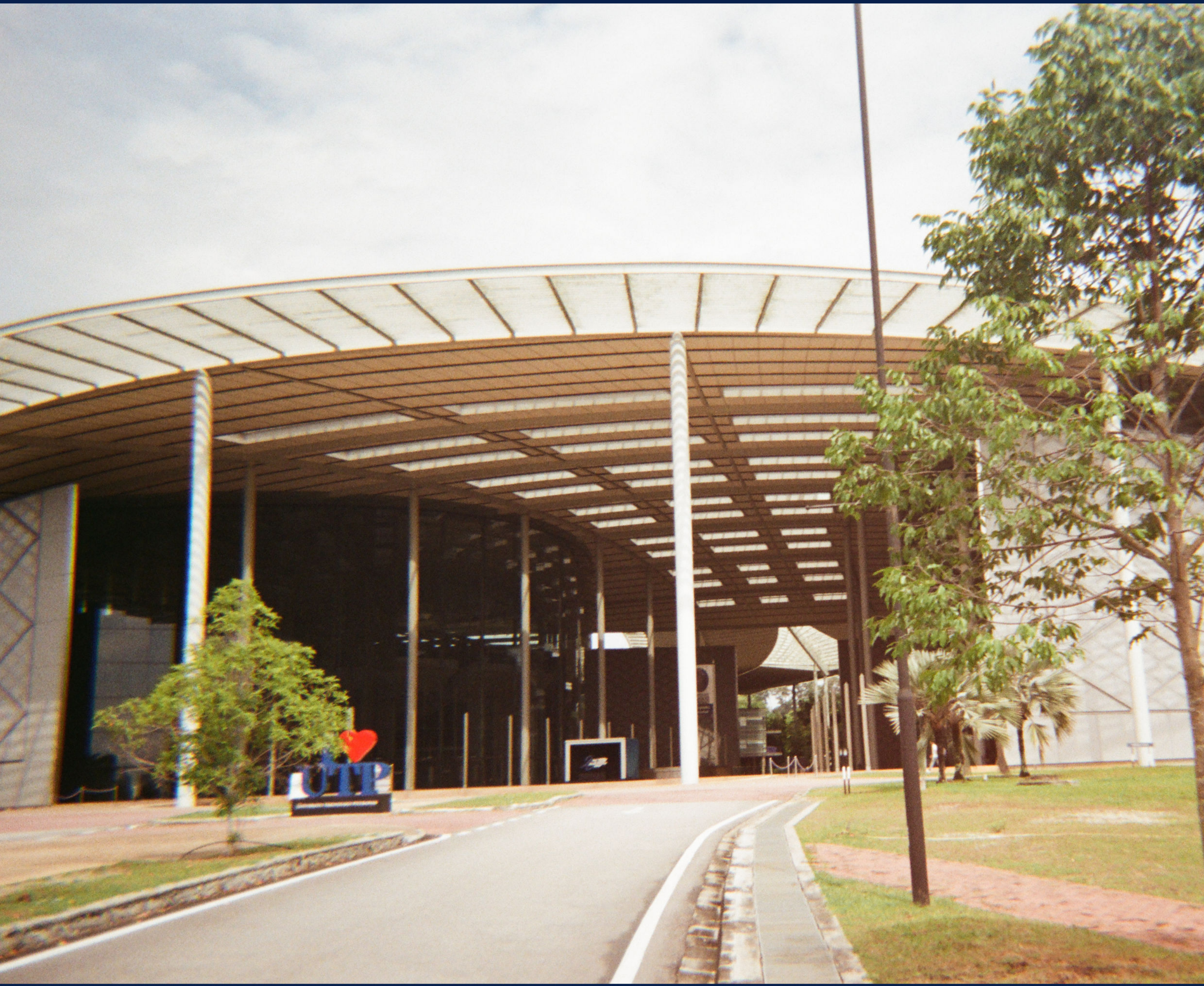
Internship Report
Diabetic Retinopathy Detection

Author: Anquetin Romain
Promotion Quantum, IE

UTP tutor: Dr. Iznita Binti Izhar Lila
ESIREM tutor: Dr. Marquié Patrick



UTP
UNIVERSITI TEKNOLOGI PETRONAS

ESIREM
ÉCOLE SUPÉRIEURE D'INGÉNIEURS
NUMÉRIQUE ET MATÉRIAUX

École associée
Polytech

Contents

List of Figures

List of Tables

1 Acknowledgement

First and foremost, I would like to thank my tutor Dr. Iznita Binti Izhar Lila who guided me in doing these projects, and providing the equipment which we needed.
Besides, we would like to thank Dr. Marquié Patrick who helped us to organize this internship in Malaysia.
I am dearly obliged to Dr. Journaux Ludovic for giving precious advice.
Also, I would like to thank my family and friends for their support.
At last but not in least, we would like to thank everyone who helped and motivated us to work on this project.

2 Work location

Universiti Teknologi PETRONAS (UTP) was established on 10 January 1997 and is a leading private university in Malaysia.
The campus is built on a 400 hectare (1,000 acres) site strategically located at Bandar Seri Iskandar, Perak Darul Ridzuan, Malaysia. The university is a wholly-owned subsidiary of PETRONAS, the national oil and gas company of Malaysia.[22]
The university conducts research activities in collaboration with PETRONAS on six research areas : self-sustainable building, transport infrastructure, health analytics, hydrocarbon recovery, contaminant management and autonomous system.[23]

UTP Chancellor complex

3 Introduction

The internship take place in the Electronic Engineering department of UTP. The internship project is to detect one eye disease affecting diabetic peoples.The diabetic retinopathy (DR) it can lead to permanent vision loss if not treated. People with diabetes rise to 171 million in 2020 to 366 million in 2030[27]. Currently, detecting DR is a time-consuming and manual process that requires a trained clinician to examine and evaluate digital color fundus photographs of the retina. By the time human readers submit their reviews, often a day or two later, the delayed results lead to lost follow up, miscommunication, and delayed treatment.
The objective of this work is to develop a computer aided diagnosis algorithm. This algorithm will help expert and allow them to review more patient. To tackle this problematic, we used deep neural network.

3.1 Diabetic retinopathy

Anyone with any kind of diabetes can get diabetic retinopathy (DR). The diabetic retinopathy is caused by a high level of sugar in the blood. Over time, this excess of sugar can damage the retina. Diabetes damages blood vessels all over the body and the eye are vulnerable. The damage to eyes start on the tiny blood vessel that goes to the retina, causing them to leak fluid or blood. If not treated, DR can cause vision loss or blindness. DR has multiple stages :

- Background retinopathy – tiny bulges develop in the blood vessels, which may bleed slightly but do not usually affect your vision.
- Pre-proliferative retinopathy – more severe and widespread changes affect the blood vessels, including more significant bleeding into the eye.
- Proliferative retinopathy – scar tissue and new blood vessels, which are weak and bleed easily, develop on the retina; this can result in some loss of vision.

Diabetic people should do eye screening once a year, to detect the disease in the early stages. The screening test involves examining the back of the eyes and taking photographs. Currently, the picture is taken and analyze by a doctor. [16] [15]

3.2 Problem to solve

As said in the introduction 3, the number of people with diabetes will increase in a near future. This lead to a problem of the increase amount of eye fundus images to potentially detect early sign of DR. This amount of data will require a lot of effort and time to analyze. This can lead to an erosion in overall quality of DR screening or impact other eye disease. Developing tools to maintain or possibly improve the quality of DR screening, by reducing the amount of work and time to analyze the data, with a computer aided diagnosis algorithm. This is an image classification problem.

3.3 Scope of work

Deep learning algorithm achieve high sensitivity and specificity compared to other standard technics [8]. Recent work in deep learning computer vision as found a new promising image processing technics call vision transformer (ViT). This new technics has first been introduced in language processing [24] and later adapted to image with Vision Transformer [4]. The particularity of this new technics is it ability to create long range dependency in the image, enabling a better understanding and thus accuracy.
This work will focus on the usage of transformer to detect and classify DR.

4 State of the art

4.1 Introduction

In this section, we will discuss of the recent advance in DR detection technics going through neural network architecture and image processing to help detection. Deep neural networks will be studied because of their higher performance compared to standard machine learning.

4.2 Convolutional neural network

Convolutional neural network (CNN) are made around the convolution operation, see figure 1. First introduced by Lecun and al [11, 10]. CNN has demonstrated through time excellent performance in image application. The powerful learning ability of deep CNN is primarily due to the use of multiple feature extraction stages that can automatically learn representations from the data. The availability of a large amount of data and improvement in the hardware technology has accelerated the research in CNNs, and recently interesting deep CNN architectures have been reported.

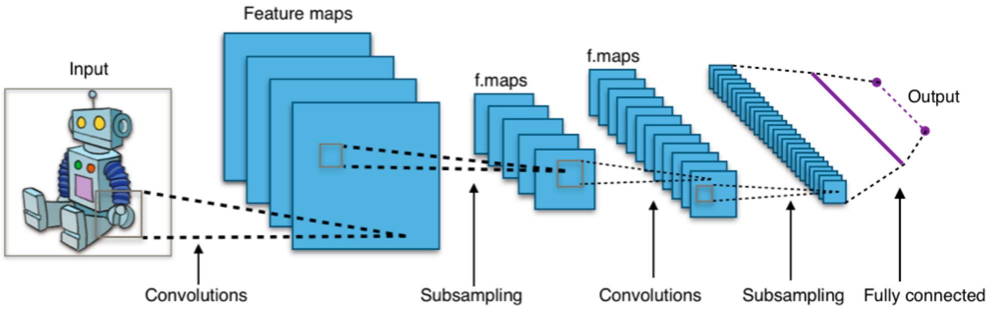


Figure 1: Typical CNN network for image classification source is from Wikipedia ”Convolutional neural network” article [26]

In DR detection, the most utilized architectures are VGG16, inception family and ResNet [21]. These networks are well understood by their democratization in other fields. They are easily trainable and can be used quickly on new data through a lot of available frameworks and git repository. Due to their popularity, people have developed tools to explain the prediction of these networks to reduce the problem of black box.

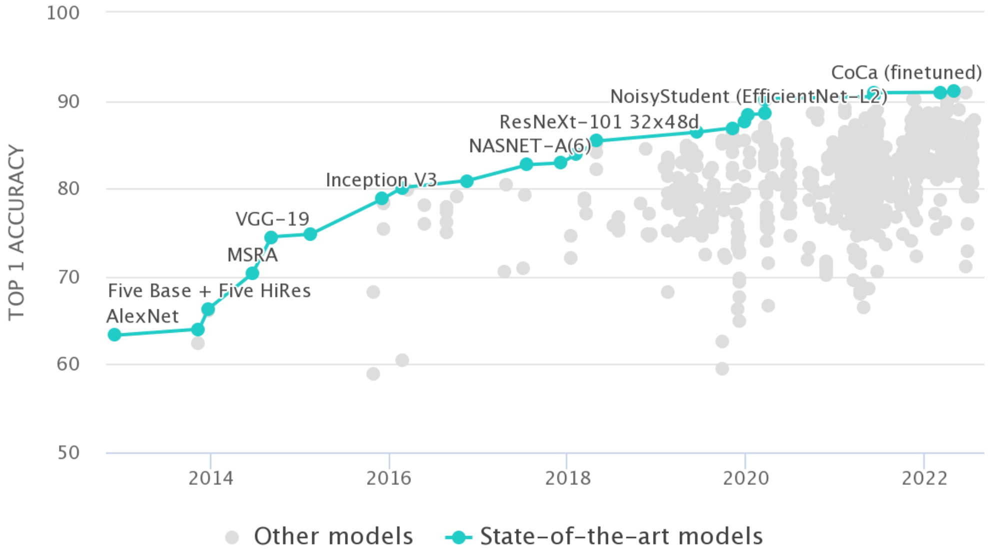


Figure 2: State-of-the-art network in Image Classification task on ImageNet [3] Dataset — PaperWithCode[17]

Specialized CNN architectures made to detect DR are still used and are performing better than some general CNN. Some studies used the severity correlation between the two eyes of a patient to develop CNN models leveraging this particularity [6, 25]. By using two eyes, the network has more information to used and thus has better prediction. One good example of a network is TSBN [18]. This network outperforms some DL networks using one eye. Using two eyes can be a quick and easy way of increasing the accuracy.
The most famous DR detection CNN network is the Zoom-in-Net [25]. This network works by extracting suspicious locals patches from the image. Thus, making a prediction with the global image and the locals patch. Notice the use of the two patient eyes that most standard networks don't use. It achieves very good precision at 0.865/0.854 on val/test set of Eyepacs [5] dataset.
The most recent CNN architecture explores new concepts like attention. This allows networks to have long range dependency. This dependency helps the CNN to have awareness of all the available data. One example of this new type of CNN with attention is PatchConvNet-S60 [20]. Efforts are made to modernize older CNN architecture to maintain them as a state of the art. One of them is the network ConvNet [14] which modernize the ResNet [7] architecture with lessons learned since its creation.

4.3 Vision Transformer

Transformers network originally appeared in the paper "Attention is all you need" [24]. The concept comes from natural language processing(NLP). NLP needs long term dependency between words to make a good prediction through long text. Transformers are based on attention mechanism and are highly parallel to speed up the training and inference time has figures 3 show. This type of attention is call scaled dot product attention, the input consists of queries and keys of dimension d_k and values dimension d_v. In practice, queries, keys and values are packed together into matrix named Q, K, V respectively.

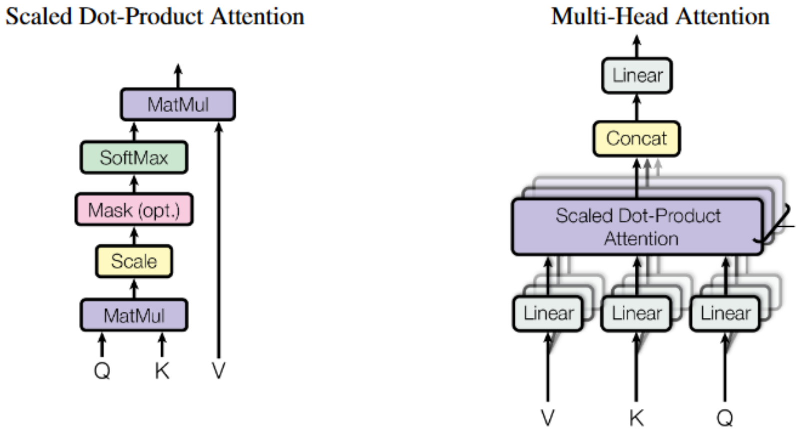


Figure 3: (left) Scaled Dot-Product Attention. (right) Multi-Head Attention consists of several attention layers running in parallel. The illustration come from "Attention is all you need" [24]

The first usage of transformer in vision task comes from the paper "An Image is worth 16x16 words" [4] with the ViT network 4. As figure 4 shows, they split an image into fixed-size 16x16 patches, linearly embedding each of them, add position embeddings, and feed the resulting sequence of vectors to a standard Transformer encoder. In order to perform classification, they use the standard approach of adding an extra learnable “classification token” to the sequence.

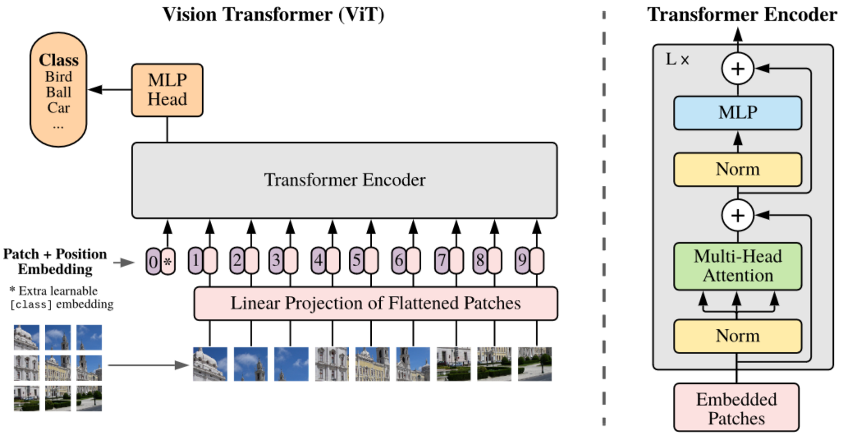


Figure 4: ViT model overview. The illustration come from "An Image is worth 16x16 words" [4]

From this paper, new architectures have been created. Solving some of the problem of the ViT network. First the amount of data needed to train the transformer is paramount, some dataset with 3 billion images can be considered small. Also, the time of training is somewhat long, less than 3 days. Another type of network is the Swin transformer [13, 12] it improves the network's ability to take larger image input by rethinking the 16x16 patch and using a shifted window.

4.4 Contrast enhancement for DR detection

Contrast enhancement is a very important step in the medical image classification task. In the literacy, there are multiple way to enhancing the quality of an image. To evaluate each method we will use the data and metrics of [28]: They use 5 metrics to evaluate the performance of enhancing algorithm: histogram, entropy, absolute mean brightness error, Signal to Noise Ratio (SNR) and Peak Signal to Noise Ratio (PSNR). In the table 1, we present a comparative list of different algorithm or technics to enhance the quality of an image.

Algorithm	Observation
Histogram equalization	Increases the global contrast of the image, but neglects the local variations across the image, thus losing information, lowest PSNR.[28, 21]
Adaptive histogram equalization	Can enhance the noise in an image, to preserve the brightness at some extent.[28]
Contrast Limited adaptive histogram equalization (CLAHE)	Can enhance the noise in an image. CLAHE preserve the brightness and has better PSNR.[28]
Exposure based Sub-Image Histogram Equalization	Enhances fundus image quality, low PSNR thus introducing/enhancing noise.[28]
Non-Local Means Denoising	Good to reduce noise coming from the enhancement algorithm, but can also destroy information.[21]

Table 1: List of enhancing algorithms and observation from [28, 21]

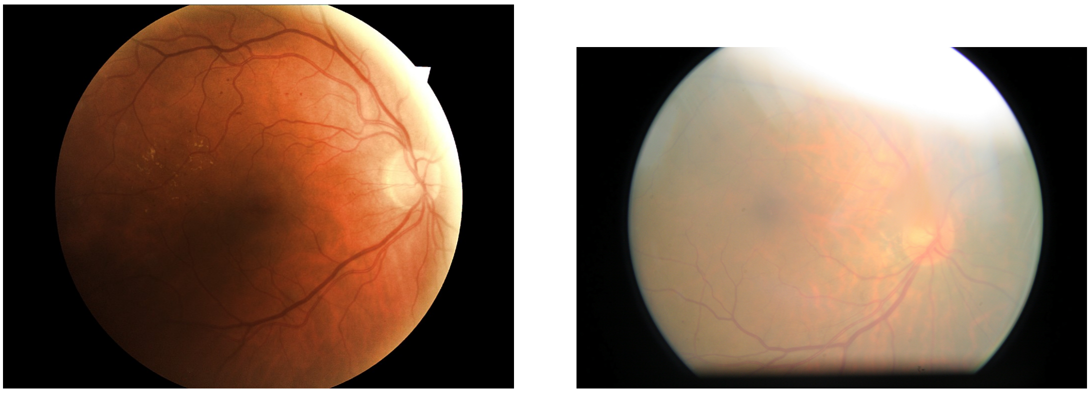
5 Available data

To train our algorithm, data is needed. There is multiple dataset available for different task. For our task, we used the Eyepacs dataset [5] downloaded from Kaggle [9].The Eyepacs [5] dataset is publicly available. Also, it's the most important one by its size with 88 702 images and in terms of variation of image quality. Higher quality dataset are available, but the number of image is limited. One solution to this problem is to first train the network on the Eyepacs [5] dataset and fine-tune it on the dataset with fewer images.

5.1 Image presentation

The figure 5 represent two images from the Eyepacs [5] dataset. The different parameter in quality between the two images are :

- Lighting
- Focus
- Centered eye fundus
- Resolution
- Type of imaging device



(a) Exemple of Eyepacs image [5]

(b) Exemple of overexposed Eyepacs image [5]

Figure 5: Exemple of two images from Eyepacs dataset [5]

5.2 Dataset Analysis

In the Eyepacs dataset [5], each image on a scale of 0 to 4, according to the following scale:

0. - No DR
1. - Mild
2. - Moderate
3. - Severe
4. - Proliferative DR

The figure 6 help to understand the distribution of each class in the dataset. The NO DR class is overrepresented. This will lead to a learning bias toward this class. Moreover, other class has at least 1000 images each. This will be enough to correctly train a network. The most severe class are less represented, this can be explained by people getting their

DR diagnostic before the most extreme cases. This unbalance in class representation is also more accurate in terms of real data distribution.
At the pixel level, the most common class is NO DR pixel.
All this point will lead network to learn more NO DR class feature than the other class. Increasing its ability to differentiate between 0 and [1-4] classes. Its accuracy to classify between [1-4] classes will be inferior.

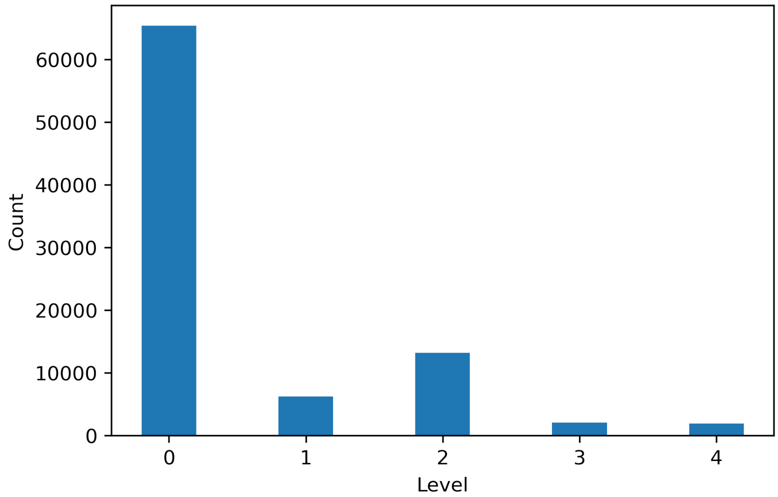


Figure 6: Histogram of class distribution in the Eyepacs dataset [5]

The figure 7 help us to understand the representation of the data and how resize operation will work. Because the resize operation is important to get a square image before feeding it to the network. Changing the scale of an image can be a source of loss in information on precise detail.
For example : Let use a 1024x512px image, and resize it to 512x512px. The final image will be squeezed on it's X axis.
In this case, figure 7 informs us that all images are mostly square, a resize operation can be done without any concern.

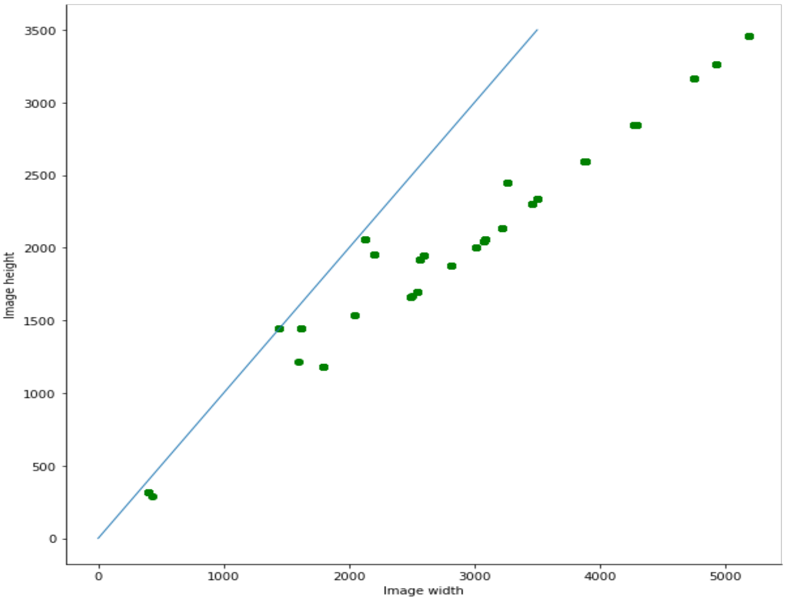


Figure 7: Scatter plot of the image ratio in the Eyepacs dataset [5]

6 Technical choices

6.1 Introduction

In this section, technical choices for the neural network, the image processing algorithm and the data augmentation operation are explained.

6.2 Neural Network

From recent advancement in deep learning in CNN and ViT network. We have chosen to use the network ConViT [2]. The ConViT network is motivated by a property of Self Attention (SA) layer [1]. This property link SA layers and convolution layers, they prove that a multi-head self-attention layer with sufficient number of heads is at least as expressive as any convolutional layer.
The ConViT network has a new form of SA layer, named gated positional self-attention (GPSA), which one can initialize as a convolutional layer. Each attention head, then has the freedom to recover expressivity by adjusting a gating parameter [2].This means we can benefit from both the transformer and the convolution layer.
One motivation of choosing a ViT network is the ability of transformer to learn from a big set of data without harming the learned feature, as shown in the figure 8. From the figure 8 more parameter in a network is beneficial for both types, but having too many samples can be harmful to the CNN base network.

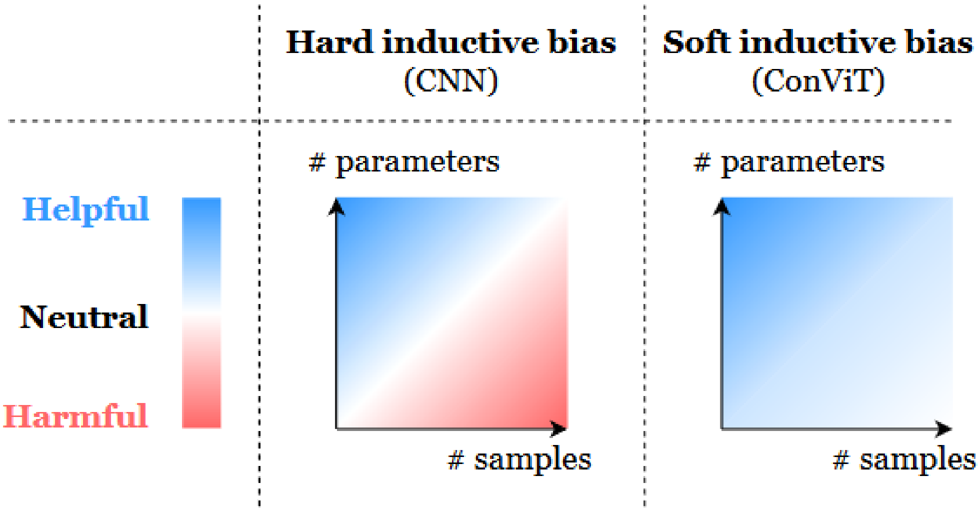


Figure 8: Soft inductive biases can help models learn without being restrictive.The image come from the ConViT [2]

6.3 Image processing

Our choice of image possessing technic is the Contrast Limited adaptive histogram equalization (CLAHE). We choose CLAHE for its ability to enhance the detail with a moderate increase in noise. The figure 9 present effect of CLAHE on eye fundus image, the first one is from the class 2 and the last one is from the class 0.

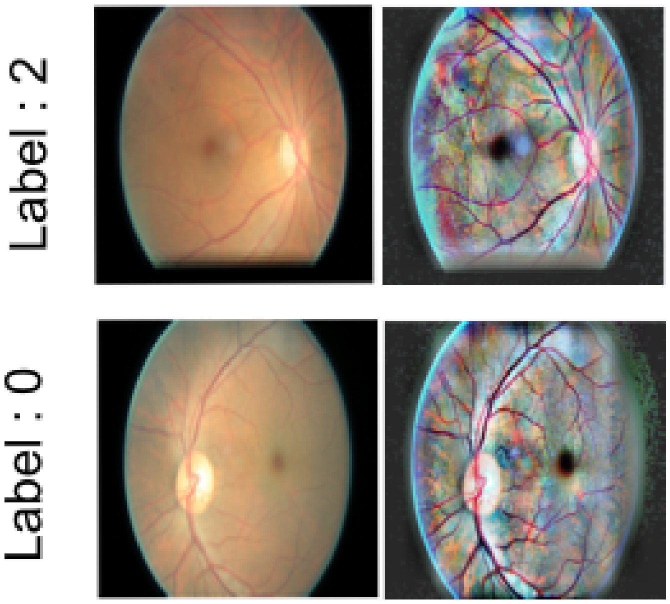
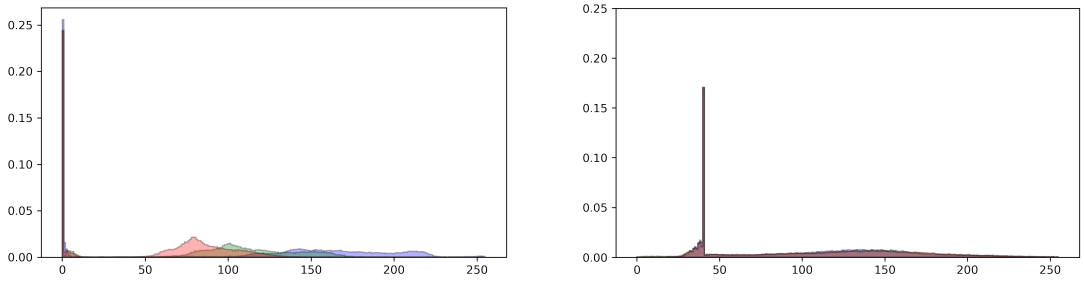


Figure 9: Effect of CLAHE on eye fundus images of different disease level

Ordinary histogram equalization uses the same transformation derived from the image histogram to transform all pixels. This works well when the distribution of pixel values is similar throughout the image. However, when the image contains regions that are significantly lighter or darker than most of the image, the contrast in those regions will not be sufficiently enhanced.
Adaptive histogram equalization (AHE) improves on this by transforming each pixel with a transformation function derived from a neighboring region. In its simplest form, each pixel is transformed based on the histogram of a square surrounding the pixel.
Ordinary AHE tends to over amplify the contrast in near-constant regions of the image. CLAHE limit of the amplification by clipping the histogram at a predefined value. The value at which the histogram is clipped, depends on the normalization of the histogram and thereby on the size of the neighboring region. The part of the histogram that exceeds the clip limit is redistribute equally among all histogram bins. In the figure 10, we can see the effect of the CLAHE operation on the histogram.



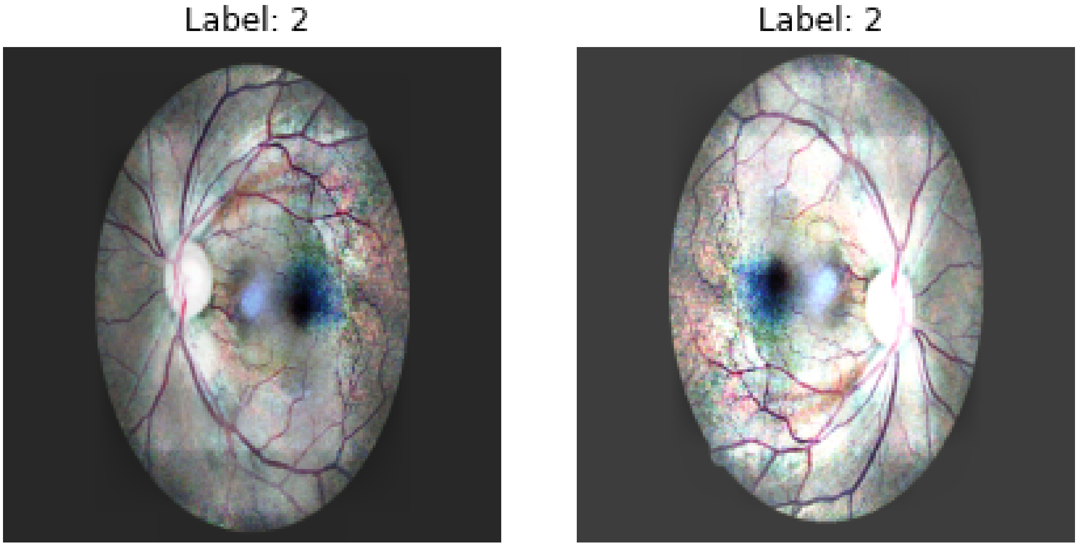
(a) RGB Histogram on an unprocessed eye fundus image

(b) RGB Histogram on an eye fundus image processed with CLAHE

Figure 10: Difference between histogram (a)before CLAHE (b)after CLAHE

6.4 Data Augmentation

Data augmentation is used to artificially increase the number of samples by making variations in the images. Each operation is randomly applied to each image.
One outstanding step is the color jitter operation, this will slightly change the color of an image. This helps the network to generalize more without focussing on the color. Also, the operation of blurring is important because multiple image in the dataset and in real data can be out of focus and thus blurry. Some time, the eye fundus pictures are cut on the edge, a good way to help the network learn better on this case is to randomly crop the image during training. This will artificially create more cropped eye fundus image and increase the accuracy on naturally cropped images.



(a) Image without data augmentation

(b) Image with data augmentation

Figure 11: Difference between two image (a)before data augmentation (b)after data augmentation

7 Implementation

7.1 Introduction

In this section, the implemented and used technical choice are exposed. Starting by explaining how the dataset is split into 3 sub sets, train, test, validation. Then with which library/framework we have worked.

7.2 Data and Data Processing

The EyePacs dataset [5] has 3 parts, train, public validation and private validation because of its use in competition. The train is accessible by every participant, the validations are only visible on the leaderboard, one public and one private. The private leaderboard is not visible until the end of the competition.

Our train dataset is split into 80%/20%. 80% for the training and 20% for the test. To have a comparable metrics and data to other work, the private and public are use as validation.

To use our data, CLAHE is first use on each image, because of the implementation of CLAHE the value for each pixel has to be between [0,1]. To do it, the whole image is divided by 255. After the CLAHE processing, the standardization step applies to our data by calculating the mean and the standard deviation of each pixel in the train set. Then subtracting the mean and dividing by the standard deviation on each pixel for the whole dataset.

$$X' = \frac{X - \mu}{\sigma} \tag{1}$$

With μ the mean and σ the standard deviation.

7.3 Library and Framework

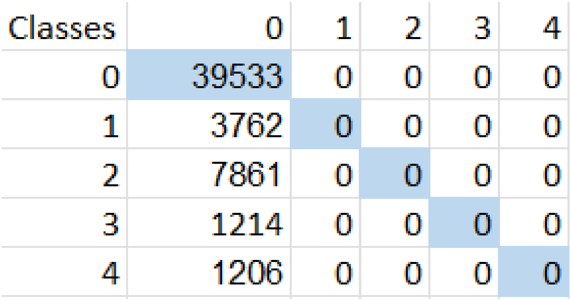
The programming language for this project is python. The two main frameworks are PyTorch and PyTorch Image Models (timm). PyTorch is an open machine learning framework. PyTorch Image Models (timm) is a library and is a collection of computer vision models, layers, utilities, optimizers, schedulers, data-loaders, augmentations and also training/validating scripts.

7.4 Cloud technology and Training

Our network is trained on a Gradient Paperspace notebook. The EyePacs dataset [5] has a size of 83 GB, and therefore we needed to find a way to host the dataset online. We used the ”Amazon S3 - Cloud Object Storage” service.
The pretrain ConViT base network [2] is used to start the fine-tuning on our Eyepacs dataset [5]. The modification of one layer from the network was mandatory because of a mismatch between our classes (5) and the original classes of the network (1000).
The trainings took approximately 4h each on RTX5000 with 16GB GPU RAM and a batch size of 100. The images are resized to 224x224 pixels.
Wandb is used to monitor the training metrics. WandB is a central dashboard to keep track of your hyperparameters, system metrics, and predictions, so you can compare models live, and share your findings.

8 Result and Discussion

During the training, we fine-tuned ConViT network [2], to achieve an accuracy at 73% on the test data. This value is below every paper in the literacy [6, 25, 21, 18, 19]. We noticed a hard limit at 73% of accuracy, signifying a limit in the learning of the network on the data.
After further validation step, the result are not great, accuracy is 20% and precision is 14.76%, the figure 12 help explain what is happening. These results are below expected, the network instead of learning the information on image, it has learned distribution information on classes. To minimize its error, outputting the class 0 is relevant because of its overrepresentation in the dataset.
A way to overcome this problem, will be to reduce the number of image belonging to the class 0 in the training dataset, but this will reduce the amount of data seen by the network. The usage of weighted loss in function of the class can be used to address this issue. Indeed, increasing the loss of classification of underrepresented classes will help the network to focus more on these classes. Another method is to show the network more of the underrepresented classes by showing the same data to artificially balance the dataset in terms of classes, the advantage to this technics is the ability to applies more data augmentation on each image.



Classes	0	1	2	3	4
0	39533	0	0	0	0
1	3762	0	0	0	0
2	7861	0	0	0	0
3	1214	0	0	0	0
4	1206	0	0	0	0

Figure 12: Confusion matrix on the public validation dataset

Another way to increase the accuracy is to tune network and training parameter (hyperparameter), this optimization involves running multiple training with modification in hyperparameter. The hyperparameter is optimized with a grid search algorithm. This hyperparameter search is very expansive in time and hardware.
We also didn't have time to analyze the explainability of the network. This can be done by an ablation study. This means to remove part of the network and analyze how they work and which feature they extract.
Another limitation of our technics is eye fundus image is a disk into a rectangle, this mean there are a lot of black pixels around it. This could be solved by doing a topology transformation to transform the circle into a rectangle to take the whole image.
Compared to other work on transformer network, which they use huge amounts of data, 3 billion images, our studies show the ability of the ConViT network [2] to learn from a smaller dataset with a pretrained on another image recognition dataset.

9 Conclusion

Deep learning gives successful result in disease detection. In this work, we have shown possible usage of vision transformer network to detect the diabetic retinopathy. The learning ability of transformer network can be used on this problematic, even if some datasets are considered as small, by using fine-tuning technics. In future studies, the algorithms will be more precise and more explicable.

10 Personal conclusion

During this internship, I have acquired a strong experience in living abroad in Malaysia. Discovering new cultures is thriving and instructive in the way I saw the world before. Being in an international place like UTP is also fascinating by it multicultural nature. I have also learn new competence in communication and management by traveling and speaking with Malaysians. For a technical standpoint, I have deeply increased my ability to program in python for data science and deep learning. I have explored and learn new methods in those fields by analyzing the state of the art.

11 Project evolution

This section is here to have a clear delimitation between the more scientific report and the difficulties encountered during the project.
The project was divided into multiple phases, see table 2. The limitation on the number of phases in the project is directly linked to a lack of time. For example, other steps could have been: Hyperparameter optimization, Explainability of the network or training on larger bigger images.

Phase	Observation
Literature analysis	In this part, literature analysis was done with a selection of 40 papers
Data analysis	Data analysis on the Eyepacs[5] dataset was to learn how the data can be transformed and processed.
Development of the code base	The development of the code base was made with two frameworks: PyTorch and a higher level API named PyTorch Lightning. This API allows reusing code easily and simplify some non-essential code like manually programming a training. I had to learn those frameworks. On importing the ConViT network [2] from PyTorch to PyTorch Lightning, some custom training function wasn't available and had to be manually set in the code base, thus inducing problem in the first training.
Cloud technology setup	The cloud technologies used are Gradient Paper space and AWS S3. Because of the size of the dataset, the upload took longer than expected and crashed many times. Also, the budget allowed to AWS was exhausted on the first month just by the amount of data stored and retrieved. In fact, AWS don't verify if a file already exists and overwrite it, thus writing 2 files. Some development effort has been made to solve this problem and efficiently upload the dataset.
First training	The first batch of training wasn't successful, the network didn't learn. This was due to the custom training process in the original code of the ConViT network [2].
Back tracking on development	A new development effort was made to solve the past issue by directly taking the code base from the network and insert our custom dataset and preprocess in the training pipeline. This took time to understand how the original code works.
Second training	Second training has been somewhat successful, except a hard limit at 73% of accuracy on the test data.

Table 2: List of phases during the internship, in order

Project management was in total freedom, a meeting with my supervisor each 2 weeks. The work was done alone, with no code review or analysis of process from exterior sources. I didn't get the chance to work with researchers or professors during this internship.

12 Annexes

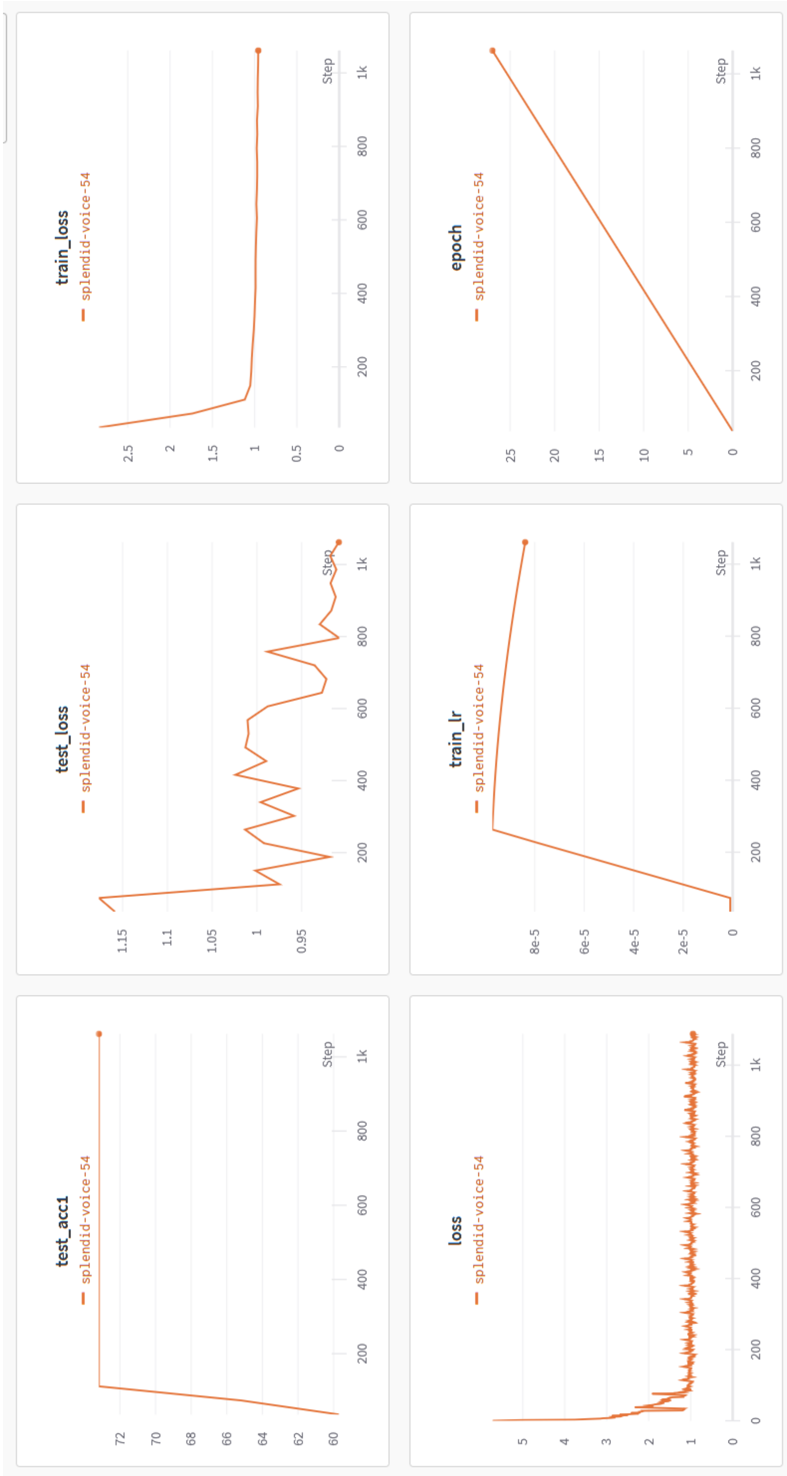


Figure 13: Example of charts generated during training

References

[1] Jean-Baptiste Cordonnier, Andreas Loukas, and Martin Jaggi. *On the Relationship between Self-Attention and Convolutional Layers*. Jan. 10, 2020. arXiv: 1911.03584[cs,stat]. URL: http://arxiv.org/abs/1911.03584 (visited on 08/29/2022).

[2] Stéphane d'Ascoli et al. *ConViT: Improving Vision Transformers with Soft Convolutional Inductive Biases*. Number: arXiv:2103.10697. June 10, 2021. arXiv: 2103.10697[cs,stat]. URL: http://arxiv.org/abs/2103.10697 (visited on 06/27/2022).

[3] 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.

[4] Alexey Dosovitskiy et al. *An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale*. June 3, 2021. arXiv: 2010.11929[cs]. URL: http://arxiv.org/abs/2010.11929 (visited on 08/23/2022).

[5] Eyepacs. *Eyepacs*. In: URL: https://www.kaggle.com/c/diabetic-retinopathy-detection/.

[6] Ben Graham. "Kaggle Diabetic Retinopathy Detection competition report". In: (), p. 9.

[7] Kaiming He et al. *Deep Residual Learning for Image Recognition*. Dec. 10, 2015. arXiv: 1512.03385[cs]. URL: http://arxiv.org/abs/1512.03385 (visited on 08/23/2022).

[8] Md Mohaimenul Islam et al. "Deep learning algorithms for detection of diabetic retinopathy in retinal fundus photographs: A systematic review and meta-analysis". In: *Computer Methods and Programs in Biomedicine* 191 (July 2020), p. 105320. ISSN: 01692607. DOI: 10.1016/j.cmpb.2020.105320. URL: https://linkinghub.elsevier.com/retrieve/pii/S0169260719311010 (visited on 05/11/2022).

[9] Kaggle. *Kaggle*. In: URL: https://www.kaggle.com/ (visited on 09/04/2022).

[10] Y. LeCun et al. "Backpropagation Applied to Handwritten Zip Code Recognition". In: *Neural Computation* 1.4 (1989), pp. 541–551. DOI: 10.1162/neco.1989.1.4.541.

[11] Y. Lecun et al. "Gradient-based learning applied to document recognition". In: *Proceedings of the IEEE* 86.11 (1998), pp. 2278–2324. DOI: 10.1109/5.726791.

[12] Ze Liu et al. *Swin Transformer V2: Scaling Up Capacity and Resolution*. Number: arXiv:2111.09883. Apr. 11, 2022. arXiv: 2111.09883[cs]. URL: http://arxiv.org/abs/2111.09883 (visited on 06/27/2022).

[13] Ze Liu et al. *Swin Transformer: Hierarchical Vision Transformer using Shifted Windows*. Aug. 17, 2021. arXiv: 2103.14030[cs]. URL: http://arxiv.org/abs/2103.14030 (visited on 08/24/2022).

[14] Zhuang Liu et al. *A ConvNet for the 2020s*. Mar. 2, 2022. arXiv: 2201.03545[cs]. URL: http://arxiv.org/abs/2201.03545 (visited on 08/23/2022).

[15] NHS. *Diabetic retinopathy National Health Service (NHS)*. In: URL: https://www.nhs.uk/conditions/diabetic-retinopathy/.

[16] NIH. *Diabetic Retinopathy National Institutes of Health (NIH)*. In: URL: https://www.nei.nih.gov/learn-about-eye-health/eye-conditions-and-diseases/diabetic-retinopathy.

[17] PaperWithCode. *PaperWithCode*. In: URL: https://paperswithcode.com/ (visited on 09/04/2022).

[18] Peisheng Qian et al. “Two Eyes Are Better Than One: Exploiting Binocular Correlation for Diabetic Retinopathy Severity Grading”. In: *2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)* (Nov. 1, 2021), pp. 2115–2118. DOI: 10.1109/EMBC46164.2021.9630812. arXiv: 2108.06763. URL: http://arxiv.org/abs/2108.06763 (visited on 05/11/2022).

[19] Shreya Shekar, Nitin Satpute, and Aditya Gupta. “Review on diabetic retinopathy with deep learning methods”. In: *Journal of Medical Imaging* 8 (2021), p. 32.

[20] Hugo Touvron et al. *Augmenting Convolutional networks with attention-based aggregation*. Number: arXiv:2112.13692. Dec. 27, 2021. arXiv: 2112.13692[cs]. URL: http://arxiv.org/abs/2112.13692 (visited on 06/27/2022).

[21] Nikos Tsiknakis et al. “Deep learning for diabetic retinopathy detection and classification based on fundus images: A review”. In: *Computers in Biology and Medicine* 135 (Aug. 2021), p. 104599. ISSN: 00104825. DOI: 10.1016/j.compbiomed.2021.104599. URL: https://linkinghub.elsevier.com/retrieve/pii/S0010482521003930 (visited on 05/11/2022).

[22] *Universiti Teknologi Petronas about*. In: URL: https://www.utp.edu.my/Pages/The-University/About-Us.aspx.

[23] *Universiti Teknologi Petronas Wikipedia*. In: URL: https://en.wikipedia.org/wiki/Universiti_Teknologi_Petronas.

[24] Ashish Vaswani et al. *Attention Is All You Need*. Dec. 5, 2017. arXiv: 1706.03762[cs]. URL: http://arxiv.org/abs/1706.03762 (visited on 08/23/2022).

[25] Zhe Wang et al. *Zoom-in-Net: Deep Mining Lesions for Diabetic Retinopathy Detection*. Number: arXiv:1706.04372. June 14, 2017. arXiv: 1706.04372[cs]. URL: http://arxiv.org/abs/1706.04372 (visited on 05/31/2022).

[26] *Wikipedia CNN*. In: *Wikipedia*. 2022. URL: https://en.wikipedia.org/wiki/Convolutional_neural_network#/media/File:Typical_cnn.png (visited on 08/24/2022).

[27] Sarah Wild et al. “Global Prevalence of Diabetes”. In: *Diabetes Care* 27.5 (May 1, 2004), pp. 1047–1053. ISSN: 0149-5992, 1935-5548. DOI: 10.2337/diacare.27.5.1047. URL: https://diabetesjournals.org/care/article/27/5/1047/27412/Global-Prevalence-of-DiabetesEstimates-for-the (visited on 07/30/2022).

[28] Sharad Kumar Yadav et al. “Comparative analysis of fundus image enhancement in detection of diabetic retinopathy”. In: *2016 IEEE Region 10 Humanitarian Technology Conference (R10-HTC)*. 2016 IEEE Region 10 Humanitarian Technology Conference (R10-HTC). Agra, India: IEEE, Dec. 2016, pp. 1–5. ISBN: 978-1-5090-4177-0. DOI: 10.1109/R10-HTC.2016.7906814. URL: http://ieeexplore.ieee.org/document/7906814/ (visited on 05/11/2022).