Eye Disease Classification Using Vision Transformer: A Deep Learning Approach

Jayakrishna Vuppalapati
Dept. of Computer Science
University of South Dakota
jayakrishna.vuppalap@coyotes.usd.edu

Tinku Rao Kotha
Dept. of Computer Science
University of South Dakota
tinkurao.kotha@coyotes.usd.edu

Krishna Prithvi Battula
Dept. of Computer Science
University of South Dakota
krishnaprithvi.battu@coyotes.usd.edu

Praveenkumar Kavali
Dept. of Computer Science
University of South Dakota
praveenkumar.kavali@coyotes.usd.edu

Abstract—**Timely and accurate identification and diagnosis of retinal diseases are essential to prevent eventual vision loss, which may be temporary or permanent. Here, we present a Vision Transformer model, called retinal ViT, which improves medical image processing using the self- attention mechanism. The key purpose of this study is to determine the transformer-based model to substitute convolutional modules used in CNN models while matching or surpassing their performance. Retinal ViT, differs from traditional CNN-based convolutional neural networks through its usage of self-attention to capture longer dependencies in retinal images. This approach builds on transformer's capacity to identify complicated patterns and connections over spatial axes, reportedly increasing the model's discriminating power across retinal conditions. Ultimately, the retinal ViT architecture is framed around a multi-class classification into which the predictor feeds and assigns a given retinal image to one of many illnesses based on presence. This classification consists of simplified activations via sigmoid functions to predict multiple tags concurrently. The classification psyches use the capabilities of deep neural network's capacity for expressive power to predict retinal disorders with high confidence. In our experiments, extensive analysis of the ViT's performance will be done on the most used public dataset, which contains a wide variety of retinal images. Ultimately, substantial comparisons with contemporary CNN models were done to illustrate comparative metrics for both models. In summary, the proposed retinal ViT approach will show promising hope for medical analysis of images with a signal of how the self-attention capabilities foster potential methodologies to improve transformational model performance for data perception and management functions. Thus, this study lays a foundation for future enhancements concerning self-attention for accurate and efficient detection and treatment of retinal disorders.**

Keywords—Vision Transformer, Deep Learning, Eye Disease Classification, Image Classification, Healthcare, Pytorch

I. Introduction

Eye diseases represent a serious health risk worldwide and could lead to vision loss or blindness, if left undiagnosed and untreated. To avoid permanent vision impairment, prompt and accurate diagnosis is crucial. Ophthalmologists in the past have diagnosed eye diseases with manual interpretation of medical records. However, it can be time consuming and be susceptible to human error.

Deep learning, a branch of artificial intelligence that has the potential to develop automated systems which would help doctors diagnose diseases more effectively, especially in medical image analysis, is emerging as an important tool for this.

Vision Transformers, a kind of deep learning architecture, have demonstrated encouraging outcomes in a variety of image classification tasks. Vision Transformers, as opposed to conventional convolutional neural networks (CNNs), use self-attention processes to identify global relationships in images, which makes them ideal for applications like medical image analysis where context is crucial. This will make them especially suitable for applications such as medical image analysis, where it is essential to understand the context of different images to diagnose accurately.

II. Literature Review

Various approaches to classifying eye diseases, ranging from traditional machine learning techniques to more recent methods of deep learning have been examined in previous research. While traditional methods have achieved some success, they often rely on handcrafted features extracted from images. These features are not able to precisely replicate the detailed patterns found in healthcare images, leading to a limitation of accuracy.

As a powerful tool for medical image analysis, deep learning approaches, in particular convolutional neural networks (CNNs) have emerged. CNNs can automatically learn relevant features directly from the data, overcoming the limitations of handcrafted features. But CNNs have their drawbacks, for example the difficulty of capturing dependencies over time in images. Vision Transformers (ViTs) address this challenge. In medical image analysis tasks such as diagnosis of eye disease, they use self-attention mechanisms to analyze relationships between different parts of the image, which allow them to detect complex and long-range critical dependencies for accurate classification.

III. METHODOLOGY

A dataset of eye photographs was used in this study covering a wide variety of ocular diseases, including cataracts, diabetes retinopathy, glaucoma, and healthy vision. Training and test sets for training and evaluation of the performance of our model have been set out in this dataset.

We used a Visual Transformer Technology architecture and pretrained it on a large-scale image dataset. Pretraining enables the model to be trained with essential image representations from a large database, enabling it to adapt its specialized task of classifying eye diseases.

We used preprocessing techniques to standardize input pictures before the model was trained. To ensure consistency of the data that is sent to a model, techniques such as resizing, and normalization have been used. In addition, to artificially expand the dataset, we used data augmentation techniques. To increase the model's ability to deal with unseen changes during testing, it was necessary to generate variations of existing images, such as rotations and flips.

We used the Adam optimizer, a popular optimization algorithm, to update the model's internal parameters during the training process, and to minimize errors in classification. The cross-entropy loss function, a common metric used to measure the difference between models' predictions and actual labels, has also been employed. To optimize the performance of this model, we have adapted hyperparameters such as learning speed and batch size.

IV. EXPERIMENTS

To assess the performance of our Vision Transformer model in a classification task for eye diseases, we have carried out studies. The experiments were carried out on a computer with sufficient computational power to carry out the training efficiently. We used standard performance metrics common to image classification tasks when evaluating the model's performance. These metrics include:

- **Accuracy:** The overall percentage of correctly classified images is measured in this metric.
- **Precision:** Among all positive predictions by the model avoiding false positives, this metric considers a proportion of real negatives.

In both training and testing datasets, we assessed the model's capabilities. This enabled us to evaluate the model's ability to learn from training data and generate a good understanding of unknown data during testing.

V. RESULTS

The performance of the model is assessed using a set of independent images which are not observed during training. To evaluate how well a model can classify images into the correct disease categories, performance metrics such as accuracy, precision are calculated.

Loss Curve: The loss curve shows how well our computer program is learning, as it looks at more examples. Loss is a measure of our program's predictions being incorrect. In the long run, we'd like to see our losses decrease to show that our program has become more efficient at accurately predicting events. That means our program is learning well if the loss curve continues to fall steadily. But if it's going up or down in an inconsistent manner, that suggests our program may not be able to learn effectively.

Accuracy Curve: The accuracy curve shows us how often our program makes correct predictions as it learns. The accuracy of our predictions compared to the overall number we make is a measure of how many are accurate. To show that our program is getting better at detecting eye diseases, we want it to get even more precise over time. If the accuracy curve continues to rise, then our program will improve its detection of diseases.

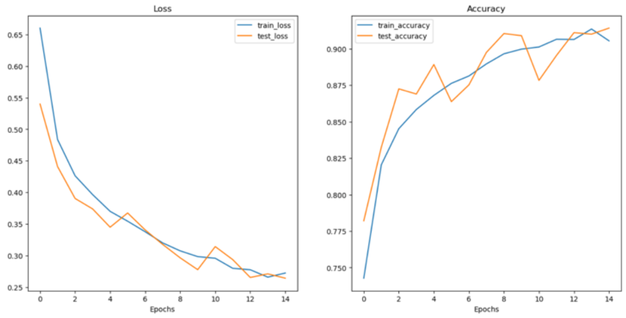

Fig. Plot of Loss and Accuracy achieved over 10 epochs.

To predict the classes of individual images, a trained model is used. For example, the model can predict whether an eye has cataract, diabetic retinopathy, glaucoma, or normal vision based on its input image. The ability of a model to generalize to new, unseen images is assessed by these predictions.

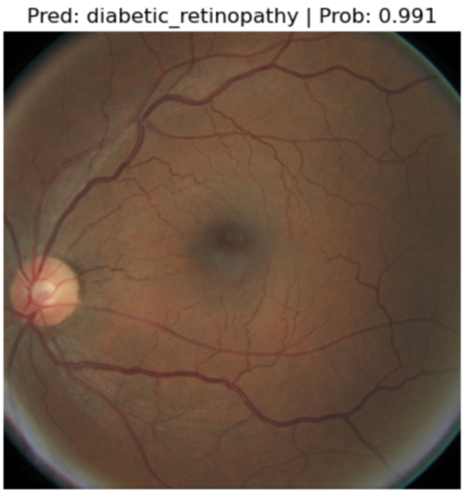

Fig. ViT prediction for Diabetic Retinopathy class

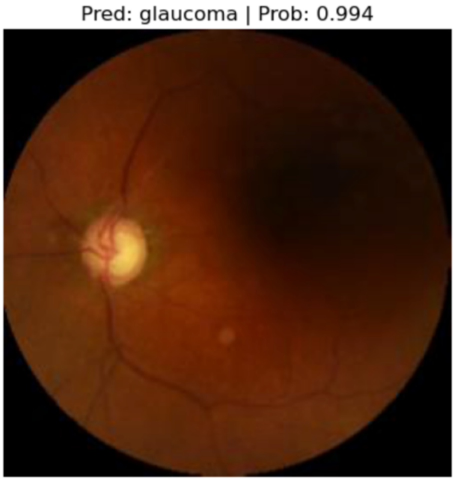

Fig. ViT prediction for Glaucoma class

VI. DISCUSSION

The effectiveness of Vision Transformers for the accurate classification of eye diseases has been shown by our experiments. To help ophthalmologists diagnose eye diseases, our model has achieved high accuracy and promising performance metrics suggesting its potential for real world applications in clinical practice. However, some limitations and challenges related to this approach were also found during the experimental phase.

- **Data Availability:** To train robust Deep Learning Models, large and varied data sets are essential. The size or diversity of the dataset used may have been a limitation of our study. The model's generalization capability could be improved by a larger and more complete set of data, which covers a wide range of eye diseases and variations.

- **Computational resources:** In many cases many computing resources are needed to train complex deep learning models, like ViTs. This may be an obstacle to wider adoption, particularly in context of limited resources**.**

Despite these challenges, our study showed a great potential for classifying eye disease through Vision Transformers. Several benefits could be achieved by the possibility to automate certain parts of diagnostic procedures that may assist healthcare professionals in making important decisions, for example:

- **Improved patient outcomes:** Early and more accurate diagnosis may lead to timely treatment interventions, which could improve the outcome of patients or prevent vision loss.
- **Reduced healthcare costs:** Early detection and treatment may reduce the need for complex and costly interventions later in life, leading to reduced health care costs.

To address the identified challenges and explore other ways of improvement, further research is necessary:

- **Data collection and sharing:** Efforts to collect, share large, varied or well anonymized datasets specific for classification of eye diseases would be useful. This would facilitate the development of even more robust and generalizable models.
- **Model Optimization:** Research into optimizing Vision Transformer architectures for medical image analysis tasks, potentially reducing computational requirements, could make this technology more accessible and widely applicable.

VII. CONCLUSION

Finally, the potential of Vision Transformers as an accurate diagnostic tool for eye diseases was demonstrated by this study. We can develop efficient and effective systems for early diagnosis and treatment of various eye diseases by using ViT and pretrained models. The importance of ongoing research and development in the field of medical image analysis is highlighted by these findings. We can continue to advance this area and improve the health outcomes of patients around the world using advanced deep learning architectures such as Vision Transformers.

VIII. FUTURE WORKS

The following possibilities could be considered as future work for this project:

- To explore the use of transfer learning with different pretrained models to see if there is an improvement in performance.
- The impact of different data enhancement techniques on model performance shall be investigated.
- To further demonstrate its effectiveness, the model must be tested on a more extensive and diverse data set.
- Developing a user interface for the ophthalmologists so they can use this model in their practice.

These areas will allow us to enhance the use of Vision Transformers for classification of eye diseases, thus leading to a more effective and precise diagnosis process in patients.

REFERENCES

[1] Dataset link: https://www.kaggle.com/datasets/gunavenkatdoddi/eye-diseases-classification

[2] Dong Wang, Jian Lian, Wanzhen Jiao. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10800810/

[3] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., ... & Houlsby, N. (2020). An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. arXiv preprint arXiv:2010.11929. https://paperswithcode.com/paper/an-image-is-worth-16x16-words-transformers-1

[4] https://www.v7labs.com/blog/vision-transformer-guide

[5] https://ieeexplore.ieee.org/document/9956086/

[6] PyTorch. (2021). TorchVision Models. Retrieved from https://pytorch.org/vision/stable/models.html