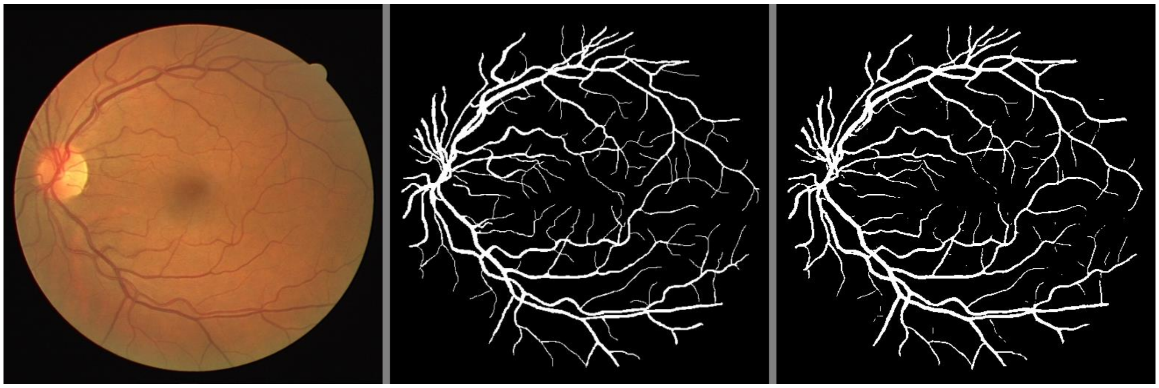


Eye Vessel Prediction

PIXEL-WISE SEGMENTATION WITH U-NET ARCHITECTURES

Derek Dixon | Capstone Project | August 9, 2022

Abstract

One common use-case for deep learning in healthcare is that of semantic segmentation on images. Typically, this comes in the form of identifying anomalies in X-rays or various forms of scans of the body. In this paper, I implement two different neural network architectures to perform semantic segmentation on a small dataset of images of the retina. The models learn to identify which pixels in the images make up the blood vessels found in the eye. I use a data augmentation strategy to increase the number of images used in model training. I train the models from scratch and contrast the difference in performance between the two models. Lastly, I explore the implications this line of research has for the healthcare industry.

Introduction

With the advent of deep learning and its applications to computer vision, a popular task with many different manifestations is that of semantic segmentation. Semantic segmentation refers to the act of predicting the class of each individual pixel in an image. Practical applications of this idea typically involve detecting objects within an image. However, semantic segmentation goes beyond simply producing a binary yes/no prediction of the object's presence. Instead, it outputs an image of the same shape as the input, but with non-class pixels in black and class pixels in white (or a predetermined color if performing multi-class segmentation).

One important real-world application involves identifying the blood vessels in the images of the retina. This is very useful to optometrists and other medical practitioners in the field, as eye scans become more popular in common vision check-ups. Many measures of the identified blood vessels, such as length, width, and branching patterns, are useful for the diagnosis of cardiovascular and ophthalmologic diseases such as diabetes, hypertension, and diabetic retinopathy. Furthermore, the pattern of the retinal vascular tree is known to be unique for everyone, like a fingerprint, and can be used for biometric identification.

The first convolutional neural network designed specifically for medical image segmentation was the U-Net [1]. Many later approaches to neural network design adopted the 'U' shape and attempted to improve upon it by adding various flavors of attention and gating mechanisms [2] [3] [4] [5] [6]. The innovation of the Transformer took the sequence modeling sub-field of deep learning by storm with the seminal paper "Attention Is All You Need" [7] in 2017, laying the foundation for the era of large language models in natural language processing. More recently, researchers began applying the Transformer to image modeling as well [8]. Much research has been done in combining the strength of the convolutional network to learn short-range spatial context with the strength of the Transformer to learn long-range spatial context [9] [10]. Lastly, much work has also been done in combining the basic structure of U-Net with the advent of the Transformer [11] [12].

In this project, I analyze the DRIVE dataset and contrast the performance of two neural network architectures on segmenting blood vessels. The DRIVE dataset is an open-sourced dataset originating with the DRIVE Grand Challenge, an open machine learning competition that took place in 2019 wherein participants competed for who could produce the most accurate predictions for blood vessel segmentation.

Methods

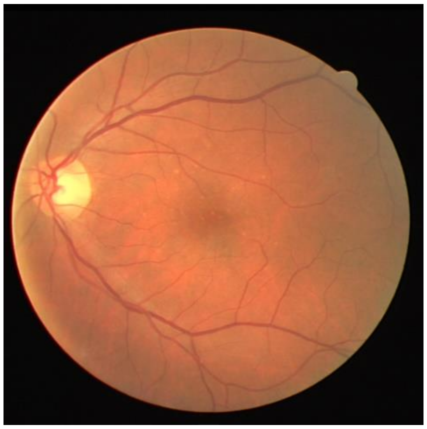
Data

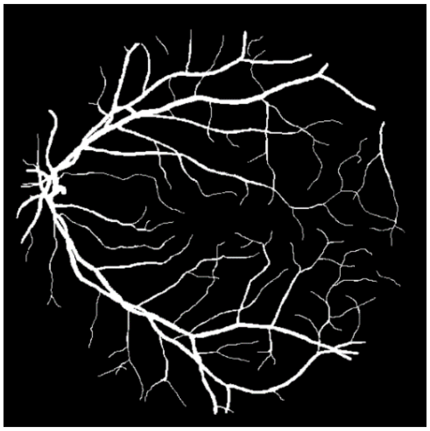
The Digital Retinal Images for Vessel Extraction (DRIVE) dataset is a set of 40 RGB images of the retina obtained from a diabetic retinopathy screening program in The Netherlands [13]. The screening population consisted of 400 diabetic patients between 25-90 years of age. Forty images have been randomly selected, 33 do not show any sign of diabetic retinopathy and 7 show signs of mild early diabetic retinopathy.

The 40 images have been divided into a training and testing set of 20 images each. Each image consists of the RGB image and a manual segmentation map of the vasculature serving as the truth label to be predicted. More information regarding the dataset and how the human labelers generated the maps can be found at the Grand Challenge website.

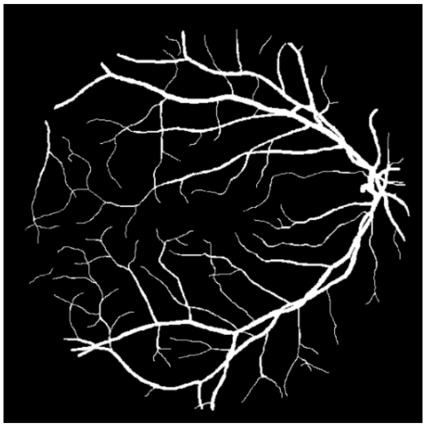
Twenty images is considered to be a very small dataset for training a neural network. To somewhat lessen the impact of this problem, I employed a set of simple augmentation on each image in the training set, generating 3 new images for each original image, thus expanding the size of the training set to 80 images. The augmentation consisted of a horizontal flip, a vertical flip, and a 45-degree rotation. Below is an example of an original image and its horizontal flip:

Original Image and label mask:





Horizontal flip augmentations:



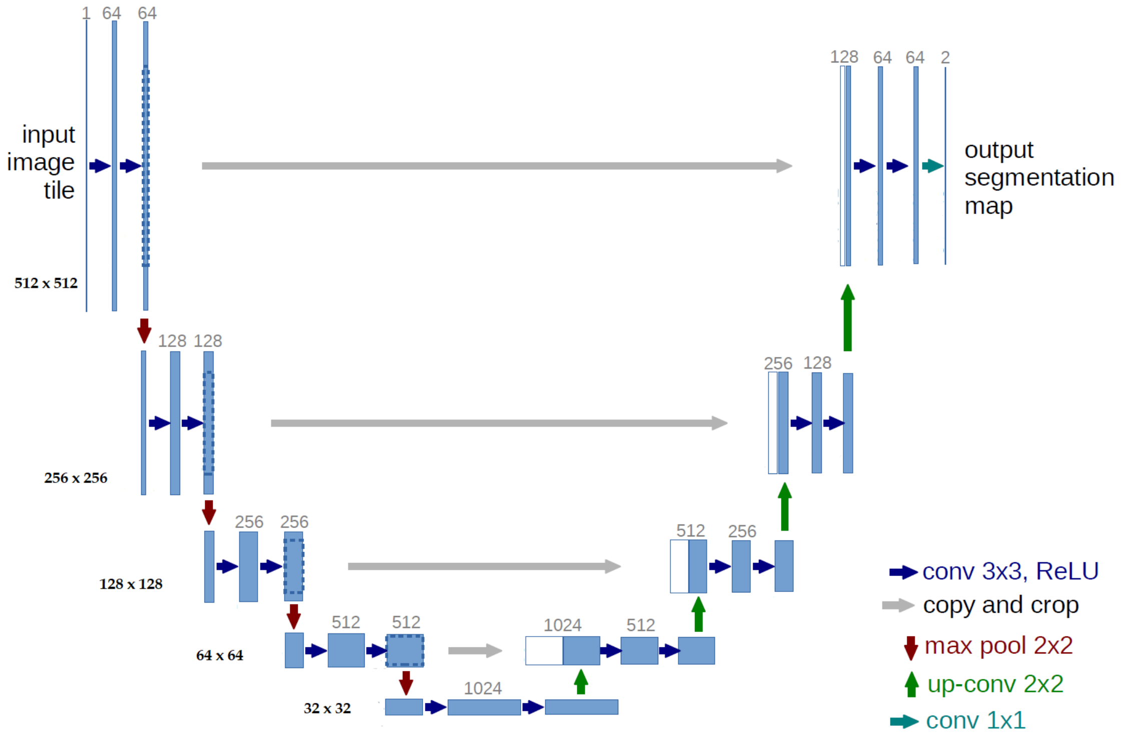
The images are 512x512 pixels. However, for the FCT model only, I needed to rescale the images to 256x256 so that the batches would fit in memory during model training. With augmentation applied, the final data set and splitting scheme was 80 training images with 20 testing images.

Model Architecture Overview

I chose two different neural network architectures for this task. The first is the classic U-Net architecture, first implemented in the influential 2015 paper [1]. The U-Net design has been influential for image segmentation tasks in the biomedical field specifically because, when paired with data augmentation, it can achieve very good performance with very small datasets, which are very common in healthcare. Many subsequent neural network architectures adopted its basic U-shaped design and improved upon it. One very recent such example is my second model choice: the Fully Convolutional Transformer (FCT), released in June 2022 [14]. The FCT is a novel transformer-based model designed specifically for medical image segmentation and adopts the 'U' shape originated by U-Net. It improves upon U-Net by using transformer blocks as the basis of each layer, which are comprised of novel arrangements of convolution and attention. Below I explore in more detail how these architectures work. For even more in-depth explanations, refer to their respective papers.

U-Net

The U-Net is a fully convolutional encoder-decoder network that employs residual connections in such a way as to give the U-Net it's 'U' shape. The encoder and the decoder are comprised of four blocks each. In the encoder blocks, the shapes of each feature map shrink down with each convolution layer, while the number of filters expand. The output of each block is pass directly to its mirror image in the decoder, as well as passed through a max pooling layer on its way to the next block in the encoder. The best way to understand this is by viewing the architecture diagram from the original paper.

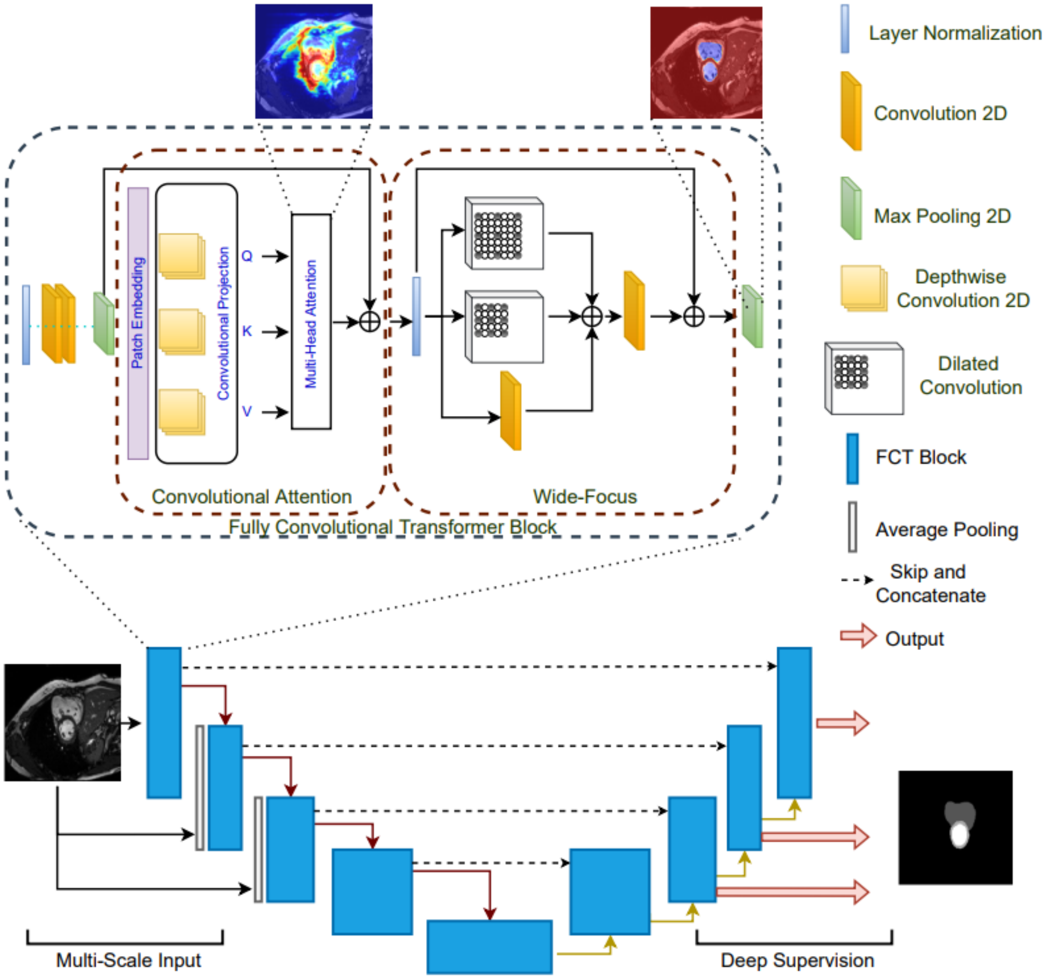


The encoder (left side of the 'U') works to capture context by learning abstract, salient features from the images. The decoder (right side of the 'U') works to enable precise localization. In the decoder blocks, upsampling is performed on the previous blocks output via transposed convolution[1] before being concatenated with the skip connection of the mirror image encoder block.

[1] See this article for an in-depth exploration of transposed convolution: Transposed Convolution Demystified | by Divyanshu Mishra | Towards Data Science

Fully Convolutional Transformer

FCT is an attempt to combine the proven success of the transformer [7] to learn long-term dependencies and context within the data, with the framework of U-Net. It is the first fully convolutional transformer-based architecture in the medical imaging literature.



The high-level structure looks very similar to U-Net. The difference is that each individual block in the encoder and the decoder are no longer sequential applications of convolution but are now novel "Fully Convolutional Transformer" blocks.

The FCT block begins by first applying layer normalization, two 3x3 convolution layers, and a max pooling layer before being passed off to the patch encoding layer. The patch encoding layer creates tokens from the images by convolving over the input with a 3x3 kernel, with 'same' padding to preserve the shape of the output, then passing them through another layer normalization operation. The patches are flattened into vectors to form the final embedded tokens.

The tokens are then passed through a modified multi-head self-attention (MHSA) module. It is modified from typical applications of MHSA by virtue of replacing linear projections with depthwise-convolutions. This better leverages contexts from images, reduces computational costs, as well as removes the need for positional encoding.

The next module in the process is the 'Wide-Focus' block. This block attempts to capture the fine-grained nature of the medical image problem by employing three distinct convolution layers that work in parallel. The output of the MHSA layer is passed to each one independently. The first performs standard convolution. The others perform a dilated version of convolution with increasing receptive fields to gain better spatial context. The output of these three layers is concatenated with the output of the MHSA layer in a skip connection before going through a final standard convolution operation and a final max pooling.

On difference to highlight between the encoder and decoder blocks. In the decoder blocks, upsampling is performed via nearest neighbors before concatenating with the skip connection from the mirror-image encoder block. It is then passed through a FCT block as normal. Also, each encoder block also as input a rescaled version of the input image, in addition to the output of the previous block. However, the authors of the paper note that the model achieves state-of-the-art results even without this feature.

Model Training Scheme

Training was performed using one NVIDIA GeForce RTX 3090 GPU for all experiments using Python 3.9.12, while all model implementations are in PyTorch. Slightly different training schemes were employed between the two models.

For U-Net, I trained for a maximum of 100 epochs, using early stopping with a patience of 5. A learning rate of 1e-4 was used with a learning rate scheduler that reduced the learning rate upon plateau. The Adam optimizer was used with otherwise default parameters. For data loading, I used a batch size of 2 images. For the loss function, a combination of Dice Loss and binary cross entropy (BCE) was used. Patience of 5 was reached on Epoch 63 with a training duration of ~12 minutes.

For FCT, I trained for a maximum of 100 epochs, using early stopping with a patience of 5. A learning rate of 1e-3 was used, again with a learning rate scheduler and the Adam optimizer. A batch size of 1 image needed to be used, even with the images rescaled to half their original size, to fit in memory during training. The loss function was binary cross entropy (BCE). Patience of 5 was reached on Epoch 46 with a training duration of ~8 minutes.

Results

Performance across five different metrics[2] was measured on the 20-image testing set for both models. Results are summarized below:

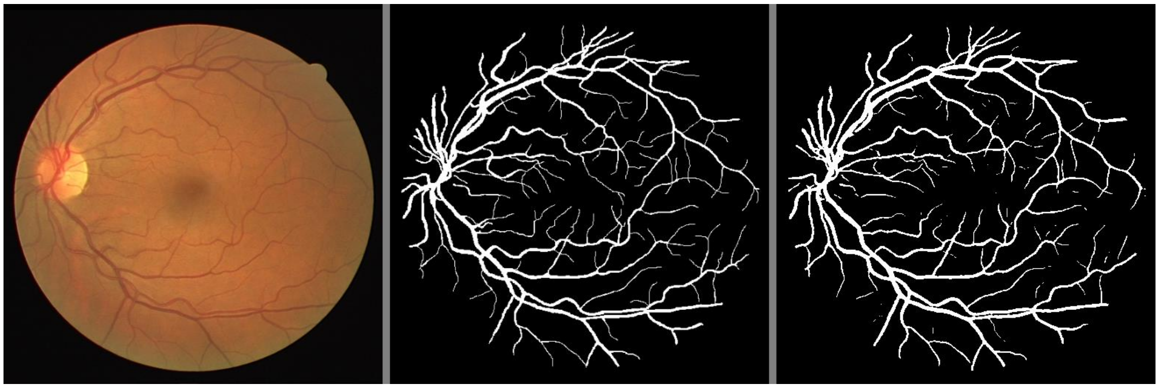
Metric	U-Net	FCT	SOTA
Jaccard	0.6662	0.5581	Not Measured
F-1 Score	0.7994	0.7154	0.8690
Precision	0.8163	0.8180	Not Measured
Recall	0.7878	0.6404	0.7927
Accuracy	0.9657	0.9581	0.9790
AUC ROC	0.8854	0.8137	0.9887

I have also included metrics for the current known state-of-the-art[3] [15] for reference. While not measuring up to SOTA, both model implementations do a reasonably good job, and diminishing returns seem to be dominating to achieve ever higher performance gains towards SOTA. My two implementations were trained using fewer data augmentations with more simple architectures. Researchers were able to improve performance by fine tuning their augmentation strategy [16].

One curious fact is that FCT seems to be performing worse than the basic U-Net despite FCT being a supposed improvement upon U-Net. My main hypothesis for this is because I needed to shrink the scale of the input training images by half for FCT to train it on my GPU. Given the main advantage of FCT is it's supposed ability to harness the signal in the fine details of the image, it would make sense that reducing the scale, and thus eliminating much of the fine-grain information, would lead to worse performance.

Below are examples of the segmentation map generated from each model, in the order of original image, target map, and predicted map:

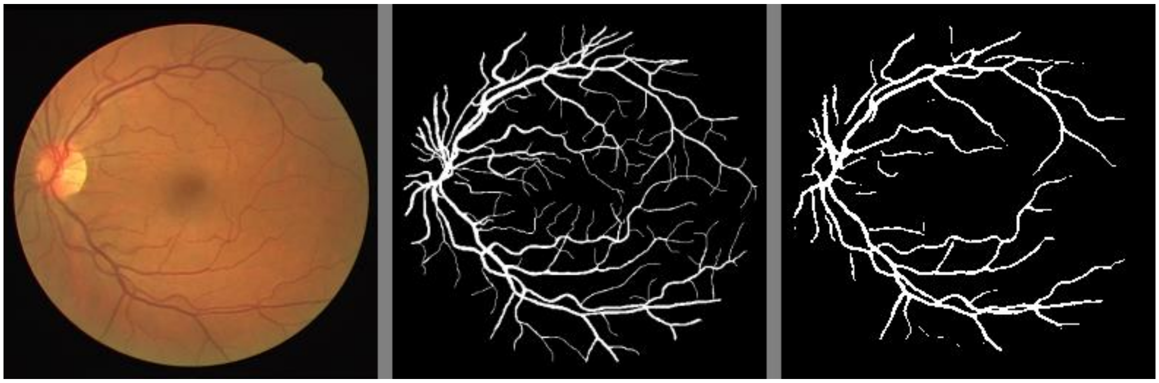
U-Net:



[2] See appendix for links to definitions of these metrics.

[3] SOTA documented on PapersWithCode: DRIVE Benchmark (Retinal Vessel Segmentation) | Papers With Code

FCT:



From this comparison, it's easy to see how FCT fails to map the finer blood vessels.

Discussion

Artificial intelligence has many important applications for the healthcare generally and optometry specifically. Further advancements in models like U-Net and FCT will contribute directly to healthcare practitioners being able to more accurately diagnose and quickly treat conditions related to the eye. As stated, being able to identify the blood vessels in retinal images, and therefore measure attributes like length, width, tortuosity, and branching patterns, can aid healthcare professionals in detecting various cardiovascular and ophthalmologic diseases such as diabetes, hypertension, atherosclerosis, and choroidal neovascularization. It can also help advance performance in computer-assisted laser surgery. I hope that my work here shows that implementing one's own machine learning system to segment blood vessels is easy to do, works fairly well, and can be done with very few labeled training examples.

References

[1] O. Ronneberger, P. Fischer and T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," Arxiv.org, 18 May 2015. [Online]. Available: https://arxiv.org/pdf/1505.04597v1.pdf.

[2] C. Kaul, S. Manandhar and N. Pears, "FocusNet: An Attention-based Fully Convolutional Network for Medical Image Segmentation," arxiv.org, 8 February 2019. [Online]. Available: https://arxiv.org/pdf/1902.03091.pdf.

[3] C. Kaul, N. Pears and H. e. a. Dai, "FocusNet++: Attentive Aggregated Transformations For Effecient and Accurate Medical Image Segmentation," arxiv.org, 7 April 2021. [Online]. Available: https://arxiv.org/pdf/1912.02079.pdf.

[4] Y. N. S. L. N. O. GM Venkatesh, "A Deep Residual Architecture for Skin Lesion Segmentation," 2018. [Online]. Available: https://doras.dcu.ie/22685/1/Deep-Residual-Architecture_Camera_Ready.pdf.

[5] C. Wang, Z. Zhao, Q. Ren, Y. Xu and Y. Yu, "Dense U-Net Based on Patch-Based Learning for Retinal Vessel Segmentation," 21 February 2019. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7514650/.

[6] Z. Zhou, M. Siddiquee, N. Tajbakhsh and J. Liang, "U-Net++: A Nested U-Net Architecture for Medical Image Segmentation," 18 July 2018. [Online]. Available: https://arxiv.org/pdf/1807.10165.pdf.

[7] A. Vaswani, N. Shazeer and e. all, "Attention Is All You Need," arxiv.org, 12 June 2017. [Online]. Available: https://arxiv.org/pdf/1706.03762.pdf.

[8] A. Dosovitskiy, L. Beyer and e. al., "An Image is Worth 16x16 Words: Transformers for Image Recognition At Scale," 3 June 2021. [Online]. Available: https://arxiv.org/pdf/2010.11929.pdf.

[9] H. Wu, B. Xiao, N. Codella and e. al., "CvT: Introducing Convolutions to Vision Transformers," 29 March 2021. [Online]. Available: https://arxiv.org/pdf/2103.15808.pdf.

[10] "Swin Transformer: Hierarchical Vision Transformer using Shifted Windows," arxiv.org, 17 August 2021. [Online]. Available: https://arxiv.org/pdf/2103.14030.pdf.

[11] J. Chen, Y. Lu, Q. Yu and e. al., "TransUNet: Transformers Make Strong Encoders for Medical Image Segmentation," arxiv.org, 8 February 2021. [Online]. Available: https://arxiv.org/pdf/2102.04306.pdf.

[12] A. Hatamixadeh, Y. Tang and V. e. a. Nath, "UNETR: Transformers for 3D Medical Image Segmentation," arxiv.org, 9 October 2021. [Online]. Available: https://arxiv.org/pdf/2103.10504.pdf.

[13] "DRIVE," 2019. [Online]. Available: https://drive.grand-challenge.org.

[14] A. Tragakis, C. Kaul, R. Murray-Smith and D. Husmeier, "The Fully Convolutional Transformer for Medical Image Segmentation," Arxiv.org, 1 June 2022. [Online]. Available: https://arxiv.org/pdf/2206.00566.pdf.

[15] S. Kamran, K. Hossain and e. al., "RV-GAN: Segmenting Retinal Vascular Structure in Fundus Photographs using a Novel Multi-scale Generative Adversarial Network," arxiv.org, 14 May 2021. [Online]. Available: https://arxiv.org/pdf/2101.00535v2.pdf.

[16] E. Uysal, B. Zaza and e. al., "Exploring the Limits of Data Augmentation For Retinal Vessel Segmentation," arxiv.org, 30 May 2021. [Online]. Available: https://arxiv.org/pdf/2105.09365v2.pdf.

Appendix

GitHub repository containing all python code (www.github.com/ddixonAI/SLU_Capstone)

Performance Metric Definitions:

- Jaccard Score
- F1-Score
- Precision
- Recall
- Accuracy
- AUC ROC

Loss Functions:

- BCE - Dice Loss
- BCE Loss (Binary Cross Entropy)