Electronic Letters on Computer Vision and Image Analysis 0(0):1-7, 2000

XAI for Retinopathy

Shravan Chandra* and Gowri Srinivasa*

* *PES Center for Pattern Recognition, PES University, Bengaluru, India*
Received 1 January 2000; revised 1 January 2000; accepted 1 January 2000

Abstract

The recently emerged field of explainable artificial intelligence (XAI) attempts to shed lights on 'black box' Machine Learning (ML) models in understandable terms for human. As several explanation methods are developed alongside different applications for a black box model, the need for expert-level evaluation in inspecting their effectiveness becomes inevitable. This is significantly important for sensitive domains such as medical applications where evaluation of experts is essential to better understand how accurate the results of complex ML are and debug the models if necessary. The aim of this study is to experimentally show how the expert-level evaluation of XAI methods in a medical application can be utilized and aligned with the actual explanations generated by the clinician. To this end, we collect annotations from expert subjects equipped with an eye-tracker while they classify medical images and devise an approach for comparing the results with those obtained from XAI methods. We demonstrate the effectiveness of our approach in several experiments.

Key Words: Explainable AI, XAI, Retinopathy, Attention Models, CAM

1 Introduction

Roughly four hundred and twenty million people worldwide have been diagnosed with diabetes mellitus. Of those with diabetes, nearly one-third are foreseen to be diagnosed with diabetic retinopathy (DR), a permanent eye disease that can advance to unchangeable vision loss. Early detection, which is crucial for a good diagnosis, relies on experienced professionals and is both labor and time-intensive, which poses a difficulty in regions that traditionally lack access to skilled clinical equipment. Furthermore, the manual nature of DR screening systems increases extensive divergence among readers. Lastly, given an increment in the ubiquity of both diabetes and associated retinal complexities throughout the world, old-fashioned methods of diagnosis may be inadequate to keep pace with the requirement for screening services.

Automated techniques for diabetic retinopathy diagnoses are imperative to answer these problems. With the arrival of deep learning architectures, CNN-based networks have displayed abilities that have transcended conventional feature engineering in various image recognition tasks, which can be ascribed to the competence of such networks to automate the process of feature extraction. But the black-box nature of deep learning architectures has withheld professionals from relying on the predictions made by these models. Professionals need to know the reason for the grading made by the models to have a deeper understanding of the retina and the patient's condition.

Correspondence to: <E-mail>

Recommended for acceptance by <name>
ELCVIA ISSN:1577-5097
Published by Computer Vision Center / Universitat Autònoma de Barcelona, Barcelona, Spain

2 Data

We use two datasets in this study: the EyePACS dataset and the Asia Pacific Tele-Ophthalmology Society (APTOS) 2019 dataset. The EyePACS dataset includes 35,126 images of various dimensions, with sufficient variability for each class, and was obtained from Kaggle: EyePACS. The APTOS contains 5590 annotated images of dimension 2136x3216. This dataset is representative of the Indian population, captured and graded by retinal specialists at the Aravind Eye Hospital, made available on Kaggle: APTOS. Since both datasets follow similar imaging protocols and convention for grading the severity of DR, we can combine the data sets. Table 1 presents a summary of the numerical categories, the label they correspond to and the number of images available in each class. We note that the 'No DR' class (or the class of images that is considered 'healthy' for the purpose of this study) is overwhelmingly over-represented compared to the other categories. Among the four categories that represent different grades of DR, we note the largest number of images for Moderate DR in both data sets.

Class	Label	EyePACS	APTOS
0	No DR	25810	1805
1	Mild DR	2443	370
2	Moderate DR	5292	999
3	Severe DR	873	193
4	Proliferate DR	708	295

Table 1: Dataset overview: Number of images in each class

3 Methodology

Deeper layers in CNNs capture higher-level visual information and retain spatial information that is lost in fully connected layers, so it is common to expect the last convolutional layer to offer the best compromise between high-level semantics and detailed spatial information, and the neurons in these layers look for class-specific semantic information in the input image.

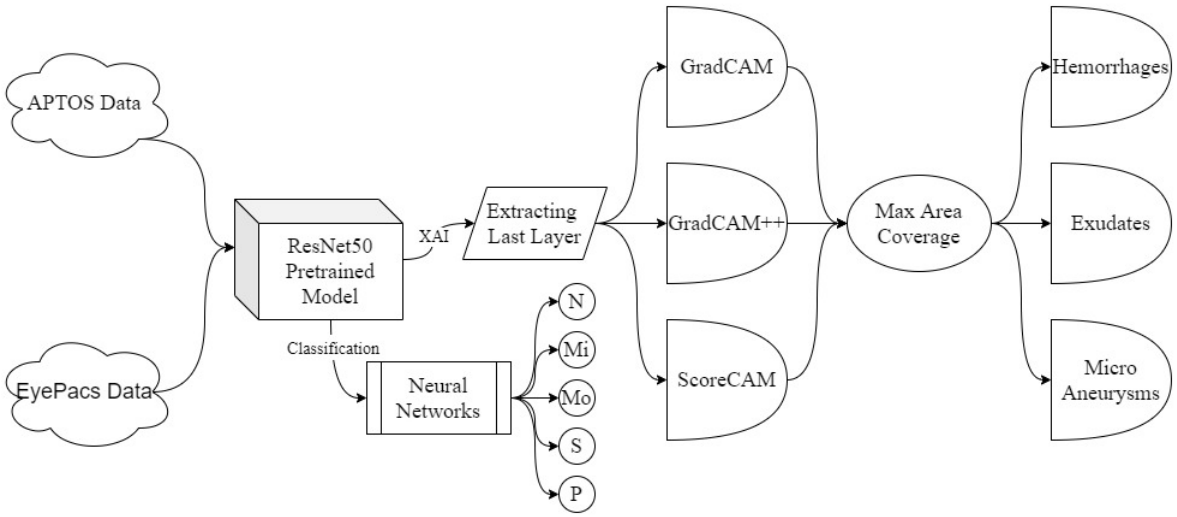

Figure 1: A schematic diagram of the solution approach

3.1 CAM

Class Activation Map is a technique used to describe the working of CNN by projecting back the weights of the output layer on the convolutional feature maps obtained from the last convolution layer. One of the shortcomings is that it requires feature maps to precede the softmax layers, so it applies to CNN architectures that perform global average pooling over convolutional maps before prediction. Hence the preferred network for training the model was ResNet.

3.1.1 Residual Networks

Residual Networks, commonly known as, ResNets was developed by *He and Zhang*. [**?**] The essence of ResNet is the introduction of "identity shortcut connection" that skips one or more layers. This circumvents degradation of performance with the stacking of more layers. By simply stacking identity mappings, the resulting architecture would perform equivalent to one which includes layers that are not doing anything. This symbolizes that the deeper model should not present a training error higher than its shallower equivalents. We used a version of ResNet called ResNet50 (with 50 indicating the number of layers of the residual network) and provides an embedding size of 2048.

3.2 Grad-CAM & Grad-CAM++

Grad-Cam uses the gradient information flowing into the last convolutional layer to interpret each neuron for a judgment. To obtain the class discriminative localization map, we compute the gradient of the score for the class and its softmax with respect to feature maps of the convolutional layer. These gradients flowing back are global average-pooled to obtain the neuron importance weights for the target class. We later perform a weighted combination of activation maps and follow it by ReLU, which results in a coarse heatmap of the same size as that of the convolutional feature maps. We apply ReLU as we are engrossed in the features with a definite impact on the class of interest.

Grad-CAM++ uses a weighted mixture of the positive part derivatives of the last convolutional layer feature maps for a specific class score as weights to create a visual explanation for the class label under consideration.

3.3 Score-CAM

The Score-Cam doesn't use gradients because the propagated gradients are unstable and generate random noise in gradient-based saliency maps. In contrast to GradCam and GradCam++, which use the gradient information flowing through the last layer of the CNN to represent the importance of each activation map, Score-Cam uses the weights of the score obtained for a specific target class. Hence, Score-Cam can get rid of the dependence on the gradient and works as a more general framework as it only requires access to the class activation map and output score of the model.

3.4 U-Net

The UNET was developed by Olaf Ronneberger et al. for Bio Medical Image Segmentation [2]. The architecture contains two paths. First path is the contraction path (also called as the encoder) which is used to capture the context in the image. The encoder is just a traditional stack of convolutional and max pooling layers. The second path is the symmetric expanding path (also called as the decoder) which is used to enable precise localization using transposed convolutions. Thus it is an end-to-end fully convolutional network (FCN), i.e. it only contains Convolutional layers and does not contain any Dense layer because of which it can accept image of any size.

4 Experiments AND Results

Once the saliency maps were extracted using different CAM methods, the area of each map was calculated and the map with maximum area was chosen to extract the features from. This was done to have the best-case scenario, as in once the features are extracted, we wanted to make sure more area is covered by the model for professional's reference, than less area.

Once the map was selected, the Region of Interest (RoI) was extracted from the original retina fundus, and passed through a UNet to extracted prominent features like Exudates, Hemorrhages and MicroAneurysms. This was compared with the same features extracted from the original retina fundus with no preprocessing.

As per our observation, the maximum area was either from Grad-CAM++ or from Score-CAM, with over 90+% overlap between the both.

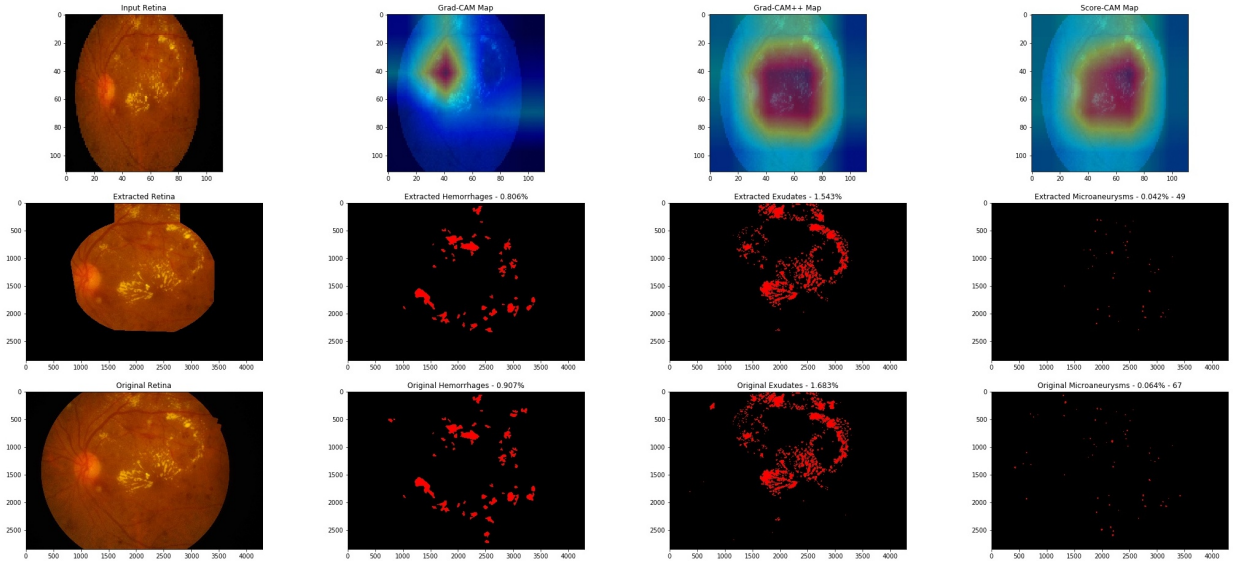

Figure 2: Sample Image of the Output.

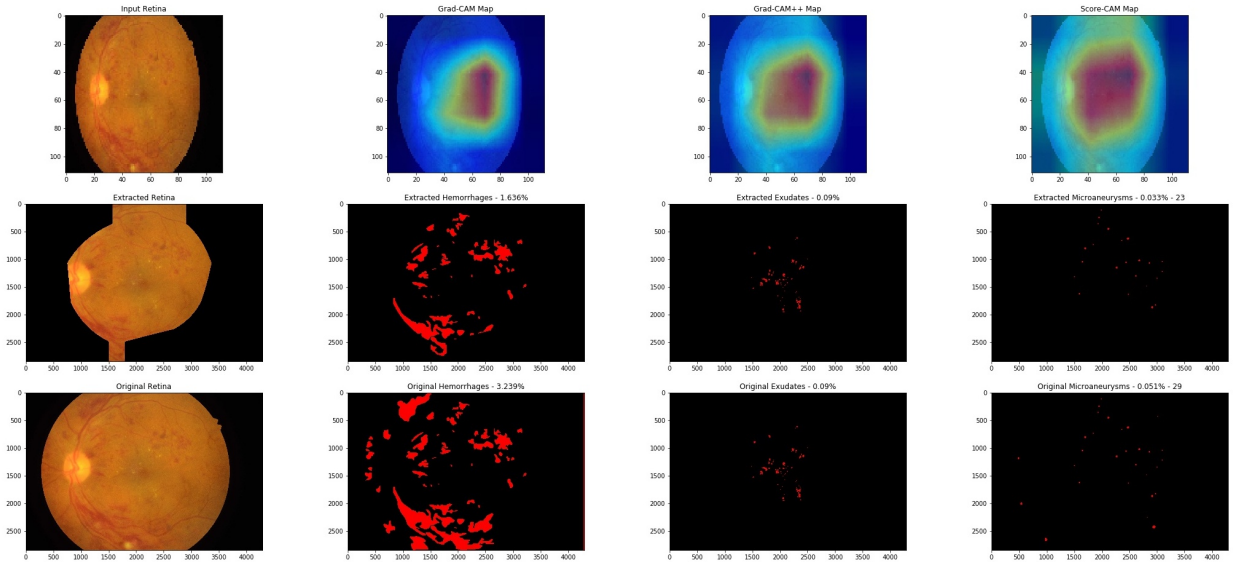

Figure 3: Sample Image of the Output.

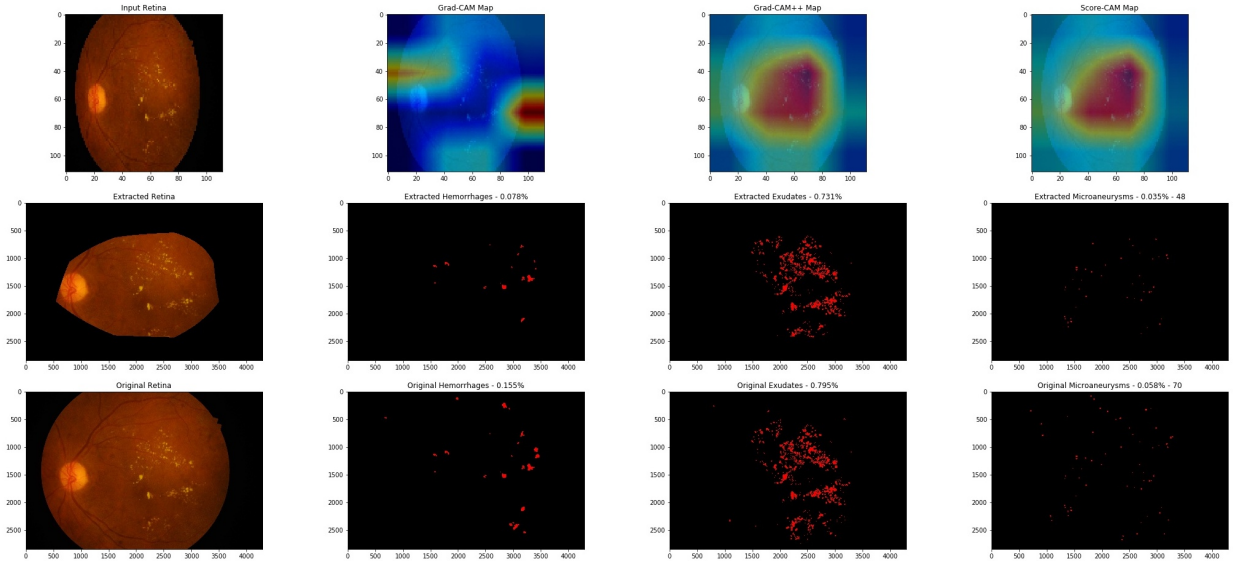

Figure 4: Sample Image of the Output.

Feature	Average Coverage (%)
Hemorrhages	79.81
Exudates	75.23
MicroAneurysms	74.40

Table 2: Average Features Coverage by the Saliency Maps compared to Original Features

5 Reproducible Research

In the spirit of reproducible research, we have used publicly available data (EyePACS and APTOS) and made all the code used in this study available here.

6 Conclusion

We have presented a framework for the automated detection and grading of the severity of diabetic retinopathy, as well as in interpretability of the decision with support. The proposed method leverages the power of CNN-based architectures pretrained with ImageNet and trained using EyePACS data, to achieve high accuracy and quadratic kappa scores on hold out data and combines that with various CAM models to create saliency maps for better understanding by the professionals. We also leverage U-Net to extract prominent features like hemorrhages, exudates, and micro-aneurysms. We have ensured there is no overfitting in the training phase through tracking the validation loss at the end of each epoch and computing the standard deviation across five-fold cross-validation. We have worked on publicly available data and made the code used in this study available to facilitate reproducing the results and, hopefully, improving upon these, towards making effective automated pre-screening for DR affordable and accessible.

References

[1] Y. Hatanaka, T. Nakagawa, Y. Hayashi, Y. Mizukusa, A. Fujita, M. Kakogawa, K. Kawase, T. Hara, H. Fujita (2007), ”CAD scheme to detect hemorrhages and exudates in ocular fundus images”, *(2007 SPIE Symposium, vol. 65142M, 2007)*

[2] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton (2012), "Imagenet classification with deep convolutional neural networks, *(Advances in Neural Information Processing Systems 25, F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, Eds., pp. 1097–1105. Curran Associates, Inc., 2012.)*

[3] L. W. Yun, U. R. Acharya, Y. V. Venkatesh, C. Chee, L.C. Min, and E.Y.K. Ng (2007), "Identification of different stages of diabetic retinopathy using retinal optical images", *(Information Sciences, vol. 178, pp. 106-121, 2008)*

[4] S. B. Hathwar and G. Srinivasa (2019), "Automated Grading of Diabetic Retinopathy in Retinal Fundus Images using Deep Learning", *(2019 IEEE International Conference on Signal and Image Processing Applications (ICSIPA), Kuala Lumpur, Malaysia, 2019)*