FRONTIERS IN DIABETIC RETINOPATHY SCREENING: DEVELOPMENT OF A RETINAL IMAGE PROCESSING PIPELINE

Anonymous authors
Paper under double-blind review

ABSTRACT

Breakthroughs in the use of AI for Diabetic Retinopathy(DR) diagnosis, have made headway in making DR treatment more accessible but image and camera variability significantly affects the reproducibility of these machine learning algorithms. In an effort to improve the reproducibility of ML algorithms, we attempt to build a retinal image processing pipeline to quantify image quality taking into account luminance and blurriness, discarding poor quality images based on these metrics. Our pipeline further standardizes all images by cropping and resizing. To test the impact of our processing pipeline, we document the results of a 5-fold cross validation with and without the pipeline. Running images through the pipeline shows an increase in AUC performance attributable to an increase in image quality.

1 INTRODUCTION

The success of deep learning analysis for fundus images represents a great milestone both for AI and global health. It has created new interest in the development of low-cost fundus cameras, and mobile phone technologies to capture fundus images in low resource constrained areas. Although the concept of portable AI driven diagnostic platforms is very attractive, algorithms trained on images from specific cameras are not generalizable. The reproducibility of these algorithms is limited due to the varying qualities of images. We attempt to quantify two image qualities, blurriness and luminance, and determine thresholds for good and poor quality images based on our algorithims. All poor quality images are discarded. To standardize images, we crop images around the fundus area to reduce image noise and resize images to an optimal size derived from experiments to find an optimal size against processing time and accuracy trade offs.

2 DATASET

Kaggle, cognizant of the need for comprehensive, automated screening methods of DR screening, hosted a competition sponsored by the California Healthcare Foundation. It provided a dataset of fundus images taken in a wide variety of imaging conditions to mimic a real world dataset. Images in the dataset were graded by clinicians on a 0 - 4 scale, 0: No DR, 1: Mild DR , 2, Moderate DR, 3: Severe DR, 4: Proliferative DR.

3 METHODOLOGY

We ran our experiments on a simple CNN architecture consisting of an Inception V3 layer (loaded with pre-trained imagenet weights), a batch normalization layer and a global average pooling layer. For this study, we define Referable Diabetic Retinopathy (RDR) as anything more severe than mild diabetic retinopathy, with or without macula edema. The learning task was a binary classification task to separate no RDR from RDR. After undersampling the majority classes, we run all images without preprocessing through the model and attain an avg. AUC of 0.8676 which served as our benchmark performance.

We quantify two qualities: brightness and blurriness, both of which affect the visibility of microvascular structures essential to the diagnosis of RDR. To compute the brightness metric of an image, we read the image on the LAB color space and afterwards calculate the average value of the L channel. To analyze image blurriness, we define a Fast Fourier Transform (FFT) blur metric based on the concentration of low and high frequencies in an image. We hypothesized that the sharper an image is the larger the amount of high frequency content and vice versa. To test this hypothesis, we artificially applied gaussian blurs at different thresholds to a sample image to quantify how high and low frequency content vary with different gaussian blurs. We indeed found that the higher the pixel blur, the higher the drop in high frequency content.

For both metrics, we study the distribution of metrics for the kaggle dataset and define thresholds for poor quality images based on the visibility of micro-vascular structures. Using both metrics we discard all images that don't fall within the high quality range. All images that pass this first test are afterwards cropped to reduce noise, and resized to an optimal image size (512x512 pixels).

4 RESULTS

As described, we develop a retinal image processing pipeline that filters poor quality images based on luminance and blurriness metrics. It afterwards crops the image and resizes images to a standard size before the images are run through our benchmark CNN model.

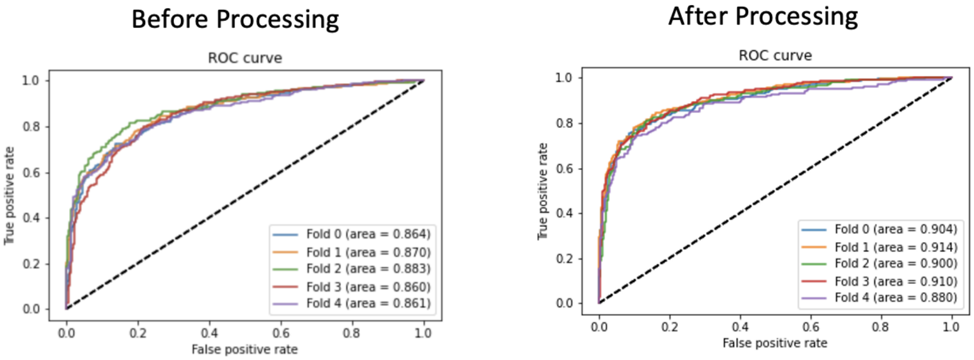

Figure 1: 5-fold cross validation results before and after running images through the retinal image processing pipeline

As shown in figure 1, we noted a net positive increase in AUC, after running images through the retinal image processing pipeline. We can attribute this performance improvement to an increase in the quality of images since the pipeline filters out all poor quality images and standardizes all images for the CNN.

5 CONCLUSION

While this work makes headway in the quest towards making AI solutions for the diagnosis more reproducible, more work still needs to be done in image analysis to qualify and quantify the factors affecting model generalization. For instance, our metrics show a performance increase, but still filter out many good images which are a result of characteristic image qualities injecting false positives. This work goes to show the possibilities in image analysis to improve generalization in the diagnosis of DR.