Using Vision Transformers to Automate the Diagnosis of Diabetic Retinopathy

Simon Ellershaw[1*], Konstantinos Balakas[2] and Andre Altmann[3]

[1] UCL Institute of Health Informatics, [2] Moorfields Eye Hospital NHS Foundation Trust, [3] UCL Centre for Medical Image Computing



Background

Recent convergence of vision and language models

- Transformer models have represented the state-of-the-art approach in natural language processing (NLP) since 2017 [1].
- Whereas in computer vision, convolutional neural networks (CNNs) have dominated the domain [2].
- Recently developed, vision transformers (ViT) [3] have challenged this split.

Application to medical imaging

- Through the use of transfer learning, CNN's have been successfully trained for use in medical imaging tasks, such as classifying cases of diabetic retinopathy (DR) from fundus images.
- However, the use of vision transformers in such a domain remains unexplored.

Therefore the aim of this is project is to analyse if there are benefits of a vision transformer-based approach to the DR classification task. If so this could change the underlying model used in the next generation of clinically deployed automated diagnosis tools

Methodology

Data

- eyePACS dataset [4] contains 88,702 labelled images
- Expertly graded on a scale of 0 (no DR) to 4 (proliferative DR).

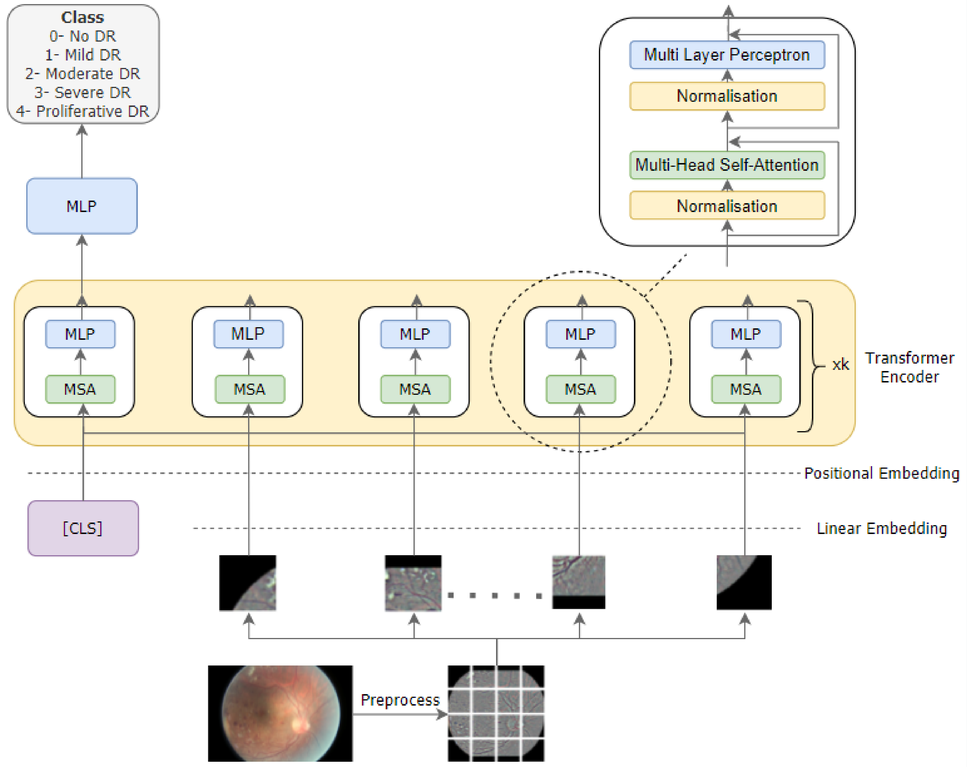
Pre-processing

- Standardise all retina radii to 500 pixels
- Subtract average local colour
- Remove boundary effects
- Rescale to 224x224
- Data augmentation includes random flips, rotations and shifts

Finetuning of models (pretrained on ImageNet 21k)

- Backpropagate batches of the training set's cross entropy loss until the validation set's loss converges
- Stochastic gradient descent optimiser with momentum 0.9
- Initial learning rate 0.001 updated over 100 epochs by cosine schedule after 10 warmup epochs
- Norm of all gradients clipped to 1

Vision Transformer



Key Properties:

1. Relaxes convolutional prior of CNNs
2. Can account for long range dependencies in early layers
3. Comparable performance to CNNs on ImageNet after large scale training
4. Faster training time due to parallelism in the architecture
5. Potential explainability via attention rollout

Attention Rollout

Given the attention matrix of each layer, A, where A_{xy} is the attention token *x* gives to token *y* .The attention rollout, $\tilde{A}$, from layers *i* to *j* is calculated recursively as

$$\tilde{\mathbf{A}} = \begin{cases} (\mathbf{A}(l_i) + \mathbf{I})\tilde{\mathbf{A}}(l_{i-1}) & \text{if } i > j, \\ \mathbf{A}(l_i) & i = j. \end{cases}$$

Evaluation

Compare ViT and CNN trained on the eyePACs DR dataset on the basis of

1. Accuracy- Quadratic weighted kappa
2. Efficiency
   i. Data- Performance with varying amounts of training data
   ii. Compute- FLOPs/average epoch time
3. Explainability- Comparison with ground truth saliency maps

Current Progress

- Data access secured
- Image pre-processing implemented
- Resnet50 model trained on UCL's HPC

References

1. Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. arXiv preprint arXiv:1706.03762, 2017
2. Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classifica-tion with deep convolutional neural networks. Advances in neural information processing systems, 25:1097–1105, 2012.
3. Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xi-aohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, GeorgHeigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale.arXiv preprint arXiv:2010.11929, 2020.
4. Diabetic Retinopathy https://www.kaggle.com/c/diabetic-retinopathy-detection/data. Accessed:04-02-2021.

*Corresponding author email: simon.ellershaw.20@ucl.ac.uk