Learning of subtle features in retinal images

Carson Lam
Stanford University
Department of Biomedical Informatics
carsonl@stanford.edu

Martin Seneviratne
Stanford University
Department of Biomedical Informatics
martsen@stanford.edu

Abstract

Diabetic retinopathy (DR) is one of the leading causes of irreversible blindness worldwide. Although effective treatments are available if detected early, the number of trained ophthalmologists able to diagnose retinal scans is far outweighed by the global burden of disease. This study applies convolutional neural networks (CNNs) to identify features of early-stage diabetic retinopathy. We show that, when trained on the entire retinal scan, CNNs have limited sensitivity for subtle pathological features such as microaneurysms. However, if CNNs are trained on clinically-relevant portions of the image, better classification results are achievable. We apply these patch neural networks to generate heatmaps across the original scan, representing the probability of pathology within each region. Finally, we show that machine-learning methods can be applied to these heatmaps to achieve binary classification accuracy of up to 0.77 using a random forest classifier. This study demonstrates the potential to use regionally-trained CNNs to generate probability maps and also output predictions in retinal scans of subtle diabetic retinopathies, which may form a basis for improved computer assisted diagnostic tools.

1. Objectives

Diabetic retinopathy (DR) is a disease of the retina which affects one in three diabetic patients in the United States, and can progress to irreversible vision loss. Manual classification of DR is highly time consuming, involving localization and grading of subtle pathological features by trained ophthalmologists. Computer-automated diagnostics for screening retinal scans would significantly reduce costs and decrease inter-observer variability, potentially allowing for more widespread screening programs and earlier detection of subtle diabetic retinopathies.

Representative retinal images of different stages of disease are seen in Fig. 1. Automated classification of diabetic retinopathy has been an active area of research in computer vision [7]. Early studies using high-bias, low-variance classification techniques performed relatively well at identifying specific features in retinal scans. Consider, for example, the top-hat algorithm used for micro-aneurysm detection [6] [11]. However, especially in early-stage retinopathy, a single feature is unlikely to be a reliable marker of disease burden. Ophthalmologists typically consider multiple features e.g. microaneurysms, dot-blot hemorrhages, cotton-wool spots, exudates, neovascularization, scarring.

More recent approaches have tried to broaden the feature extraction to include other pathological features. K-nearest neighbor [1] [8], support vector machine [10], and ensemble-based methods [2] have all yielded sensitivities and specificities of approximately 90%.

Convolutional neural networks have emerged as a promising tool for medical image classification, and there has been significant interest in their application to retinal imaging [5]. Gulshan *et al.* at Google achieved sensitivities and specificities in the range of 95% for a binary classification task of normal/mild vs moderate/severe using a private dataset of 120,000 images.

However, the detection accuracy across all four classes of the severity spectrum towards DR - no DR (R0), mild DR (R1), moderate DR (R2), and severe DR (R3) - varies significantly. While the R0 and R3 stages are able to be identified with high accuracy, R1 and R2 (representing early-stage retinopathy) are much more difficult to identify. Current accuracies for R1 and R2 stages are reported at under 50%.

This study aims to investigate the poor sensitivity of computer-assisted diagnostics in early-stage retinopathy. We evaluate the deficiencies of traditional CNN implementations and propose several methods to improve detection of subtle pathological features. Specifically, we implement a sliding-window approach using neural networks trained on clinically-selected regions of interest. This generates a probability heatmap representing the likelihood of pathology across the entire retinal scan, which can then be used as the input for a classical machine learning pipeline.

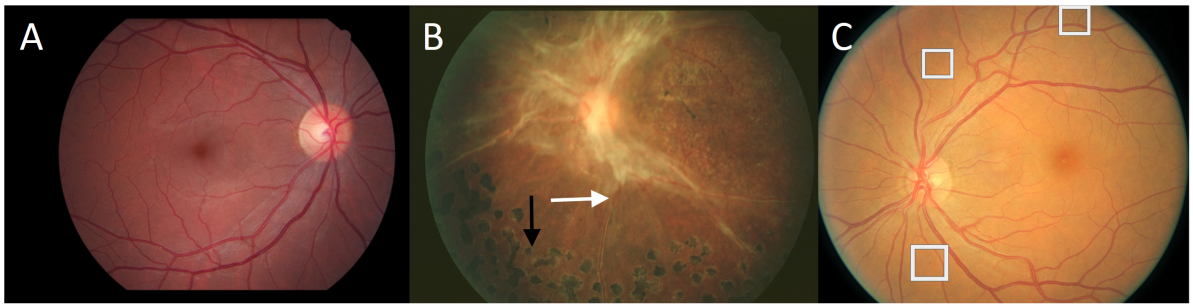

Figure 1. Representative retinal images of DR at various stages of severity : A- normal, B- end stage, C- early stage. Arrows in B point to pathological indications. White boxes in C enclose very small lesions that the CNNs have difficulty discerning. (see figure 2).

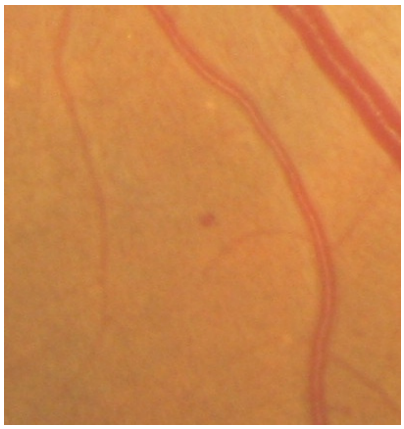

Figure 2. **Micro-aneurysm**
zoomed-in image of the lower left white box in Figure 1 Image C.

2. Data Source

We used a publicly-available Kaggle dataset of 35,000 images with 5 class labels (normal, mild, moderate, severe, and end-stage). This was combined with the Messidor-1 dataset from four French hospitals, containing 1200 fundoscopic images with 4-class labels (normal, mild, moderate, severe) [3] [4]. Both datasets consisted of color photographs that varied in height and width between the low hundreds to low thousands of pixels.

The combined dataset contained images from a diverse patient population with a significant proportion of early diabetic retinopathy. Images showed extremely varied levels of lighting, with variation independent of classification label. Black edges were cropped, images were reshaped to 512x512x3 and pixel intensities were normalized.

To demonstrate the value of image-augmentation, this technique was applied to a subset of the images (900 images in the training set and 150 in the test set).

A subset of the Messidor dataset (535 images total) was used for generating heatmaps and training classifiers, as outlined in Section 3.2

3. Methodology

3.1. Convolutional neural network (CNN)

This study used the GoogLeNet archiecture containing a mixture of low-dimensional embeddings and heterogeneously sized spatial filters [9]. The network contained convolutional blocks with activation on the top layer, followed by batch normalization after each convolution layer. As the number of feature maps increased, one batch normalization per block was introduced in succession.

The max-pooling process was performed with a kernel size of $3x3$ and $2x2$ strides. The network was then flattened to a single dimension after the final convolutional block. Dropout on dense layers was performed until a dense five-node classification layer was reached. A softmax function was used for multi-class classification. This multi-class classification was to distinguish between normal retina and one of four classical retinal pathologies (microaneurysms,dot-blot-hemorrhages, exudates and neovascularization). Cross-entropy Loss was computed for parameter updates.

3.1.1 Image augmentation

The network was first trained with the original preprocessed images. Subsequently, we augmented the number of images in real-time to improve the network's localization ability and reduce over-fitting. Augmentation was performed at each epoch by randomly augmenting images with transformations that preserved collinearity and ratios of distances. We found that contrast-enhancing adaptive histogram equalization gave a significant boost to performance, so this augmentation technique was used throughout.

3.1.2 Sliding-window CNN

To improve the sensitivity of the model to small features such as microaneurysms, we trained the network using image patches of dimensions 448x448x3, centered on features

of interest (see Fig. 3 for representative patches). Features of interest were identified by a trained ophthalmologist. Normal patches for comparison were cropped from the same image as the abnormal patches.

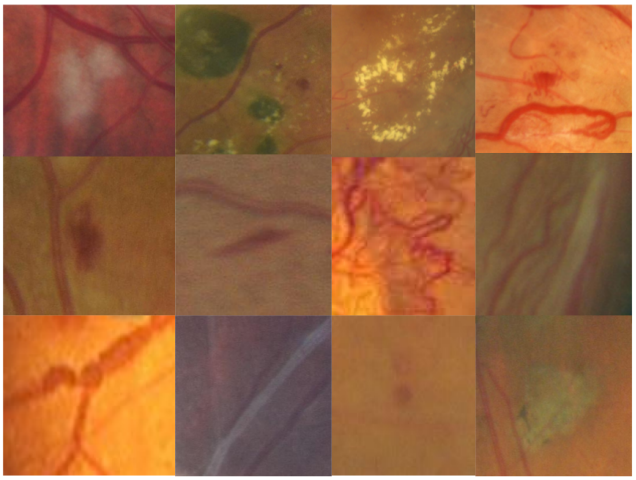

Figure 3. Representative image patches with feature of interest centered

We then defined a kernel of the same dimensions as the patch above. This kernel was convolved with the original image (full retinal scan), at each point producing a probability score of pathology within that patch.

This binary sliding-window approach was then extended to a multi-class classifier. Initially, a model was trained on regions of interest to distinguish betweeen normal retina and one of four key retinal pathologies - microaneurysms, dot-blot-hemorrhages, exudates and neovascularization. Subsequently, these trained kernels were passed over the full scan to give a multi-class probability distribution across the aforementioned pathologies.

Both the binary and multi-cass classifiers allowed us to construct heatmaps representing the probability of pathology within a given image patch (Fig. 4).

3.2. Final classification output

In order to convert these probability heatmaps into a single classification output, we applied a variety of traditional machine-learning methods to define a binary classifier to distiguish normal from pathological scans. The dataset was randomly distributed into training and test sets using a 4:1 ratio (107 samples in the test set).

Firstly, two rule-based approaches were used. (i) A probability threshold γ was assigned. For each value of γ, a classifier was generated whereby a scan was defined as positive if any patch had probability greater than γ. γ was optimised for the training set over the range 0.5 - 1.0. (ii) Scans were defined as positive if a certain number of pathces (ϵ) had probability score greater than threshold γ. This classifier was optimised over both γ and ϵ.

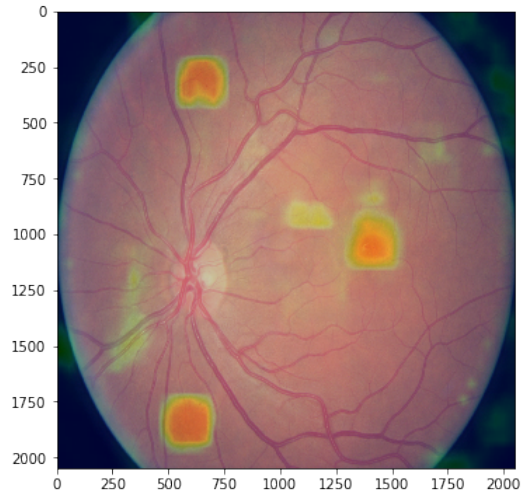

Figure 4. Sliding window heat map of micro-aneurysms in the early retinopathy image featured above, highlighting 3 of the 4 micro-aneurysms in the image

K-nearest neighbours (KNN) was applied using odd values of k between 1 and 50. For each value of k, accuracy and F-1 score were calculated using 5-fold cross validation on the training dataset.

Logistic regression was applied using gridsearch and cross-validation to select the parameter c using the test dataset. C was chosen from among 10 values within a logarithmic scale between 1e-4 and 1e4.

A support vector classifier was trained, again using gridsearch and cross-validation to select the hyperparameters c and gamma, as well as the kernel type.

Finally, a random forest classifier was trained using a gridsearch for estimators from 5 to 20 and depth from 2 to 9. The best validation set parameters were then applied to the test set.

4. Results

To explore the strengths and weaknesses of CNNs, we trained a GoogleLeNet model on a combined dataset from Kaggle and Messidor, containing over 36000 fundoscopic images. We improved on the initial accuracy of 76%, reaching up to 84% as seen in Fig. 5 using the real-time image augmentation technique outlined above. We found that our performance was limited by the inability of CNNs to detect very small pathological features such as micro-aneurysms.

To address this limitation, we trained the network using regions of interest - patches centered on key pathological features compared with equivalent-sized patches containing normal retina (see Figure 3). We found a that with only 900 images in the training set and 150 in the test set, we were able to obtain 95% accuracy with 96% sensitivity and 95% specificity (Figure 5) in *patch* classification.

Figure 5. Training Curve for model on the binary classified Kaggle data set of DR fundoscope images

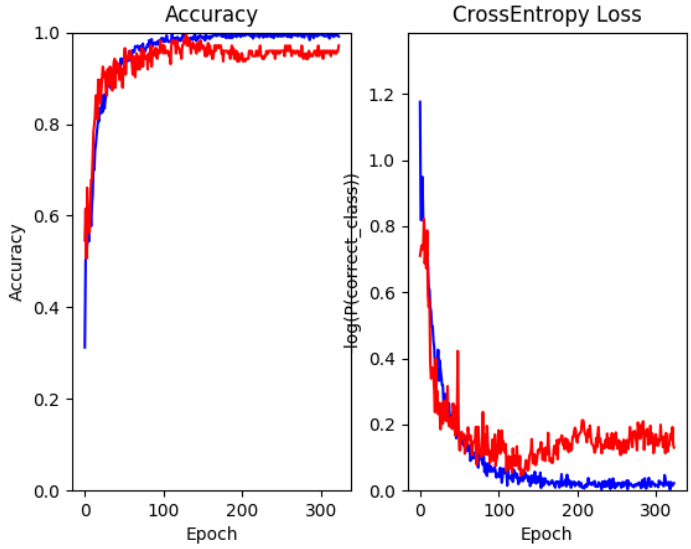

Figure 6. **Training curve on centered patches** 95% accuracy with 96% sensitivity and 95% specificity

Using the patch-trained network as a kernel and convolving over the original scan produced a heatmap with a probability distribution - in one case for the binary output of normal/pathological, and in another case for the multiclass classification across pathological features. Figures 7 and 8 show representative heatmaps for two such features: dot-blot hemorrhages and cotton-wool spots.

For the binary (abnormal/pathological) heatmaps, numerous machine-learning methods were used to build classifiers to produce a single output classification for the scan as a whole. These methods were trained on a subset of the images - 535 images total, of which 54% were positive scans.

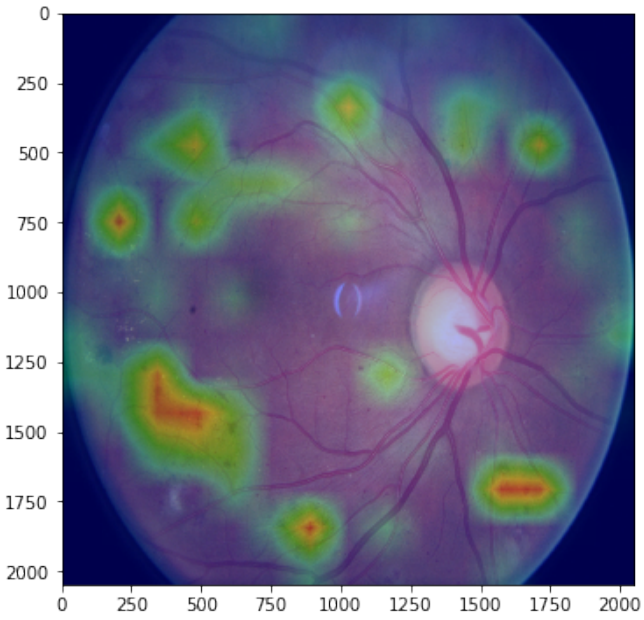

Figure 7. Sliding window heat map of dot-blot hemorrhages in a severe retinopathy image

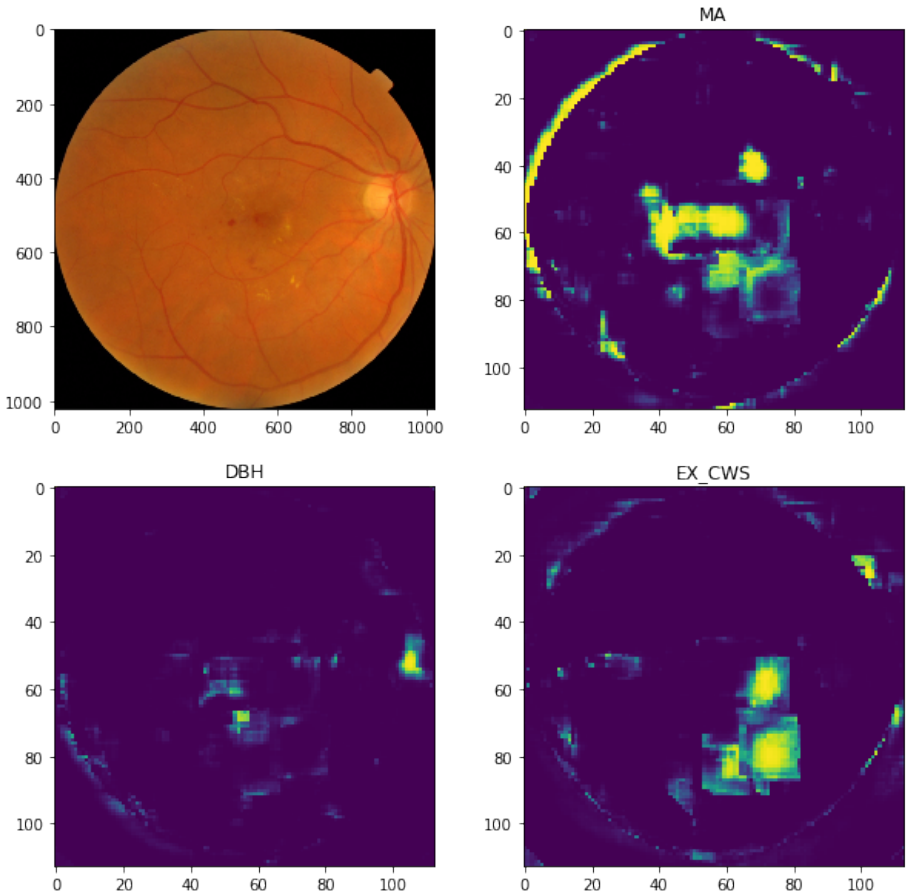

Figure 8. **Feature-wise heatmap:** original image on the upper left. Micro-aneurysms in upper right panel. Dot blot hemorrhages in the lower left panel and exudates or cotton wool spots in the lower right panel. Image from unrelated dataset from Kaggle, the Messidor dataset.

4.0.1 Rule-based classifiers

The first rule-based classifier used a single threshold value γ. Figure 9 shows the accuracy score over γ in the training set. A value of $\gamma = 0.99$ was used for the test set. There was of 0.64 in the test set, relative to the majority classifier of 0.54.

The second rule-based classifier required ϵ patches above threshold γ. Figure 10 shows subplots of the accuracy score relative to ϵ for different values of γ. A classifier was γ = 0.88 and ϵ = 20 yielded an accuracy of 0.67.

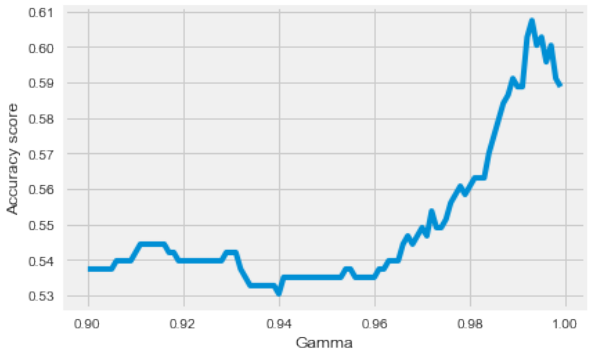

Figure 9. Accuracy score versus gamma using rule-based classifier 1.

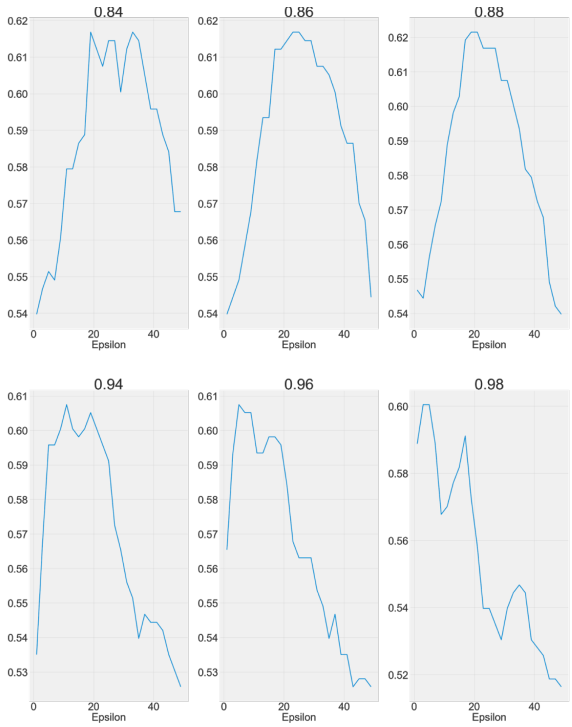

Figure 10. Accuracy score versus epsilon (number of patches above probability threshold γ) for different values of γ.

4.0.2 K-nearest neighbors

The optimal value of k was 3. Applying a 3-nearest neighbour model to the test dataset yielded a precision of 0.56, recall of 0.56 and F1-score of 0.54 on the test dataset.

4.0.3 Logistic regression

Optimization of the regularization hyperparameter c yielded a value of 0.006. Area under the curve (AUC) of the receiver-operating characteristic (ROC) was 0.76.

4.0.4 SVM

Hyperparameter tuning on the test dataset yielded the following optimal parameters with a score of 0.65:
c: 10 , γ: 0.001 , kernel: radial basis function. When applied to the test dataset, this tuned support vector classifier yielded accuracy score of 0.68.

4.0.5 Random forest

The best parameters after gridsearch and cross-validation yielding an accuracy score of 0.61, were as follows:
max_depth: 5 , n_estimators: 20. When applied to the test dataset, this random forest classifier yielded accuracy of 0.77. Figure 11 shows the receiver operating characteristic (ROC). ROC-AUC was 0.80.

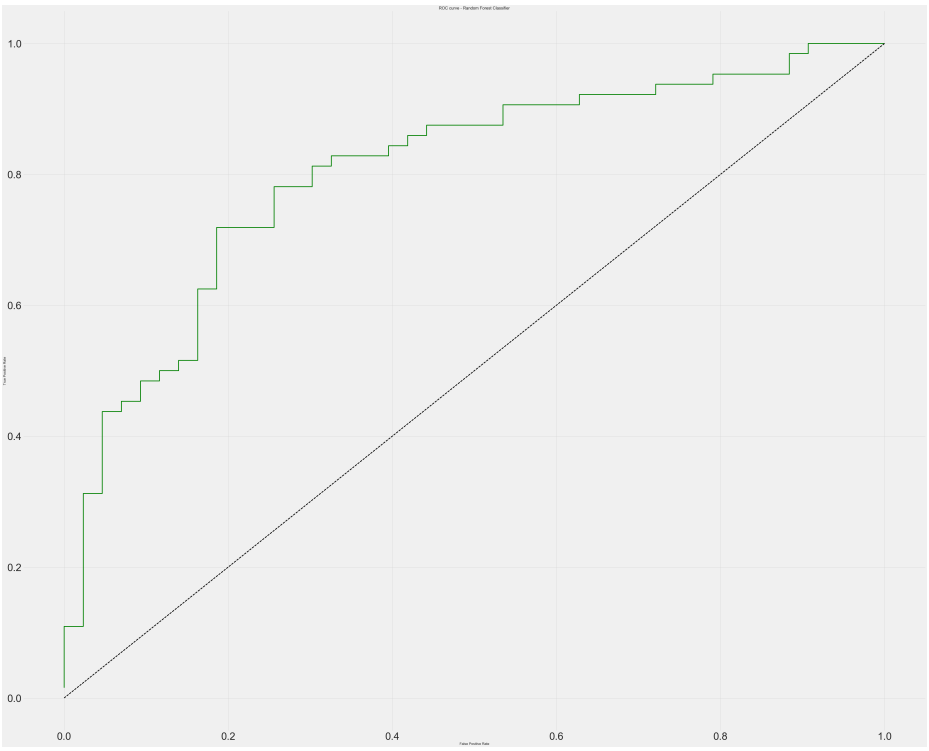

Figure 11. Receiver-operating characteristic for a random forest classifier. AUC = 0.80

Table 1. **Classifier performance** - precision, recall and f1 scores for classifying normal versus abnormal retinal images.

Classifier	Precision	Recall	F1
Rule-based classifiers			
1. Single threshold (γ = 0.99)	0.64	0.64	0.62
2. ϵ patches $>\gamma$ 0.88	0.67	0.67	0.67
Machine learning methods			
KNN	0.61	0.54	0.53
Logistic regression	0.69	0.69	0.69
SVM	0.70	0.68	0.69
Random forest	**0.77**	**0.76**	**0.76**

5. Discussion

In recent years, researchers have incorporated CNNs into the suite of algorithms used to screen for diabetic disease. The high variance and low bias of these models suggested CNNs may be able to detect subtle features of retinopathy that were not captured by traditional feature extraction.

Our study found that, despite the early promise of convolutional neural nets [5], CNNs trained on the entire retinal scan do not effectively detect the subtle pathological changes present in early-stage retinopathies. One possible reason for this is that the GoogLeNet architecture has been optimized to recognize macroscopic features such as those present in the ImageNet dataset, rather than microscopic (but highly malignant) features such as microaneurysms.

Previous work in the field has corroborated this finding, suggesting that the scale-invariance of CNNs has limited its accuracy in retinal classification. This is a problem that is not necessarily remedied with additional data [12]. For example, Gulshan *et al.* reported a 93-96% recall for their binary classification tasks; however this was not improved when training with 60,000 samples versus 120,000 samples.

Visualizations of the features learned by CNNs reveal that the signals used for classification are clearly visible by the observer [13]. Moderate and severe diabetic retinopathy contain macroscopic features at a scale that current CNN architectures are optimized to classify. However, the features that distinguish normal retinas from mild disease reside in less than 1% of the total pixel volume, a level of subtlety that is difficult for both human interpreters and CNNs to detect.

Here we propose a method of clinically-enriched training (on clinician-identified regions of interest) that improves the

capacity of CNNs to detect these subtle features. Using this method to generate heatmaps may be a valuable adjunct in ophthalmology care that could be dynamically overlaid on a retinal scan when it is taken. A technical assistant may then be able to triage patients who may need more urgent attention.

Our study explored whether these heatmaps could form the basis for a binary classification task. The probability scores within the heatmaps were vectorized and concatenated into a sample-feature matrix, on which multiple classifiers were trained. The best approach was a random forest classifier, which achieved accuracy of an accuracy score of 0.77 on a held-out test set.

Overall, machine-learning approaches outperformed the rule-based classifiers based on probability thresholds. This demonstrates that CNNs combined with traditional machine-learning can achieve superior results than the pure probability scores generated from the deep learning approach.

Limitations of the study

Bias is introduced into the deep learning model by factors such as the clinician-identified regions of interest and the predefined patch size. Preliminary work found the sliding window technique was very sensitive to changes in the window size and stride. We hypothesize that this is due to the fact that different features vary in scale, for example, hemorrhages are in general significantly larger than microaneurysms. Further work is warranted to rigorously experiment with different sliding window properties.

Furthermore, although this sliding-window CNN showed promising results, we recognize that its clinical applicability is limited by the computational time required to run several forward passes per image.

The machine-learning on the heatmaps was limited by the number of heatmaps used - 535 images total, due to the computational restrictions. We might improve the accuracy of these classifiers with a larger training set. We could also have tried alternate classifiers such as gradient boosting.

Future work

Future work may include alternative classification approaches on the heatmaps. For example, one might extract shape or texture parameters from the heatmap contours and then perform machine-learning on these features. We could also attempt the a multi-class classification task - either to determine the grade of retinopathy (1-4), or the specific pathology (microaneurysms, dot-blot-hemorrhages etc) present in the scan. Further work is also warranted to combine the classification task with segmentation of subtle pathological features.

Conclusions

In conclusion, this is an exploratory study demonstrating how CNNs trained on a targeted portion of a scan may allow for more clinically meaningful outputs, in the form of heatmaps and an overall classification output. This illustrates the importance of tailoring CNN use to the clinical context - in this case, tackling the scale-invariance issue in retinal scans with small, subtle pathological changes. Further work is needed to evaluate the utility of this patch-based approach in automated retinal screening.

References

[1] M. D. Abràmoff, J. M. Reinhardt, S. R. Russell, J. C. Folk, V. B. Mahajan, M. Niemeijer, and G. Quellec. Automated early detection of diabetic retinopathy. *Ophthalmology*, 117(6):1147–1154, 2010.

[2] B. Antal and A. Hajdu. An ensemble-based system for microaneurysm detection and diabetic retinopathy grading. *IEEE transactions on biomedical engineering*, 59(6):1720–1726, 2012.

[3] E. Decencière, X. Zhang, G. Cazuguel, B. Laÿ, B. Cochener, C. Trone, P. Gain, R. Ordonez, P. Massin, A. Erginay, et al. Feedback on a publicly distributed image database: the messidor database. volume 33, pages 231–234, 2014.

[4] B. Graham. Kaggle diabetic retinopathy detection competition report. 2015.

[5] V. Gulshan, L. Peng, M. Coram, M. C. Stumpe, D. Wu, A. Narayanaswamy, S. Venugopalan, K. Widner, T. Madams, J. Cuadros, 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.

[6] M. Niemeijer, B. Van Ginneken, M. J. Cree, A. Mizutani, G. Quellec, C. I. Sánchez, B. Zhang, R. Hornero, M. Lamard, C. Muramatsu, et al. Retinopathy online challenge: automatic detection of microaneurysms in digital color fundus photographs. *IEEE transactions on medical imaging*, 29(1):185–195, 2010.

[7] S. Philip, A. Fleming, K. Goatman, P. Mcnamee, and G. Scotland. The efficacy of automated disease/no disease grading for diabetic retinopathy in a systematic screening programme. In *British Journal of Ophthalmology*, pages 1512–1517, 2007.

[8] G. Quellec, M. Lamard, P. M. Josselin, G. Cazuguel, B. Cochener, and C. Roux. Optimal wavelet transform for the detection of microaneurysms in retina photographs. *IEEE Transactions on Medical Imaging*, 27(9):1230–1241, 2008.

[9] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2818–2826, 2016.

[10] A. UR. Decision support system for diabetic retinopathy using discrete wavelet transform. *Proceedings of the Institution of Mechanical Engineers, Part H: Journal of Engineering in Medicine*, 227(3):251–261, 2013.

[11] S. Wang, H. L. Tang, Y. Hu, S. Sanei, G. M. Saleh, T. Peto, et al. Localizing microaneurysms in fundus images through singular spectrum analysis. *IEEE Transactions on Biomedical Engineering*, 64(5):990–1002, 2017.

[12] Y. Xu, T. Xiao, J. Zhang, K. Yang, and Z. Zhang. Scale-invariant convolutional neural networks. *arXiv preprint arXiv:1411.6369*, 2014.

[13] J. Yosinski, J. Clune, A. Nguyen, T. Fuchs, and H. Lipson. Understanding neural networks through deep visualization. *arXiv preprint arXiv:1506.06579*, 2015.