CS229 Final Project: Deep Learning for Glaucoma Detection

Anne-Louise Kopff
Stanford University
MS in Comp. and Math. Engineering
alkopff@stanford.edu

Anh Hoang Nguyen
Stanford University
MS in Computer Science
anhng@stanford.edu

Qianzhong Chen
Stanford University
MS in Mech. Eng.
qchen23@stanford.edu

Abstract

This paper is our final report for the project of the CS229 (Machine Learning) class at Stanford University. Computer vision techniques have been widely employed to solve various image analysis problems, especially in the world of medecine. Deep learning and computer vision provide precious tools for early diagnosis of multiple diseases using medical imagery. In this project, we focus on developing models to detect glaucoma using OCT eye fundus images.

Our dataset consists of 650 labeled eye-fundus images and a small table of extracted features for each of the images, and both the images and the extracted features were used to run the deep learning models. Model evaluation was performed using F1 score, precision and recall metrics.

1. Introduction

Glaucoma, an asymptomatic eye disease and a leading cause of irreversible blindness, poses a significant health concern, with a projected number of patients of 112 million by 2040. This chronic neuropathy induces structural optic nerve damage with visible changes on an eye fundus image, ultimately leading to functional vision loss. Artificial intelligence offers the potential to improve diagnosis and screening for glaucoma with minimal reliance on human input. This project aims to leverage the capabilities of Deep Learning for binary classification of glaucoma using Optical Coherence Tomography (OCT) eye fundus images. The core methodology involves converting these OCT images into arrays of pixel values, serving as input for our DL models. Our dataset is composed of 650 OCT eye-fundus images. The main challenge faced in this project is the small amount of data available. Even if the final accuracy obtained is not enough for a direct medical application, the exploration of common techniques to deal with small amount of data made this project very interesting and made us learn a lot about this common problem faced in Deep Learning. We explored different approaches such as dataset splitting to deal with this limitation. We also tried the new powerful machine learning architecture Transformer and achieved great results. The evaluation metrics for our models are the F1 score, precision, and recall, which are critical in assessing the balance between the model's sensitivity and specificity, especially in a medical diagnostic context. This report will describe previous work done on the topic, present our dataset and detail the methodologies employed and the challenges faced. We will then describe the solutions implemented, ultimately presenting the results of applying Deep Learning to the early detection of glaucoma through OCT image analysis.

2. Related Work

Since early glaucoma detection is key to prevent optic nerve damage leading to blindness, computer vision on eye fundus images has been increasingly used in recent years. If the results aren't yet as reliable as the human eye, the process is automatic and wouldn't require the input of an experimented specialist. The glaucoma could be directly predicted from the medical image itself. The damage caused by glaucoma, such as neuroretinal rim thinning around the optic nerve head, can be quantified in fundus photos by measuring the vertical cup-to-disc ratio (VCDR). An elevated VCDR is considered suspicious for the glaucoma diagnosis. In this section, I will describe a few papers and the methods implemented to detect glaucoma on similar images then the ones we use in our models.

2.1. Cup-Disk Ratio Regression and ResNet classification

In the paper 'Deep learning on fundus images detects glaucoma beyond the optic disc' [9], the authors use a dataset of 37,627 stereoscopic color fundus images and corresponding meta-information from the University Hospitals Leuven (UZL) in Belgium. They are the only one that combine accurate Vertical Cup Disk Ration estimation and glaucoma detection in one study using end-to-end deep learning and that study the exact location of glaucoma symptoms on the eye fundus image using different cropping methods. The work was divided into three main task:
1. **Cropping of images using different regions and different ratios**: The images were cropped in two different ways, to prove the existence of glaucoma signs on both the optic nerve head region (ONH) and the periphery region of the fundus. For one cropping method, only the pixels in the middle of the image are kept (ONH region), and for the other cropping method only the periphery of the eye fundus is kept. For both of these croppings, different ratios are used to try to locate the region of glaucoma symptoms in the most accurate way possible.

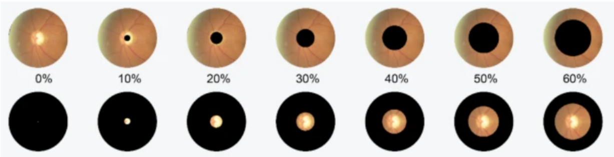


Figure 1. Cropping of the periphery region (first line) and cropping of the ONH (second line) with different ratios

2. **Vertical Cup to Disk Ratio (VCDR) Regression**: The baseline used is to predict the mean VCDR of the dataset (which is 0.67) and the mean absolute error (MAE) for this baseline is 0.19. A ResNet CNN was then trained on the cropped images of the ONH (first cropping method). A MAE of 0.079 is obtained, corresponding to an error reduction of 58% compared to baseline.

2.2. Limitations of the paper and possible improvements

However, this paper fails to mention some key challenges regarding glaucoma detection. First, the bias in the training dataset. The UZL images used for training mainly represent old people (age avergae is 60 years old) that have a suspicion of glaucoma already. The fact that approximately half of the patients in this dataset present glaucoma signs on their eye-fundus is not representative of the real-world repartition of glaucoma, that can start as early as 20 years old and only represent 0.5% of the population. In our work, we used a more imbalanced dataset containing all ages and genders to train the model on a dataset that more accurately represents the real-world repartition of glaucoma. Other improvements that we implemented are the use of other neural networks and higher performance models such as transformers. We combined the VCDR regression task and the classification in a single network and used different metrics to assess the performance of the different models.

3. Datasets and Features

Our dataset is composed of 650 publicly available eye-fundus images and of a small table of associated features. The features are the filename, the eye (left or right) the Cup-Disk Ration (CDR) and the label of the diagnosis: glaucoma (1), no glaucoma (0). Among the 650 pictures, 25% were diagnosed with the glaucoma and the average CDR is 0.57. Below is an example of one of the pictures in the dataset.

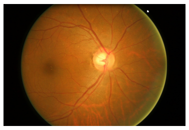


Figure 2. Eye-fundus of an eye diagnosed with glaucoma

The pictures were already preprocessed to have optimal contrast, and the pixels outside the eye were set to black for optimal performance. For our DL models our inputs are the pictures resized with size (224, 224), combined with the CDR feature given in the table. The main challenge regarding this dataset is the small amount of data. Although the final accuracy achieved is not enough for a medical diagnosis because of the small size of the dataset, we loved exploring different techniques to limit this problem.

4. Methods

4.1. Prepocessing and data exploration

For the images, the main preprocessing steps were resizing the images, and associating each image with its CDR in the table. We split the data into glaucoma positive and glaucoma negative and associated them with the features in the table. For the table, we one-hot encoded the categorical features, filled the NaN values with the mean of the feature, and performed some data exploration to get familiar with the dataset.

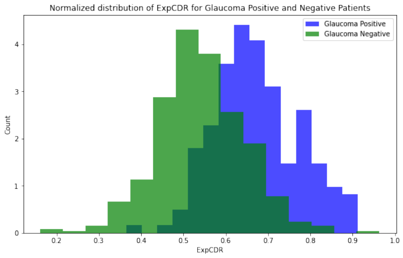


Figure 3. Repartition of the cup-disk ratio for positive and negative patients

4.2. Improved models

We selected 7 pre-trained CNN models at the first stage for pilot test and comparison: ResNet50 [4], VGG16 [6], Xception [1], ResNet101 [4], Inception [7], MobileNet [5], and EfficientNet-B7 [8]. Based on the special features of our task that we really care about decreasing false negative errors as much as possible. Because for our diseases detection task, the ”miss” caused much more serious problem than ”false alarm”. Our best score were obtained for Inception [7] and MobileNet [5]. MobileNets utilize a simplified structure incorporating depthwise separable convolutions, enabling the construction of lightweight deep neural networks. In this model, two uncomplicated global hyperparameters are used in order to effectively balance latency and accuracy.

4.3. Dealing with small amount of data

We observed clear overfitting on our initial results. To improve our scores, we decided to explore different methods to deal with small datasets. This is a common problem in Deep Learning, as it is often challenging to gather a good amount of labeled data for a specific problem, especially in medical applications where the data is confidential. This part of the project was especially interesting. The method we used was mentioned in class, and consists in reducing the size of the test set to put more data in the training set. The initial split was 520 images in the training test and 130 images in the test set. We trained our models for the following train sizes: 520, 540, 560, 580, 600 and 620. The results are reported in the next section.

4.4. Vision Transformer model

Following the explosion of transformer models for natural language processing (NLP) tasks, there have been many papers applying them to vision tasks. One of the prominent models is Visision Transformer model (ViT) [2]. Pixels in images can be treated as sequences of data that can be fed into transformers. However, the computation requirement for image tasks can scale up quickly. ViT solves this issue by splitting each image into sub-image patches, embedding each patch with a linear projection and then the sequence of these embedded patches is fed into the transformer model.

5. Experimentation and results

We conduct our experiments on a open-sourced glaucoma detection dataset [3] that has been processed the way discussed above with multiple pre-trained CNN models and different training methods. After training, we evaluated the trained model on a test set and report some key metric values.

5.1. CNN models

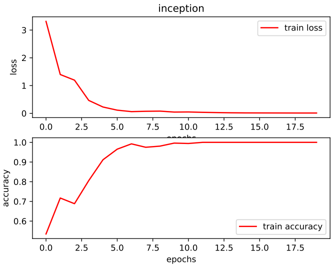
5.1.1 Training methods

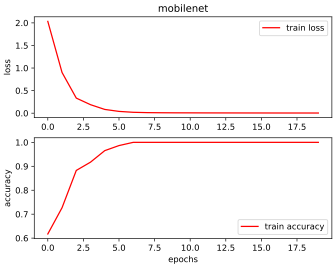
For each CNN model, we tried two ways of training: only using glaucoma images v.s. using combined inputs including glaucoma images and CDR feature data. For the first case, we used sigmoid as activation function on the output layer. While for the second case, to better fit the combined input, we designed a three-layers output module with size of 150*50*1, using relu, relu, and sigmoid as activation functions, separately. For both cases, we used binary-crossentropy as loss function, training for 20 epochs, recording the accuracy and loss during training.

5.1.2 Training results

Below are the training results of our experiments. Each subplot demonstrates the training on one CNN model using the certain method. The upper part is the loss v.s. epoch while the lower part shows the acuracy v.s. epoch.

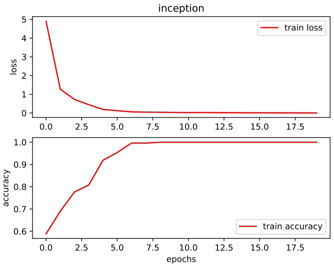
Training curves and model evaluation table for image inputs

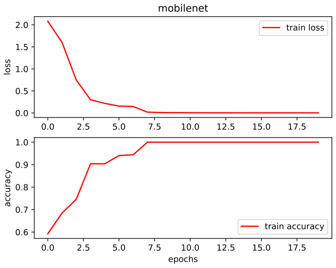




Model	Precision	Recall	F1-score
Inception	0.56	0.57	0.57
MobileNet	0.60	0.61	0.60
ViT	**0.73**	**0.75**	**0.66**

Training curves and model evaluation table for image inputs and feature table inputs





Model	Precision	Recall	F1-score
Inception	0.59	0.60	0.59
MobileNet	0.62	0.64	0.63

5.1.3 Results and discussion

After training, we test our trained models on the test set we discussed above. The test results report the precision, recall value, and F1-score of each model with both training methods. Because we have an imbalanced dataset where there are many more glaucoma-negative images than glaucoma-positive images, we added weights to the training process. Evaluation tables above demonstrate the test results of trained models.

- It is obvious that most models achieved better detection performance with combined inputs including glaucoma images and CDR features compared with only trained with glaucoma images. Therefore, we can conclude that CDR features are significantly useful in glaucoma detection task.
- Due to the limitation on the size of dataset, there still remains a fairly large room to improve for the glaucoma detection method we proposed. Both Inception and MobileNet suffered critical overfit on the training data. They achieved high and stable accuracy (both over 0.95) with a very low loss (0.0326 for Inception and 0.0046 for MobileNet) at the end of training but still did not perform ideally on test set.

5.2. Different splitting on dataset

After getting the baseline results trained on 520-images training set, to make a better use of our size-limited dataset, we re-split the dataset as Part. 4 discussed. For the training set of size 520, 540, 560, 580, 600, and 620, we trained them with CNN models Inception [7] and MobileNets [5] then got the following classification results.

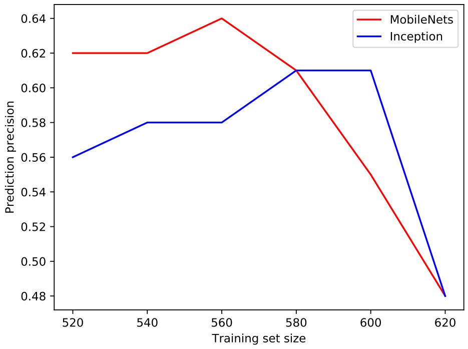


Figure 6. Test results of different splitting on dataset

According to our test results, both precision v.s. training set size curves demonstrated trends that precision first increased then dropped drastically with the increase of training set size. Precision first increased because that there are more data being included in the training set which ease the over-fit issues and lead to a better performance for the model. However, at last, the precision dropped drastically for that the test set at that time was too small and a few wrong predictions can harm the precision seriously. Therefore, there is a trade-off relationship between training set and test set size when splitting the dataset especially for tasks of small available dataset like our project. For our special case, the optimal splitting should be 560 images for training set and 90 images for test set.

5.3. Vision Transformer

In our experiments, we applied some basic data augmentation including resizing (to 72x72-pixel images), flipping, rotating and zooming images before separating each image into 6x6 patches. These patches are then processed and used as inputs to a transformer decoder according to the ViT model architecture.

After training the images with ViT, the predicted results are evaluated using the same metrics used for the CNN models (precision, recall and F1-score) and are reported in section **5.1.2**. The results of this transformer model trained only on images have already outscored the CNN models trained on both image and data features, proving that transformer is a powerful architecture and it can be applied to many domains.

6. Conclusion and Future Work

In this paper, we proposed a novel glaucoma detection framework with a relatively small sized dataset. By leveraging different methods includes fine-tuning the model, different dataset splitting, as well as data augmentation, plus utilizing the pivotal feature of Cup-Disk Ration (CDR), we throughout exploited the abilities of mainstream off-the-shelf CNN models

and achieved good glaucoma detection results. Besides, by pilot testing the detection task with Transformer framework, the results proved that Transformer has great potential in glaucoma detection task compared with current CNN models.
For future work, we plan to make better use of the ability of Transformer by exploring data features with images as the inputs for Transformer. Meanwhile, we also plan to better utilize our limited dataset by using data augmentation.

References

[1] François Chollet. Xception: Deep learning with depthwise separable convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1251–1258, 2017. 3

[2] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale, 2021. 3

[3] Sauman Das Edward Zhang. Glaucoma detection. https://www.kaggle.com/datasets/sshikamaru/glaucoma-detection/data. Kaggle. 3

[4] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 3

[5] Andrew G Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*, 2017. 3, 5

[6] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 3

[7] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826, 2016. 3, 5

[8] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, pages 6105–6114. PMLR, 2019. 3

[9] C. Lawrence Zitnick and Piotr Dollár. Edge boxes: Locating object proposals from edges. In David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars, editors, *Computer Vision – ECCV 2014*, pages 391–405, Cham, 2014. Springer International Publishing. 2