Feature Visualisation of Classification of Diabetic Retinopathy Using a Convolutional Neural Network

Harry Pratt[1] , **Frans Coenen**[3*] , **Simon P. Harding**[1,2] , **Deborah M. Broadbent**[2] , **Yalin Zheng**[1,2]

[1]Institute of Ageing and Chronic Disease, University of Liverpool, Liverpool, L7 8BX,
[2]St. Paul's Eye Unit, Royal Liverpool University Hospital, Liverpool, L7 8XP,
[3]Department of Computer Science, University of Liverpool, Liverpool, L69 3BX

sghpratt@liverpool.ac.uk, coenen@liverpool.ac.uk, sharding@liverpool.ac.uk,
dbroadbe@liverpool.ac.uk, yzheng@liverpool.ac.uk

Abstract

Convolutional Neural Networks (CNNs) have been demonstrated to achieve state-of-the-art results on complex computer vision tasks, including medical image diagnosis of Diabetic Retinopathy (DR). CNNs are powerful because they determine relevant image features automatically. However, the current inability to demonstrate what these features are has led to CNNs being considered to be 'black box' methods whose results should not be trusted. This paper presents a method for identifying the learned features of a CNN and applies it in the context of the diagnosis of DR in fundus images using the well-known DenseNet. We train the CNN to diagnose and determine the severity of DR and then successfully extract feature maps from the CNN which identify the regions and features of the images which have led most strongly to the CNN prediction. This feature extraction process has great potential, particularly for encouraging confidence in CNN approaches from users and clinicians, and can aid in the further development of CNN methods. There is also potential for determining previously unidentified features which may contribute to a classification.

1 Introduction

Convolutional Neural Networks (CNNs), a deep learning approach to image classification, can offer extremely fast classification predictions based on learning relevant features. These features are learned within the network structure itself; from labeled images that the network has 'seen'. Recently CNNs have been used to enhance accuracy on a wide range of computer vision tasks [Krizhevsky *et al.*, 2012]. This has extended to the application of automated medical image diagnosis. For example, the classification of Diabetic Retinopathy (DR) severity through the use of colour fundus images [Pratt *et al.*, 2016; Gulshan *et al.*, 2016]. The CNNs presented in these papers have learned features of DR in order to determine the level of DR severity within a fundus image using clinically labeled images.

However, the DR classification predictions presented in these papers do not offer any insight in to the reasoning behind the CNN model predictions. Although the CNN models have learned from ground truths based on a clinical grading framework the methods do not present the features that have been learned by the CNN in order to arrive at the prediction. DR feature extraction from fundus images typically involves manual algorithms [Ravishankar *et al.*, 2009; ManojKumar *et al.*, 2015] which are undertaken before the classification process commences. The extracted features then correspond to a predicted severity of the disease. In the case of CNN models we wish to implement the reverse procedure. Through dissecting the CNN model we wish to determine which features have led to the prediction.

Feature extraction is a vital process in the grading of DR because the manual process used by clinicians are typically feature based processes, for example the process prescribed in[ETDRS Study Group, 1991]. Deep learning in the clinical community is widely perceived to be *black box*. Consequently it is unclear to clinicians whether the feature based framework used in manual grading is the same as the classification framework produced by the CNN. As a result there is a lack of trust in the ability of deep learner.

In [Zhou *et al.*, 2015] Class Activation Maps (CAMs) were presented as a method of determining the regions within a CNN input image which have contributed most towards the classification. In the case of disease classification this offers insight into the areas of the image containing features of the disease under consideration. The severity of DR within a fundus image directly relates to the location of certain features [ETDRS Study Group, 1991]. These features, their location and how they relate to DR classification are presented in Table 1. The idea of saliency maps was presented in [Simonyan *et al.*, 2013]. Saliency maps offer a method of determining the most significant pixels involved in the classification prediction of an image.

This paper aims to open the CNN black-box in order to make CNNs more transparent in the context of feature based prediction of DR. Deep learning classification methods do

*Contact Author

not justify prediction values. This paper presents a novel method of extending CNN black box prediction models so that they become feature based models. Through determining the learned features and their locations we explore how the CNN reached its prediction and how this corresponds to the manual feature based grading.

2 Method

Initially a CNN was trained on fundus images to predict DR severity. Once the model had been trained the model parameters remained immutable throughout the rest of the process. The trained model was then used to produce prediction values, saliency maps and CAMs for unseen test images. Attention maps and other techniques would produce similar results to the class activation maps and saliency maps if applied to the CNN. The two selected methods were used as they compliment each other and highlighted features within the image in different manners. For evaluation, these were compared to the clinical ground truth and the features identified within the images.

2.1 Dataset

The dataset used for training and evaluation was from Kaggle [Kaggle, 2016]. The dataset is a large set of 88,702 high-resolution retina fundus images; 78,076 training, 10,626 testing. A clinician has graded the level of DR using five classes: no DR, mild DR, moderate DR, severe DR and proliferative DR. The images were provided by eyePACS [EyePacs, 2018] from a diabetic screening process. Example images from the dataset are given in Figure 1.

2.2 Convolutional Neural Network Training

The adopted CNN architecture was the well-known DenseNet [Huang *et al.*, 2016] demonstrated in Figure 2. The DenseNet weights were initialised with pre-trained ImageNet weights, a learning rate of 0.0003 was used with Adam backpropagation on a NVIDIA k40 GPU using *Keras* [Keras, 2019] library. Training was undertaken until the categorical cross entropy loss function plateaued on the test data.

2.3 Class Activation Maps

In this section, we define the procedure for producing Class Activation Maps (CAMs). CAMs require global average pooling after the final convolution layer in the CNN. Pooling provides the localisation for the region detection. Applying the trained CNN to test activated weights in the output layer depending on nodes that have beeen activated. These weights can be projected back on to the convolutional feature maps in order to identify regions of importance for a certain class. Hence, to compute the class activation maps of an input image we computed a weighted sum of the feature maps of the last convolutional layer. CAMs are defined as follows:

- Let input image I with coordinates (x, y) be $I(x, y)$
- Let $f_k(x, y)$ be the activation of a node k in the last layer of convolution
- The result of global average pooling is $F_k = \sum_{(x,y)} f_k(x, y)$
- For class c softmax input is $S_c = \sum_k w_k^c F_k$ where w_k^c is the weight for node k
- Softmax output, probability, is given as $P_c = \frac{e^{S_c}}{\sum_{c=0} e^{S_c}}$
- The weighted sum of feature maps, the CAM, is defined as,

$$CAM_c = \sum_k w_k^c f_k(x, y). \quad (1)$$

Therefore, it is clear the CAM for class c, CAM_c, directly relates to the prediction value of the class S_c. The weights w in the definition of CAM_c and S_c remain constant from the trained CNN. This therefore indicates the direct importance of the activation at pixel $f_k(x, y)$ to the prediction within the CAM of an image to class c. Therefore, for our CNN trained for DR severity, CAMs are an effective method for determining the region of pixels relating to disease severity prediction. This process is shown in Figure 2.

2.4 Saliency Maps

The idea of saliency maps is to compute the gradient of the output class with respect to the input image. This tells us how the output category value changes with respect to a small change in the input image pixels. Therefore, like CAMs, in saliency maps the weights remain unchanged. Positive values in the gradient tell us that a change to that pixel will increase the output class value. Hence, the larger the positive gradient, the more reliant on this pixel the image is in the classification process. Visualising all of the gradients, which are the same shape as the input image, produces a saliency map which highlights the salient pixels that contribute the most towards the output class. Saliency maps are described as follows:

- Let the input image be defined as I
- Let $S_c(I)$ be the class score function for the image
- We want to rank each pixel (x, y) based on its influence on S_c
- S_c is a highly non-linear function in a CNN. Hence S_c is approximated with a first-order Taylor expansion in the neighborhood of the pixel
- $S_c(I_{(x,y)}) \approx w^T(x, y) + b$
- Where w is the derivative of S_c with respect to image I at point (x, y)

$$w = \frac{\delta S_c}{\delta I}|_{I_{x,y}}$$

The computation of an image-specific saliency map for a single class is extremely quick, since it only requires a single back-propagation pass. Saliency maps differ from CAMs as they look at how changes in the input image affect the class prediction as opposed to combining feature maps in order to determine the most filtered region of an image.

3 Results

The purpose of the paper is to give an insight in how qualitative features can be derived and presented (quantitative results

Feature Grading	DR Level
No apparent retinopathy	No Retinopathy
• Haemorrhages/Microaneurysms only $<$ 2A • $<$ 6 Cotton Wool Spots in the absence of other features • $<$ 6 Cotton Wool Spots with Haemorrhages/Microaneurysms $<$2A • Single venous loop	Mild
• Haemorrhages/Microaneurysms $\geq$ 2A in 1-3 quadrants • $\geq$ 6 Cotton Wool Spots • 1 quadrant Venous Beading/Looping/Reduplication • Intraretinal microvascular abnormalities $<$ 8A	Moderate
• 4 quadrants Haemorrhages/Microaneurysms $\geq$ 2A • 2-4 quadrants Venous Beading/Looping/Reduplication • 1 quadrant Intraretinal microvascular abnormalities $\geq$ 8A	Severe
• Neovascularisation of disc $<$ 10A alone • Neovascularisation Elsewhere $<$ $\frac{1}{2}$ disc area (DA) alone • Neovascularisation Elsewhere $\geq$ $\frac{1}{2}$ DA and no Preretinal/vitreous haemorrhage	Early Proliferative
• Neovascularisation of disc $\geq$ 1/3 DA (10A) alone • Neovascularisation Elsewhere $\geq$ 1/2 DA and Preretinal/vitreous haemorrhage • Vitreous haemorrhage precluding adequate view of fundus • Traction retinal detachment (TRD)	High-risk Proliferative
Neovascularisation of disc/elsewhere have inactivated	Stable treated
Fibrovascular proliferation disc/elsewhere	Stable treated

Table 1: Clinical diagnosis of DR based on various feature types with different contributions to classification. One feature in each list is required for the equivalent DR grading. 2A, 8A and 10A refer to 'standard photographs' from ETDRS [ETDRS Study Group, 1991].

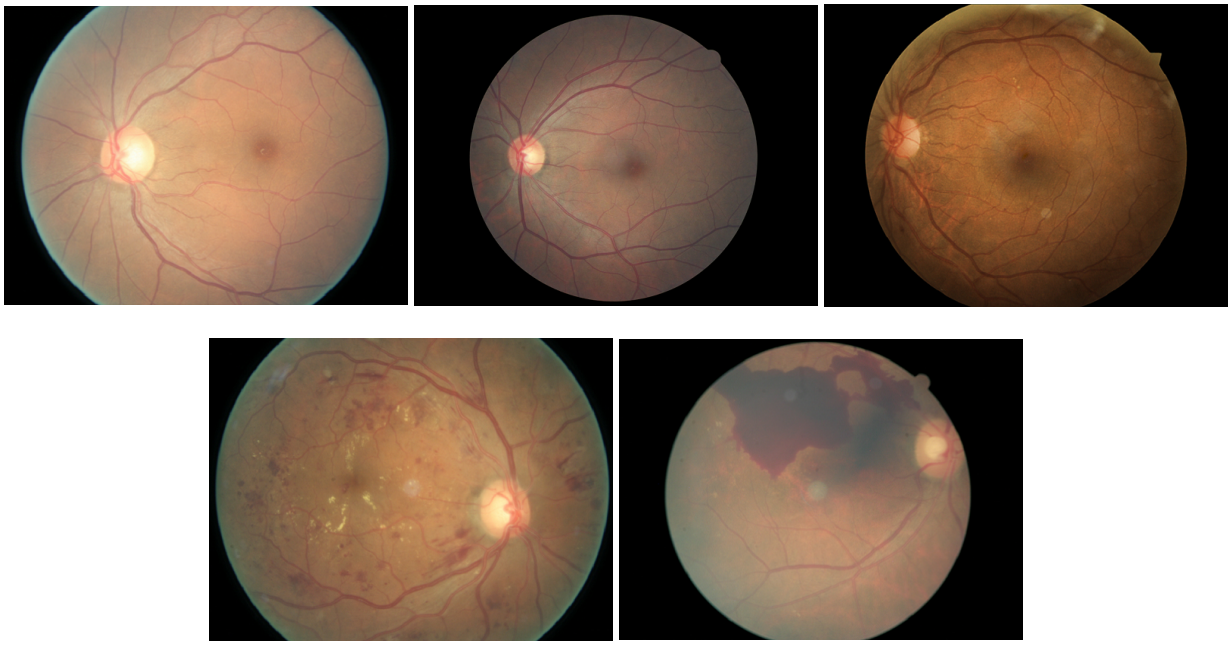

Figure 1: Fundus images from Kaggle dataset; (a) No DR (b) Mild DR (c) Moderate DR (d) Severe DR (e) Proliferative DR. Note: Little obvious difference between (a),(b) and (c), however it is important to distinguish these. CAMs and saliency maps should detect features such as Haemorrhages/Microaneurysms around the vessels.

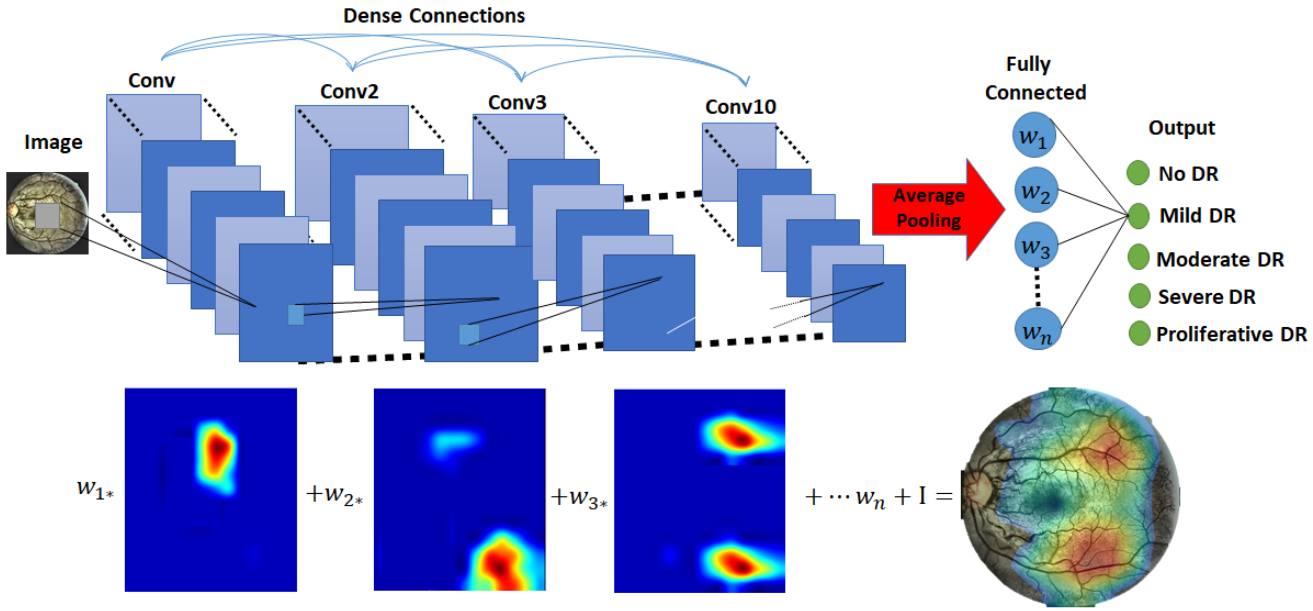

Figure 2: Top: The DenseNet architecture used for training. Bottom: The combination of the trained weights w from the final layer and the feature maps of the last convolution layer to produce the CAM. The feature maps vary depending on the input image.

have been widely discussed in the literature). However, in order to determine the level of quantitative results required in order to achieve this level of qualitative output the qualitative results must be defined. The multi-class DenseNet model achieved 0.81 quadratic weighted kappa on the test data for the multi-class problem.

CAMs from test images, with an example result for each class of DR, are presented in Figure 3. The colour range is from red to green. The closer the region is to red the more that region has contributed towards the prediction. Similarly, in the saliency maps, the lighter the pixel the more the pixel has contributed to the classification of the image.

The CAMs of each class of DR demonstrate the links between the severity ground truth and the input image that the CNN has divulged through the training process. As seen in Figure 3, the regions leading to classification of No, Mild or Moderate DR relate to the main vessel structure and tend to avoid the macula (centre of the retina). Initial signs of disease stem from the vessels in the form of haemorrhages or microaneurysms or abnormal vessels as presented in Table 1. Furthermore, it was also clear from the test image CAMs that the severe and proliferative classifications look more towards the macula. This is shown in the severe and proliferative cases in Figure 3. This corresponds to the clinical classification process as severe disease requires Haemorrhages or Microaneurysms and Venous Beading/Looping/Reduplication throughout the retina. However, the saliency maps for the proliferative case rarely took in to consideration the optic disc region in the classification prediction. This suggests that the CNN model is excluding an important marker for proliferative retinopathy; neovasularisation of the disc.

The saliency maps provide insight in to the features that have been detected through the ground truth and input image training. Figure 3 demonstrates that in the early stages of retinopathy the CNN looks along the vessel structure and looks for deviation normal vessel structures. This is shown through the lightest pixels being the vessels in the saliency map for the no DR and mild DR cases. Haemorrhages and microaneurysms from the early stage of the disease tend to lie around the vessel structure and abnormal vessels are a key distinction between no DR and mild or moderate DR. It is also apparent that the saliency maps in the moderate class have "light" pixels spread around the retina as the CNN looks for features in more than one region of the retina; which is key to moderate classification.

In the saliency maps for the severe and proliferative classifications we can see identification of features relating to clinical diagnosis. In the severe DR saliency map in Figure 4 we can identify the microaneurysms and cotton wool spots. The microaneurysms in different regions of the retina relate directly to the severe DR classification. Similarly, in the proliferative saliency map in Figure 4 the lighter pixels correspond to features that the CNN has identified. The laser spots produced through treatment to the eye remain dark in the saliency map and therefore the CNN is, correctly, not treating these as a feature of disease. An example of this is shown in Figure 4.

4 Discussion

The visualisation techniques presented in this paper demonstrate that CNN models are achieving some success in replicating the clinical process undertaken during diagnosis of fundus images. Similar features are being detected and similar regions are being related to the appropriate classes. However, in order to fully determine if the CNN has learned a similar classification process we would require fundus images

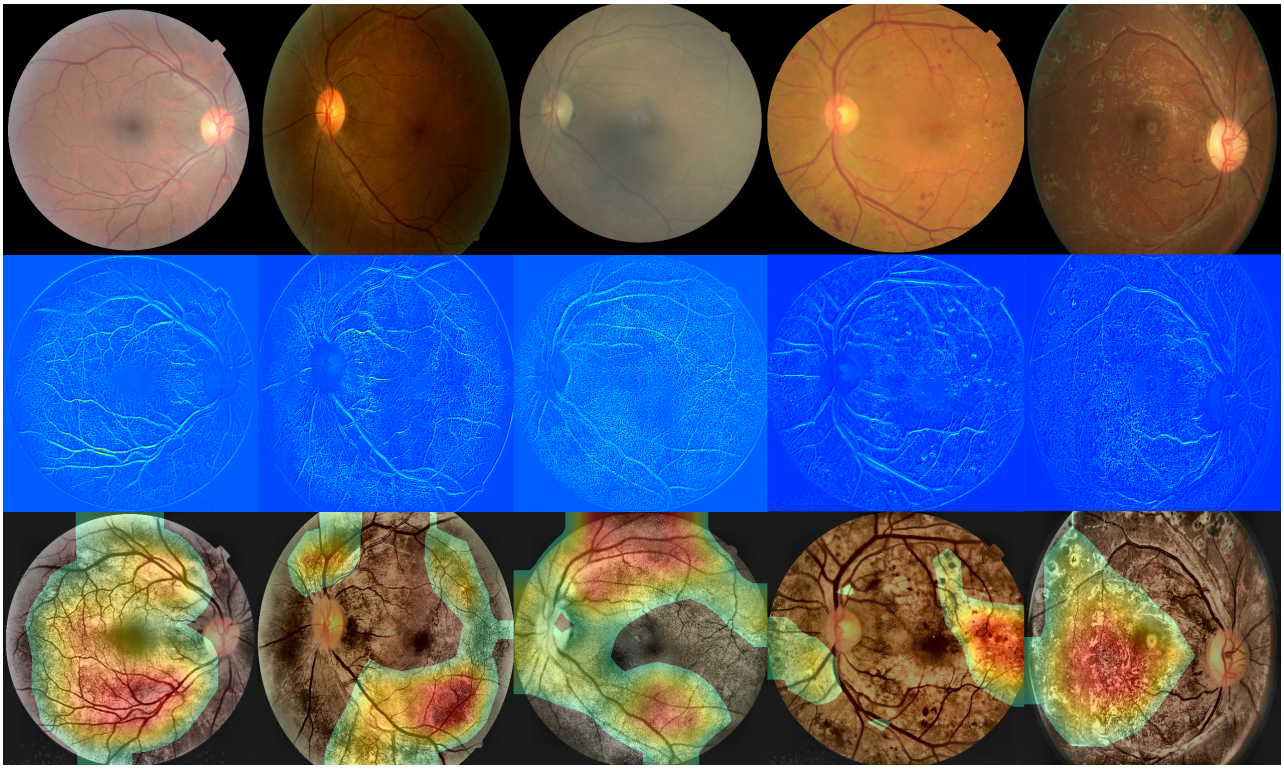

Figure 3: Left to right; No DR, Mild DR, Moderate DR, Severe DR and Proliferative DR. Top to bottom; Original Kaggle Image, Class Activation Map of preprocessed image and Saliency Map of preprocessed image.

annotated with every single feature present in the image and saliency maps annotated to the same criteria. Furthermore, the CNN is only told the severity of the image, not the combination of features involved, so it may therefore be deemed unfair that the CNN is expected to learn the precise mechanism that was used to determine the ground truth. Especially when grader agreement is often variable; complex structures of DR can become subjective when based on such minute features.

The method presented also discovers features of disease severity that are missed in the automated procedure and therefore indicates where the CNN needs to be improved; such as neovasularisation of the disc detection. This could be used to determine a general set of features that CNNs struggle to detect. During training image preprocessing techniques could be used in order to make the missed features more apparent within the image to aid CNN learning.

The methodologies have been validated on images from the Liverpool Diabetic Eye Screening Program (LDESP) in order to test their ability to generalise to other datasets. Figures 5 and 6 demonstrate the class activation maps and saliency maps abilities to generalise to unseen data. Numerous features are identified in multiple fundus images from the same eye, including images that aren't macula centred.

5 Conclusion

In conclusion, we have demonstrated that the correlation between CNN predictions and manual grading of DR can be visualised through the use of Class Activation Maps (CAMs) and saliency maps. These methods provide a useful tool to determine if deep learning classification models relate accurately to clinical diagnosis procedures. The presented methods could also be used in the screening process to reduce the time a clinician spends looking for features within a fundus image. CAMs present a good method for 'flagging' regions of disease, whereas saliency maps present a solution for feature detection.

Acknowledgment

H. Pratt would like to acknowledge everyone in the CRiA imaging team at the Institute of Ageing and Chronic Disease at the University of Liverpool. He would also like to thank the Fight for Sight charity for PhD funding and NVIDIA for providing an NVIDIA k40 GPU.

References

[ETDRS Study Group, 1991] ETDRS Study Group. Grading diabetic retinopathy from stereoscopic color fundus photographs? an extension of the modified airlie house classification: ETDRS report number 10. *Ophthalmology*, 98(5):786–806, 1991.

[EyePacs, 2018] EyePacs. A free platform for retinopathy screening. http://www.eyepacs.com/, 2018. [Online; accessed 30/05/2018].

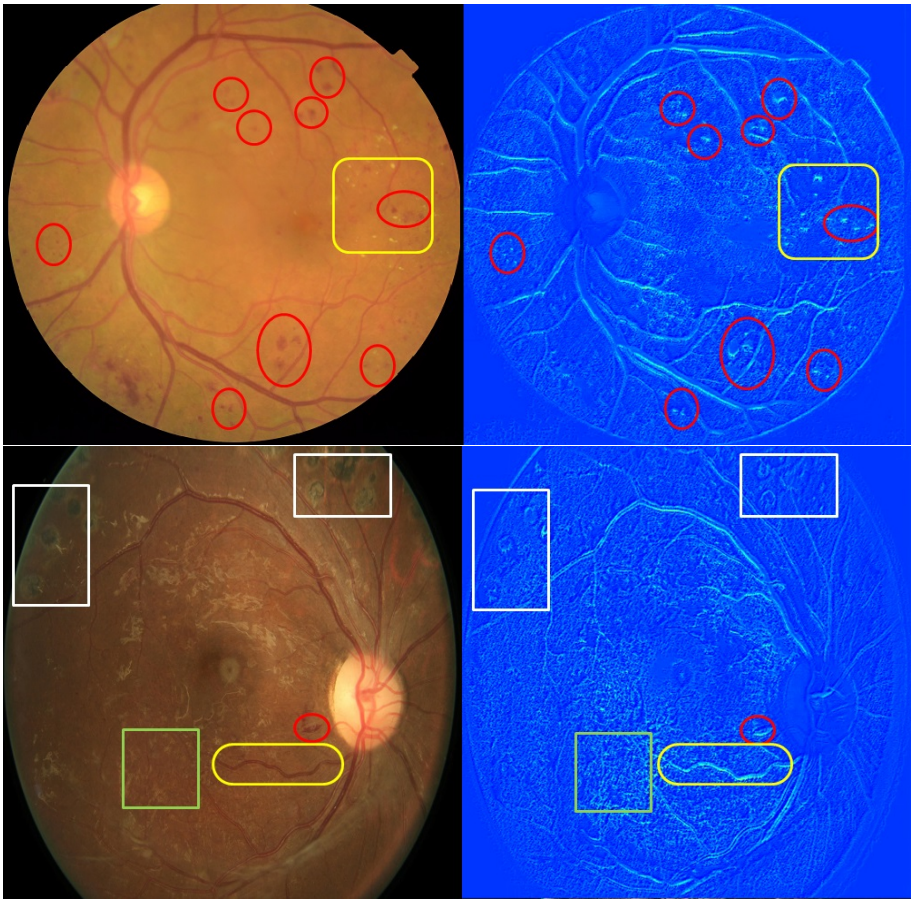

Figure 4: Top to Bottom; severe and proliferative DR. Left to right; original image with expert labelled features and saliency map with matching region overlay. Rectangles denote laser spots, circles denotes haemorrhages or microaneurysms, squares denote neovascularisation elsewhere, curved squares denote cotton wool spots and curved rectangles denotes venous reduplication.

[Gulshan *et al.*, 2016] V Gulshan, L Peng, M Coram, and et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *JAMA*, 316(22):2402–2410, 2016.

[Huang *et al.*, 2016] Gao Huang, Zhuang Liu, and Kilian Q. Weinberger. Densely connected convolutional networks. *CoRR*, abs/1608.06993, 2016.

[Kaggle, 2016] Kaggle. Kaggle: Platform for predictive modelling and analytics competitions, 2016.

[Keras, 2019] Keras. *Keras: Deep Learning library for Theano and TensorFlow*, 2019.

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

[ManojKumar *et al.*, 2015] S B ManojKumar, R Manjunath, and H. S. Sheshadri. Feature extraction from the fundus images for the diagnosis of diabetic retinopathy. In *2015 International Conference on Emerging Research in Electronics, Computer Science and Technology (ICERECT)*, pages 240–245, Dec 2015.

[Pratt *et al.*, 2016] Harry Pratt, Frans Coenen, Deborah M. Broadbent, Simon P. Harding, and Yalin Zheng. Convolutional neural networks for diabetic retinopathy. *Procedia Computer Science*, 90:200 – 205, 2016. 20th Conference on Medical Image Understanding and Analysis (MIUA 2016).

[Ravishankar *et al.*, 2009] S. Ravishankar, A. Jain, and A. Mittal. Automated feature extraction for early detection of diabetic retinopathy in fundus images. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 210–217, June 2009.

[Simonyan *et al.*, 2013] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. *CoRR*, abs/1312.6034, 2013.

[Zhou *et al.*, 2015] Bolei Zhou, Aditya Khosla, Àgata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. *CoRR*, abs/1512.04150, 2015.

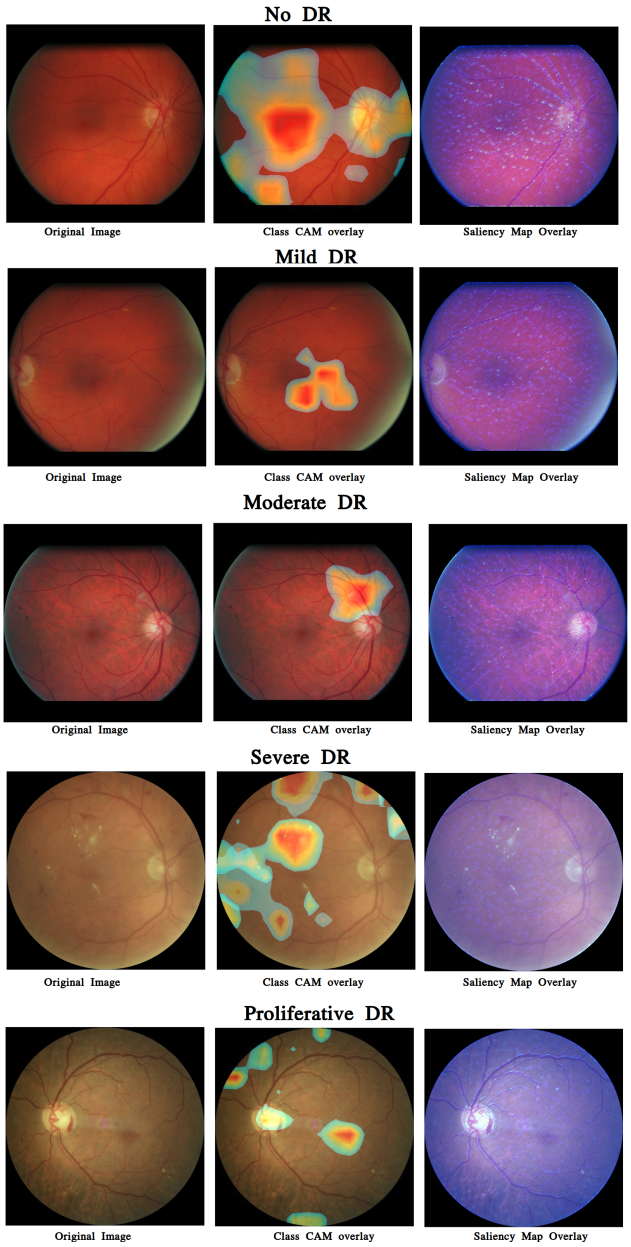

Figure 5: (Left) Fundus images from the Liverpool Diabetic Eye Screening Program (LDESP). Middle) Saliency map from the trained DenseNet multi-class DR model overlayed on the original fundus image. Right) CAMs from the trained DenseNet multi-class DR model overlayed on the original image.

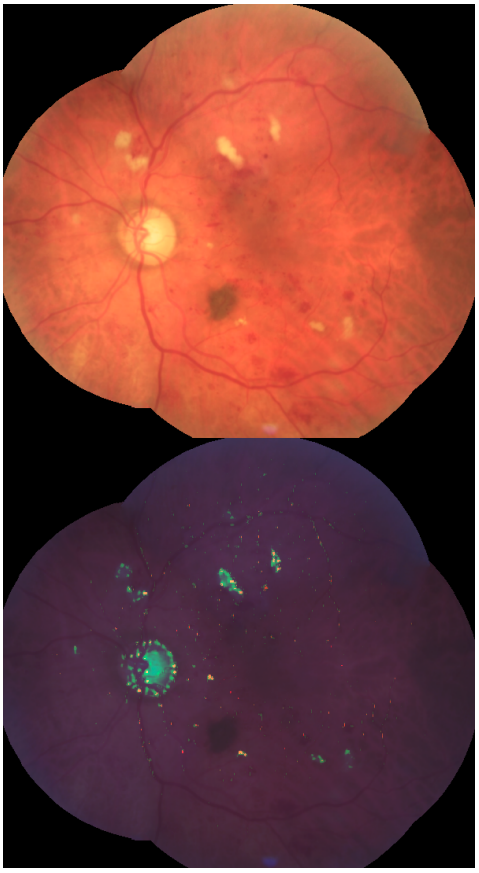

Figure 6: (Top) 4 montaged fundus images. (Bottom) Overlayed saliency map from the trained DenseNet multi-class DR model overlayed on the original fundus image.