Glaucoma-Attention: An Advanced Vision Transformer Approach for Automated Glaucoma Detection using Fundus Images

Imed-Eddine Haouli [1] Walid Hariri [2] Hassina Seridi-Bouchelaghem [3]

Deep Learning Indaba Accra - Ghana / 3rd to the 9th September 2023

Abstract

Glaucoma, a challenging disease that can cause irreversible blindness if not detected early, has garnered significant interest in utilizing deep learning and medical imaging for automated diagnosis and categorization. This poster presents a novel approach termed **GlaucomaDetector** for automatic glaucoma detection by leveraging the advancements of Vision Transformers (ViT), which have shown excellent performance in computer vision tasks following their success in NLP. Although CNNs are widely acknowledged as one of the best techniques for this task, they still face certain challenges. Therefore, extensive hyper-parameter optimizations are performed during the training phase to determine the optimal settings for each ViT model. Subsequently, a comparison is made between the results obtained from two top-performing CNN models. The study employs transfer learning on five retinal fundus image datasets. Experimental results conducted on the combined datasets demonstrate that the proposed ViT-based model is an effective tool for glaucoma detection, achieving an average accuracy of 94.13% using ViT-B32.

Introduction

Glaucoma disease is a leading cause of blindness worldwide, causing nerve fiber degeneration and causing blindness. Between 2010 and 2020, the number of glaucoma cases increased from 60.5 million to 85.6 million, with 59.5% of cases being women and 47.7% Asian. It is known as the "**silent thief of sight**" due to its gradual destruction and irreversible damage before vision loss occurs . Glaucoma is a serious eye disease if not detected and treated early. Manual diagnosis of glaucoma using color fundus imaging is challenging due to the limited availability and cost of the technology.

To address this issue, several automatic methods have been proposed to detect glaucoma based on traditional handcrafted-based approaches. However, these methods have some shortcomings, such as the need for a high level of expertise to analyze the images. Recently, deep learning (DL) has emerged as a promising approach for automatic glaucoma detection. DL methods are able to learn features from the data automatically, which eliminates the need for handcrafted features. This makes DL methods more robust and easier to use. DL methods have shown promising results in glaucoma detection, and they are expected to play an increasingly important role in the future of glaucoma diagnosis.

We present, a new ViT-based method to detect glaucoma disease [1]. To prove the superiority of the proposed method compared to top-performing CNNs, several experiments have been conducted. Accordingly, a transfer learning scenario called Non-trainable has been carried out on five retinal fundus image datasets with hyper-parameters tuning.

Objectives

This project involves creating a computer application called **GlaucomaDetector** to assist ophthalmologists in their diagnostics. Since this disease is difficult to detect at an early stage, it has become necessary to find a more effective alternative. Our goal is to provide an automatic glaucoma detection application using fundus images. After validating the proposed application on five benchmark databases, it was found that this detection is almost immediate with very high accuracy.

Contribution

The innovation brought by this project lies in the application of a very recent deep model based on (Vision Transformer - ViT) instead of conventional CNNs. Transformers were primarily proposed for natural language processing (NLP) in 2017 by Google's research team. Building on the remarkable success of this new approach in representing long-range dependencies in long sequences (words, sentences), vision transformers (proposed in 2021) have become highly competitive compared to CNN models in the field of image classification and have achieved superior performance. Furthermore, we have used a visualization technique called Grad-Cam to make the results more interpretable.

Method

To implement our method, the following steps are carried out:

- Data Collection: five benchmark datasets are combined to increase the model's robustness to variations in the data and to enhance data diversity. The datasets are: ACRIMA, RIM-ONE, Drishti-GS1, HRF, and SJCHOI86-HRF.
- Pre-processing: Image resizing by respecting the input default size for each model.
- Data augmentation on the fly during the training process by performing random transformations such as translation, horizontal flips, and zooming.
- Train models: Many ViT and CNN models have been trained on adopting the proposed training scenario called Non-trainable.
- Models explainability: a Visualization technique called Grad-Cam has been employed to study the receptive field of the models.

Experiments

To study the feature extractor's ability of ViT models over CNN models, many experiments have been conducted on adopting the proposed training scenario, Non-trainable. Furthermore, to evaluate the performance of our proposal from another point of view, different batch sizes have been tested with all models. Table 1 shows our data arrangement.

Table 1. Data arrangement

Dataset	ACRIMA	Sjchoi86-HRF	RIM-ONE	Drishti-GS1	HRF	Total
Train (80%)	564	321	364	81	36	1366
Test (20 %)	141	80	91	20	9	341

Table 2 displays the hyper-parameters used where the effect of the batch size is tested using six different values, whereas the rest of the hyper-parameters are fixed.

Table 2. Hyper-parameters.

Param	Batch size	Optimizer	Epochs	Learning rate
Value	32, 64, 128, 256, 512, 1024	Adam	30	0.001%

- **Training scenario** : The scenario named **Non-trainable** that consists of freezing all the blocks / layers of the feature extraction part, whilst in the classification part, the default layers have been replaced by two dense layers of 256 and 128 nodes, each followed by the dropout layer. Finally, an output layer comprises two nodes representing our classes. The value of dropout used is 0.5.
- **Models evaluation**: We have employed the **10-Fold cross-validation** method.
- **Technical specifications**: We have used Google Colab environment, Python 3.10 language, Tensorflow / Keras framework, and Tesla T4 GPU.

Results

Figure 1 compares the performances of ViT and CNN models using different batch size values to investigate their impact on the accuracy and training time of each model. This Figure shows that the ViT models outperform the CNN models in all the different batch size values. More specifically, the ViT-B32 model gives the best performance by 94.13% when being trained with a batch size of 32. It also gives a high accuracy when trained with 128 and 1024 batch size values of 90.90% and 91.20%, respectively. The subfigure on the right shows that the ViT-B32, VGG19 and ResNet152V2 models are faster than the other models for all the batch size values. Moreover, using all the models, the training time decreases when the batch size value is augmented.

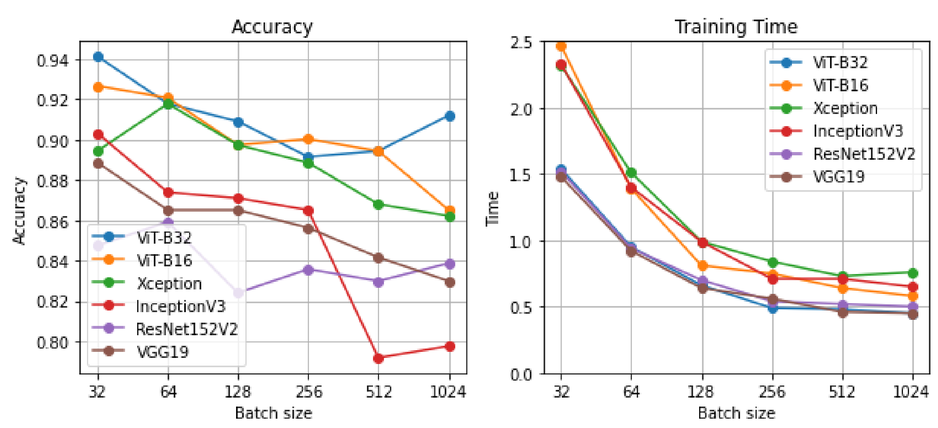
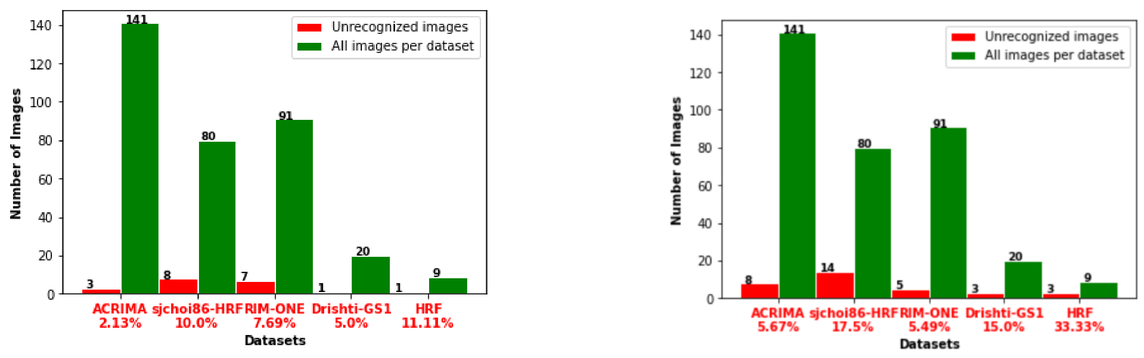


Figure 1. Effect of the batch size on the models' accuracy (Left) and training time (Right)

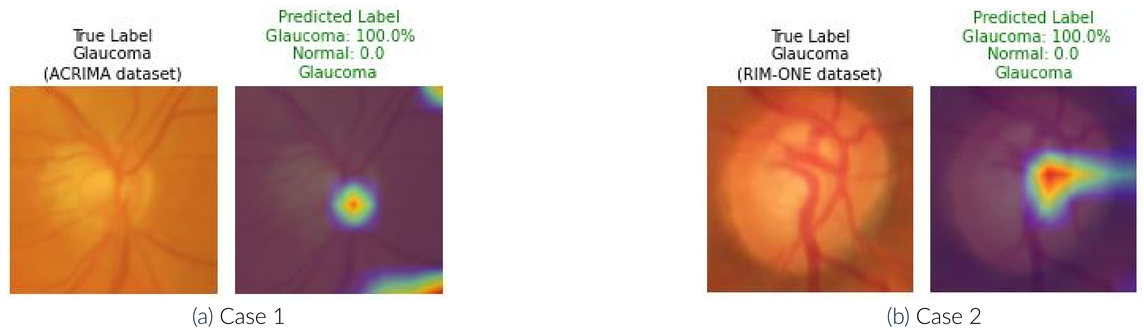
Figure 2 presents the well / wrongly recognized images per dataset for the two best models. From this figure, we can see that the Sjchoi86-HRF dataset is more challenging than the other datasets for the two models, ViT-B32 and InceptionV3, by 8 and 14 wrongly recognized images, respectively. We can also notice that the ViT-B32 model gives the best results in most datasets compared to Xception.



(a) ViT-B32. (b) Xception

Figure 2. Comparison of best models per dataset based on prediction results

The Grad-Cam visualization presented in Figure 3 mainly locates the optic disk and cup regions which are the areas responsible for this disease, it is considered as an automatic segmentation.



(a) Case 1 (b) Case 2

Figure 3. Grad-Cam visualization using ViT-B32 model

Deployment

We have deployed our high-performance ViT model, ViT-B32, in a web application using the TensorFlow.js framework. Figure 4 displays some windows of our applications

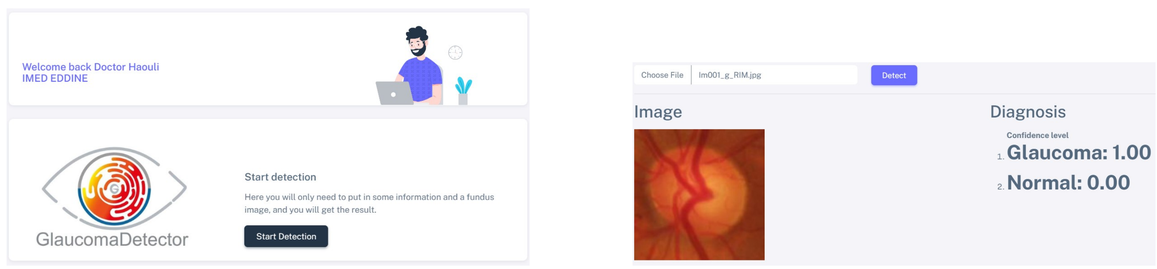


Figure 4. Main window (left) Prediction window (right)

Conclusion and Future Work

The experimental results show that the best model is ViT-B32 in terms of accuracy. These achievements are due to the **self-attention mechanism**, which is the main component of these models, giving them more robustness to scale, translation, and hyper-parameter changes. Additionally, The ViT models shown are steadier than CNN models over different batch size values. Moreover, the combination of five different datasets doesn't affect the overall performance of the proposed transformer-based network. This result further proves the proposed architectures' robustness and important anti-interference ability. As a perspective, we look to combine ViT and CNN models to drive decisions, and applying other visualization methods to further understand the receptive field of the models.

References

[1] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al.
An image is worth 16x16 words: Transformers for image recognition at scale.
arXiv preprint arXiv:2010.11929, 2020.