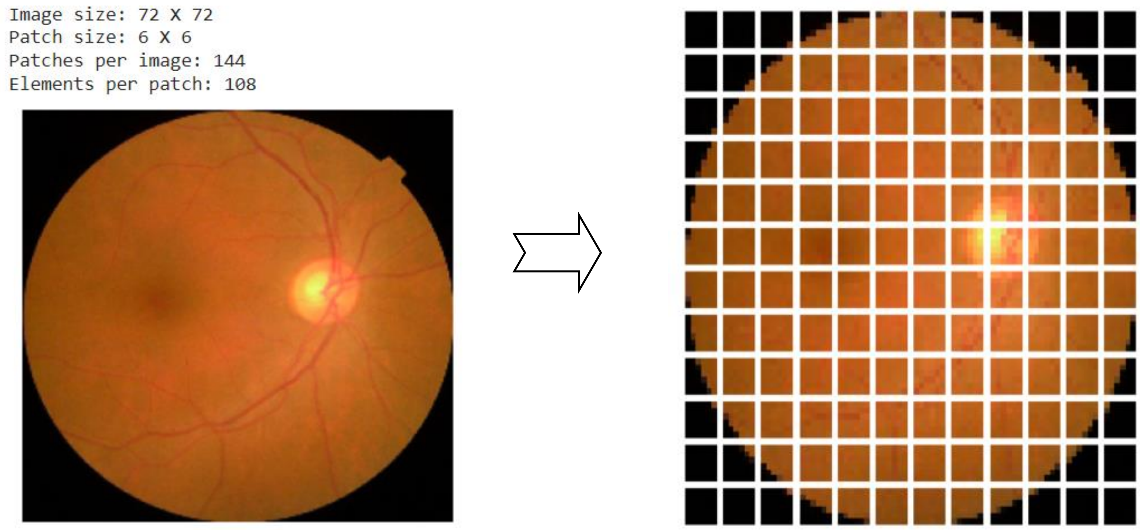
```
Text(0.5, 1.0, 'Input Image')
```

## 4.2 Resized Image:



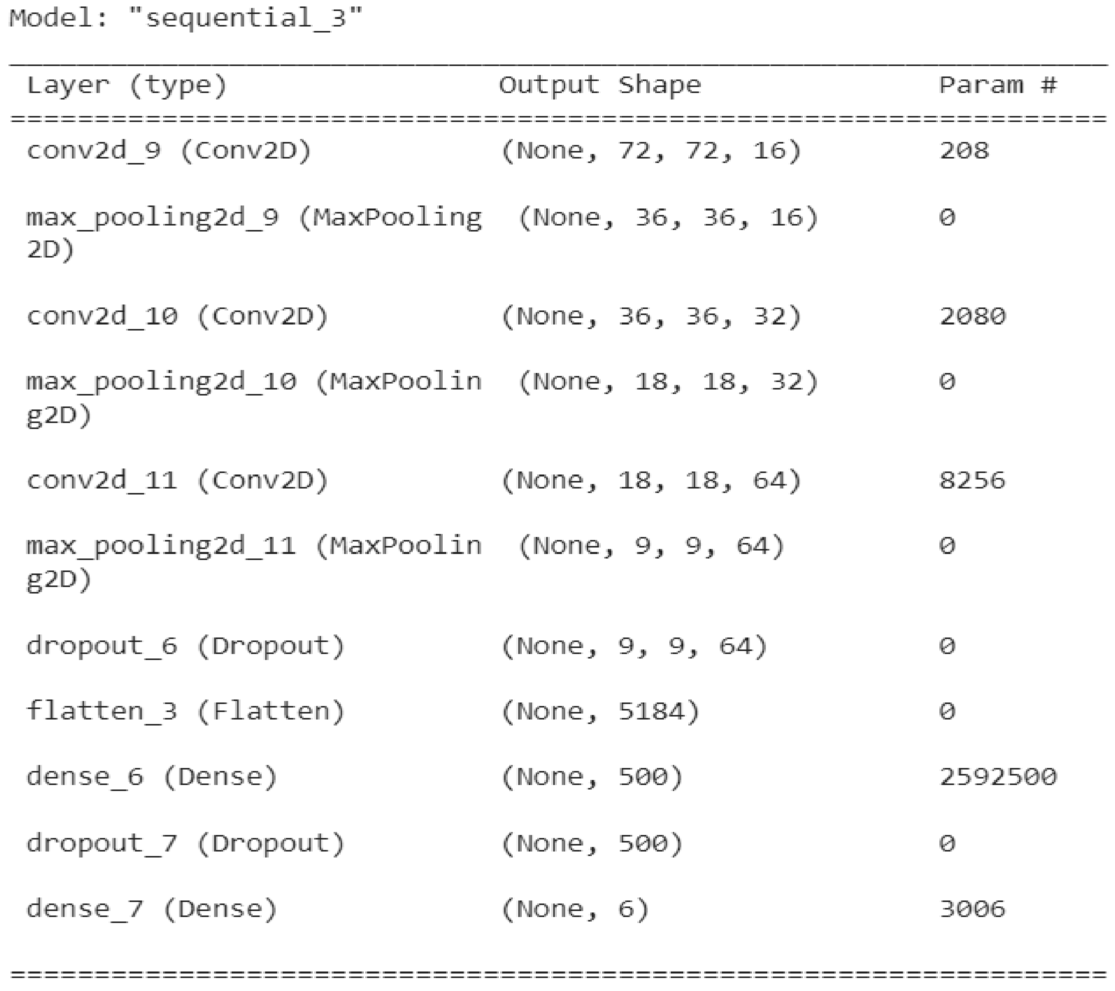
```
Text(0.5, 1.0, 'Resize Image')
```

## 4.3 Patch Formation:

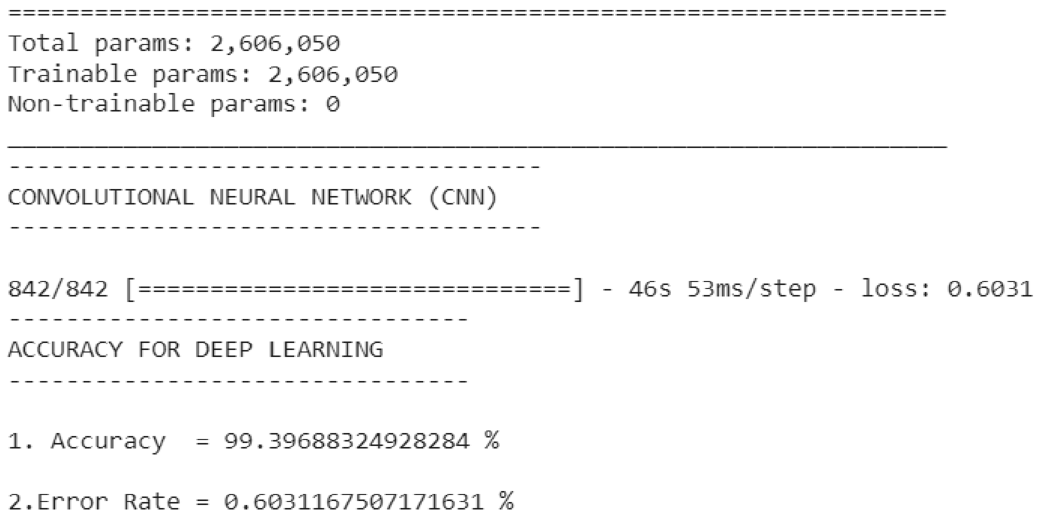


```
Image size: 72 X 72
Patch size: 6 X 6
Patches per image: 144
Elements per patch: 108
```

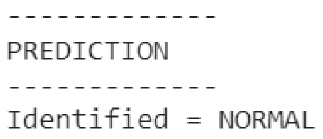
## 4.4 Model summary and prediction results:



```
Model: "sequential_3"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 conv2d_9 (Conv2D)           (None, 72, 72, 16)        208

 max_pooling2d_9 (MaxPooling  (None, 36, 36, 16)       0
 2D)

 conv2d_10 (Conv2D)          (None, 36, 36, 32)        2080

 max_pooling2d_10 (MaxPoolin  (None, 18, 18, 32)       0
 g2D)

 conv2d_11 (Conv2D)          (None, 18, 18, 64)        8256

 max_pooling2d_11 (MaxPoolin  (None, 9, 9, 64)         0
 g2D)

 dropout_6 (Dropout)         (None, 9, 9, 64)          0

 flatten_3 (Flatten)         (None, 5184)              0

 dense_6 (Dense)             (None, 500)               2592500

 dropout_7 (Dropout)         (None, 500)               0

 dense_7 (Dense)             (None, 6)                 3006

=================================================================
```



```
=================================================================
Total params: 2,606,050
Trainable params: 2,606,050
Non-trainable params: 0
_________________________________________________________________
-------------------------------------
CONVOLUTIONAL NEURAL NETWORK (CNN)
-------------------------------------

842/842 [==============================] - 46s 53ms/step - loss: 0.6031
--------------------------------
ACCURACY FOR DEEP LEARNING
--------------------------------

1. Accuracy  = 99.39688324928284 %

2.Error Rate = 0.6031167507171631 %
```



```
-------------
PREDICTION
-------------
Identified = NORMAL
```

# Chapter 5

# Conclusion

Diabetic retinopathy causes irreversible visual loss if not detected and treated on time. An improved transformer model for grading the severity levels of diabetic retinopathy is introduced. The performance of the model is improved by selecting the proper fundus image size, patch size, number of transformers, multi-layer perceptron head layers. After performing several experiments, the optimal values for training the model are selected, which gave better results with less computational cost. The model is tested on kaggle dataset. The results achieved are better than the other approaches like CNN and its variants etc. The performance of the model can be improved further by increasing the size of the database and patches at the cost of computation. Preprocessing the fundus images to identify micro aneurysms helps to increase the diabetic retinopathy classification results. In the future, the efficiency of the model can be improved with better preprocessing techniques and can be tested with a real time database.

# Bibliography

**[1] W. M. Gondal, J. M. Köhler, R. Grzeszick, G. A. Fink and M. Hirsch, "Weakly-supervised localization of diabetic retinopathy lesions in retinal fundus images," 2017 IEEE International Conference on Image Processing (ICIP), Beijing, China, 2017, pp. 2069-2073, doi: 10.1109/ICIP.2017.8296646.**

**[2] Y. Yang, T. Li, W. Li, H. Wu, W. Fan, and W. Zhang, "Lesion detection and Grading of Diabetic Retinopathy via Two stages Deep Convolutional Neural Networks" Int. Conf. Med. Image Comput. Comput.-Assist. Intervent. Springer, 2017 on for Computational Linguistics.**

**[3] E. Abdelmaksoud, S. El-Sappagh, S. Barakat, T. Abuhmed and M. Elmogy, "Automatic Diabetic Retinopathy Grading System Based on Detecting Multiple Retinal Lesions," in *IEEE Access*, vol. 9, pp. 15939-15960, 2021, doi: 10.1109/ACCESS.2021.3052870.**

**[4] N. J. Mohan, R. Murugan, T. Goel and P. Roy, "Exudate Localization in Retinal Fundus Images Using Modified Speeded Up Robust Features Algorithm," *2020 IEEE-EMBS Conference on Biomedical Engineering and Sciences (IECBES)*, Langkawi Island, Malaysia, 2021, pp. 367-371, doi: 10.1109/IECBES48179.2021.9398771.**

**[5] H. Jiang, K. Yang, M. Gao, D. Zhang, H. Ma and W. Qian, "An Interpretable Ensemble Deep Learning Model for Diabetic Retinopathy Disease Classification," *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, Berlin, Germany, 2019, pp. 2045-2048, doi: 10.1109/EMBC.2019.8857160.**

**[6] Y. Bazi, L. Bashmal, M. M. A. Rahhal, R. A. Dayil, and N. A. Ajlan, "Vision transformers for remote sensing image classification," Remote Sensing, vol. 13, no. 3, p. 516, 2021.**

**[7] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly et al.,"An image is worth 16x16 words: Transformers for image recognition at scale," arXiv preprint arXiv:2010.11929, 2020.**

**[8] S. Qummar, F. G. Khan, S. Shah, A. Khan, S. Shamshirband, Z. U. Rehman, I. A. Khan, and W. Jadoon, "A deep learning ensemble approach for diabetic retinopathy detection," IEEE Access, vol. 7, pp. 150 530–150 539, 2019.**

**[9] J. Wu, R. Hu, Z. Xiao, J. Chen, and J. Liu, "Vision transformer-based recognition of diabetic retinopathy grade," Medical Physics, vol. 48, no. 12, pp. 7850–7863, 2021.**

**[10] N. AlDahoul, H. Abdul Karim, M. Joshua Toledo Tan, M. A. Momo, and J. Ledesma Fermin, "Encoding retina image to words using ensemble of vision transformers for diabetic retinopathy grading,"F1000Research, vol. 10, p. 948, 2021.**

**[11] Alexey Dosovitskiy, L. Beyer, and K. Alexander, *An Image Is worth 16x16 words:Transformers for Image Recognition at Scale*, ICLR, Vienna, Austria, 2021.**

**[12] N. J. Mohan, R. Murugan, T. Goel, and P. Roy, "Optic disc segmentation in fundus images using operator splitting approach," in 2020 advanced communication technologies and signal processing (ACTS). IEEE, 2020, pp. 1–5.**

**[13] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N.Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need,"Advances in neural information processing systems, vol. 30, 2017.**