Retinal Disease Multiclass Classification

Bhagesh Gaur
bhagesh20558@iiitd.ac.in

Madhava Krishna
madhava20217@iiitd.ac.in

Sanyam Goyal
sanyam20116@iiitd.ac.in

Abstract

Retinal diseases like diabetes, glaucoma and age-related macular degeneration (ARMD), among others, can cause loss of vision and permanent blindness. As a result, it is imperative that early detection and diagnosis be carried out effectively and efficiently.

We aim to tackle a multilabel classification problem on the ODIR-2019 dataset, made public by Peking University, China. We implement known and new deep-learning algorithms on the dataset after proper processing and examine how transformations and model layouts change the performance on the dataset. We conducted benchmark experiments on deep neural networks baselines. We discovered that merely increasing the size of the network did not lead to significant improvements in multi-disease classification. We also found that a well-structured feature fusion method that effectively combines the characteristics of multiple diseases along with a transformer based encoder achieves good results.

1 Introduction

Fundus diseases affect millions of people worldwide. They include diseases like diabetic retinopathy, age-related macular degeneration, cataract, glaucoma and hypertensive retinopathy, among others. India has a high prevalence rate of 16.9% for diabetic retinopathy (Vashist, 2021), and 12 million people suffer from glaucoma (George R, 2010). An estimated 100 million people go blind from cataract and 8 million from age-related macular degeneration (ARMD) every year (Sen M, 2022).

Machine learning-based tools can help ease the load on the healthcare sector by assisting in diagnosis to improve accessibility and, in some cases, the accuracy of the treatment method prescribed. Deep-learning-based techniques allow for an even greater level of flexibility in the diagnostic procedure, allowing for multimodal diagnosis and taking inputs from a variety of sources (Ngiam et al., 2011).

We explore various deep-learning algorithms and techniques on the ODIR-19 dataset, provided by the Peking University, China, which has 8 classes of ocular diseases with corresponding fundus images. (ODIR2019, 2019)

2 Related Work

Corbella's MSc dissertation(Coll Corbilla, 2020) had done single-eye multilabel classification. The repository is available and featured transfer-learning methods on models like VGG16, VGG19 and InceptionV3. The full method is well-documented and available on the repository.

Wang *et al.* (Wang et al., 2020) created a similar classifier but using the EfficientNet architecture. They compared various transfer-learning approaches and gave metrics for evaluation. They even used an ensemble-learning approach wherein they co-used EfficientNet-based classifiers on RGB and Grayscale histogram-equalised images respectively, averaging the output at the terminus for class-wise prediction in a multilabel format.

Jinke *et al.* created a multilabel classifier using graph convolutional networks (Lin et al., 2021). They use the ODIR, SSL and GTest datasets, but did not mix the datasets. Li *et al.* used attention mechanisms on binocular fundus images on the ODIR dataset (Li et al., 2022). They used a ResNet50 model in a multilabel fashion.

3 Dataset details

The dataset was taken from the ODIR-2019 challenge by Peking University, China. The dataset instance used was taken from Kaggle (Larxel, 2021). The dataset has a main directory comprised of preprocessed images of dimension 512 x 512, in RGB colour format. It also has two additional directories for the test-train split (pre-split), but we do not use that in our methods. There is an excel file having the dataset description 1.

Feature	Description
ID	Identifier
Age	Integer
Sex	Male or Female
Eye	Whether left or right eye
Keywords	Diagnosis keywords
Diagnostic Columns	Features: N (normal), G (glaucoma), D (diabetes), C (cataract), A (ARMD), H (hypertension), O (others)

Table 1: Dataset description

3.1 Preprocessing stages

We focused on a single-eye image diagnosis. A patient may have multiple defects, which require parsing the diagnostic labels provided to determine the class.

With assistance in code by Jordi Corbella (Corbella, 2020), we were able to process each eye independently. There were some images that were not present in the dataset, so we excluded them. We blacklisted images as specified in his GitHub repository (low-quality and unavailable images in the CSV).

4 Methodology

The baseline systems consisted of an Inception V3 architecture and an ensemble of EfficientNet B3 models trained on RGB and histogram-equalised images. For the Inception V3 baseline, the right-eye images were horizontally flipped to simulate left-eye images.

The DeiT transformer model is an efficient transformer relying on knowledge distillation requiring less data and compute to train. The model architecture consists of a DieT-base unit for encoding the text, 1x1 convolution network on top of the encodings, after which the convolved features are flattened and concatenated with the age and sex information. The vector is then passed through a fully connected network for classification. ReLU activations were used for all convolutional and the output fully connected layer.2

We explored transformer-based approaches to classification and attempted to include the age and sex information of the patients as well. The proposed architecture is based on an ensemble of Swin-T transformers (Jia et al., 2021) with the age and sex information added to the generated image embeddings in an early-fusion fashion 1. The pooled outputs from the transformer are concatenated with the additional age and sex features and fed into a classification fully-connected network powered by ReLU activations.

The Swin Transformer is engineered to possess a high degree of scalability, allowing it to handle large input images or sets of images with remarkable efficiency. This attribute is especially valuable in scenarios where high-resolution images must be processed, enabling the model to effectively manage the computational burden of analyzing vast amounts of visual data. It has a hierarchical structure which enables it to process both local and global features. We thought this would be instrumental to our use case of classifying fundus images.

5 Experimental Setup

For all models, we used the Adam optimiser with a learning rate of 0.00001. We used a weighted BCE loss with weights $[1, 1.2, 1.5, 1.5, 1.5, 1.5, 1.5, 1.2]$ respective to N, D, G, C, A, H, M and O classes.

The training was done on Google Colab notebooks and on a system with a Ryzen 9 CPU and an RTX 3090 graphics card.

For EfficientNet B3, we performed ensembling as given in the paper. One of the models was for RGB colour images, and the other used images which underwent histogram equalisation to enhance them.

For the transformer models, we used 224x224 RGB (or Histogram Equalised images, as per requirement). The parameters For each architecture, the best-performing model was saved via an early-stopping mechanism. The reported metrics depict a 'micro' style averaging, not taking into account the class distributions.

The hyperparameters for DeiT and Swin-T can be found in 2.

6 Observation and Future Work

6.1 Observations

We observe that the convolution-based ensembled EfficientNet B3 architecture beat the proposed architectures handily. Despite varying the underlying transformer model, the performance did not surpass that offered by the CNN-based architectures. This could be because the model would be failing to extract features of relevance from the ocular scans. The results can be found in table 3.

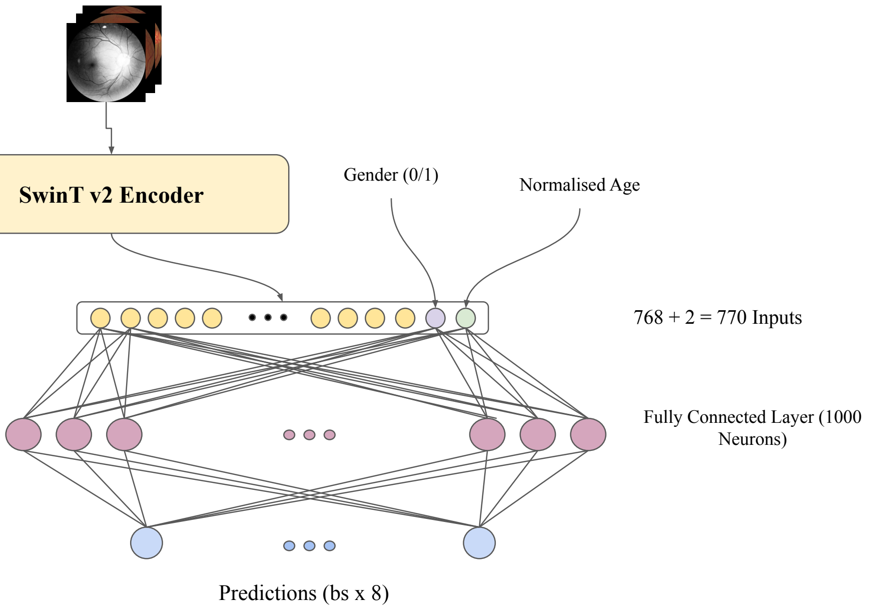

Figure 1: Proposed Architecture for the Swin-T Ensemble Component. Two of them were used in conjunction. *bs* denotes the batch size. Embeddings from the Swin transformer get concatenated with the age and gender features and passed onto a fully connected, ReLU-activated network.

Swin-T and DeiT Hyperparameters	
Hyperparameter	Value
Input Size	224x224
Transforms	-
Batch_Size	32
Epochs	20
Early Stopping	Patience : 4
Optimizer (Adam)	lr = 1e-5
Loss	BCE Loss

Table 2: Hyperparameters for the Swin-T and the DeiT models.

6.2 Error Analysis

Comparing images could not be possible because of the nature of the problem: it is fairly nontrivial to visualise multilabel images in a report. Hence, we attach a confusion matrix depicting the performance.

We observe that for some classes, the loss is negligible: the confusion matrices (available in the appendix in the plots directory) had very minor false classifications: true for glaucoma, cataract, ARMD, myopia. We observe that the images for the normal and diabetes classes could have been improved upon: so perhaps placing a higher weightage on the losses of those components may provide a better result.

The dataset was fairly imbalanced with the normal, diabetes and other diseases dominating, so augmenting the data could with a greater number of relevant transformations could also be explored in the future.

6.3 Future Work

We observe interesting results here: the transformer based-models usually perform lower than convolutional-neural networks. The next step would be discovering why and how this is happening precisely, and if possible, circumvent that by making necessary changes to the architecture.

References

Jordi Coll Corbilla. 2020. Reconeixement intel·ligent de malalties oculars mitjançant arquitectures d'aprenentatge profound.

Jordi Corbella. 2020. Ocular disease intelligent recognition through deep learning architectures. `shorturl.at/oxJQZ`.

Vijaya L George R, Ve RS. 2010. Glaucoma in india: estimated burden of disease. *J Glaucoma*, 19(6):391–7.

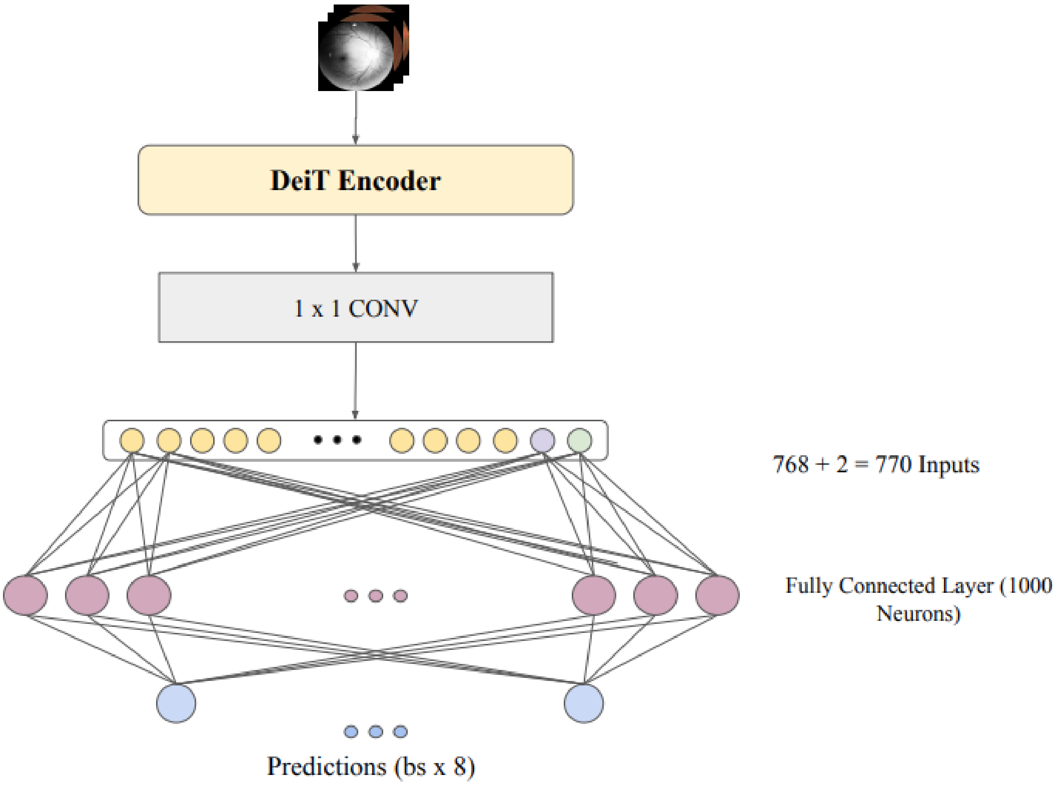

Figure 2: Proposed Architecture for the DeiT component.

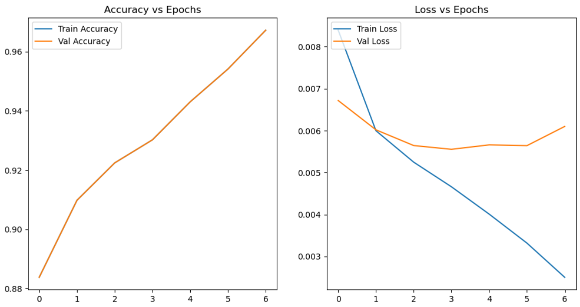

Figure 3: Loss and Accuracy curves for Swin-T

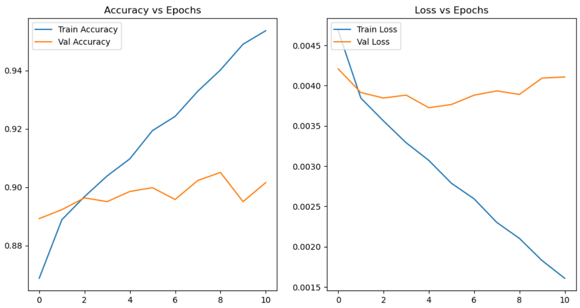

Figure 4: Loss Curve for DeiT model

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V. Le, Yunhsuan Sung, Zhen Li, and Tom Duerig. 2021. Scaling up visual and vision-language representation learning with noisy text supervision.

Larxel. 2021. Ocular disease recognition. https://www.kaggle.com/datasets/andrewmvd/ocular-disease-recognition-odir5k.

Zhenwei Li, Mengying Xu, Xiaoli Yang, and Yanqi Han. 2022. Multi-label fundus image classification using attention mechanisms and feature fusion. *Micromachines*, 13(6).

Jinke Lin, Qingling Cai, and Manying Lin. 2021. Multi-label classification of fundus images with graph convolutional network and self-supervised learning. *IEEE Signal Processing Letters*, 28:454–458.

Jiquan Ngiam, Aditya Khosla, Mingyu Kim, Juhan Nam, Honglak Lee, and Andrew Y Ng. 2011. Multimodal deep learning. In *Proceedings of the 28th international conference on machine learning (ICML-11)*, pages 689–696.

ODIR2019. 2019. https://odir2019.grand-challenge.org/.

Honavar SG Sen M. 2022. Eye care for all.

Suraj S; Gupta Vivek; Manna Souvik; Gupta Noopur; Shamanna B R1; Bhardwaj Amit; Kumar Atul; Gupta Promila2 Vashist, Praveen; Senjam. 2021.

Model Performance						
Metrics	**Accuracy**	**Precision**	**Recall**	**F1-Score**	**AUC**	**#Parameters**
InceptionV3 (Baseline)	0.8984	0.6021	0.552	0.8984	0.8838	27,218,640
EfficientNet B3 Ensemble (Baseline)	**0.929**	**0.802**	0.578	**0.672**	0.779	**21,416,336**
DeiT Convolution (Proposed)	0.901	0.637	0.498	0.559	0.918	87,031,068
Swin-T Ensemble Median (Proposed)	0.920	0.725	**0.588**	0.648	**0.945**	56,714,324

Table 3: Comparative performance of the models. While the number of parameters in the baseline models is lower, they took significantly more time to train. The EfficientNet ensembles took almost 20 minutes to train compared to the Swin-T Ensemble, which took half of that and did not saturate the memory as much.

Prevalence of diabetic retinopathy in india: Results from the national survey 2015-19. 69(11):3087–3094.

Jing Wang, Liu Yang, Zhanqiang Huo, Weifeng He, and Junwei Luo. 2020. Multi-label classification of fundus images with efficientnet. *IEEE Access*, 8:212499–212508.

7 Contributions

While the work was shared equally, these are the significant contributions:

- Bhagesh: EfficientNet B3, Swin-T transformer, Literature survey.
- Madhava: Preprocessing, DeiT, exploring models.
- Sanyam: EfficientNet B3, Inception V3.

8 Appendices

- Code folder: Google Drive
- Confusion Matrix