Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 90 (2016) 200 – 205

Procedia Computer Science

International Conference On Medical Imaging Understanding and Analysis 2016, MIUA 2016, 6-8 July 2016, Loughborough, UK

Convolutional Neural Networks for Diabetic Retinopathy

Harry Pratt[a,*], Frans Coenen[b], Deborah M Broadbent[c], Simon P Harding[a,c], Yalin Zheng[a,c]

[a]*Department of Eye and Vision Science, Institute of Ageing and Chronic Disease, University of Liverpool, Apex Building, 6 West Derby Street, Liverpool L7 9TX, United Kingdom*

[b]*Department of Computer Science, University of Liverpool, Ashton Street, Liverpool L69 3BX, United Kingdom*

[c]*Royal Liverpool University Hospital, St. Paul's Eye Unit, Prescot Street, Liverpool L7 8XP, United Kingdom*

Abstract

The diagnosis of diabetic retinopathy (DR) through colour fundus images requires experienced clinicians to identify the presence and significance of many small features which, along with a complex grading system, makes this a difficult and time consuming task. In this paper, we propose a CNN approach to diagnosing DR from digital fundus images and accurately classifying its severity. We develop a network with CNN architecture and data augmentation which can identify the intricate features involved in the classification task such as micro-aneurysms, exudate and haemorrhages on the retina and consequently provide a diagnosis automatically and without user input. We train this network using a high-end graphics processor unit (GPU) on the publicly available Kaggle dataset and demonstrate impressive results, particularly for a high-level classification task. On the data set of 80,000 images used our proposed CNN achieves a sensitivity of 95% and an accuracy of 75% on 5,000 validation images.

Peer-review under responsibility of the Organizing Committee of MIUA 2016

Keywords: Deep Learning, Convolutional Neural Networks, Diabetic Retinopathy, Image Classification, Diabetes

1. Introduction

Diabetic Retinopathy (DR) is one of the major causes of blindness in the western world[1,2]. Increasing life expectancy, indulgent lifestyles and other contributing factors mean the number of people with diabetes is projected to continue rising[3]. Regular screening of diabetic patients for DR has been shown to be a cost-effective and important aspect of their care[4]. The accuracy and timing of this care is of significant importance to both the cost and effectiveness of treatment. If detected early enough, effective treatment of DR is available, making this a vital process[5].

Classification of DR involves the weighting of numerous features and the location of such features[6]. This is highly time consuming for clinicians. Computers are able to obtain much quicker classifications once trained, giving the ability to aid clinicians in real-time classification. The efficacy of automated grading for DR has been an active area

* Harry Pratt. Tel.: +447428611330
E-mail address: sghpratt@liverpool.ac.uk

Peer-review under responsibility of the Organizing Committee of MIUA 2016
doi:10.1016/j.procs.2016.07.014

of research in computer imaging with encouraging conclusions[7,8]. Significant work has been done on detecting the features of DR using automated methods such as support vector machines and k-NN classifiers[9]. The majority of these classification techniques are on two class classification for DR or no DR.

Convolutional Neural Networks (CNNs), a branch of deep learning, have an impressive record for applications in image analysis and interpretation, including medical imaging. Network architectures designed to work with image data were routinely built already in 1970s[10] with useful applications and surpassed other approaches to challenging tasks like handwritten character recognition[11]. However, it wasn't until several breakthroughs in neural networks such as the implementation of dropout[12], rectified linear units[13] and the accompanying increase in computing power through graphical processor units (GPUs) that they became viable for more complex image recognition problems. Presently, large CNNs are used to successfully tackle highly complex image recognition tasks with many object classes to an impressive standard. CNNs are used in many current state-of-the-art image classification tasks such as the annual ImageNet and COCO challenges[14,15].

Two main issues exist within automated grading and particularly CNNs. One is achieving a desirable offset in sensitivity (patients correctly identified as having DR) and specificity (patients correctly identified as not having DR). This is significantly harder for national criteria which is a five class problem in to normal, mild DR, moderate DR, severe DR, and proliferative DR classes. Furthermore, overfitting is a major issue in neural networks. Skewed datasets cause the network to over-fit to the class most prominent in the dataset. Large datasets are often massively skewed. In the dataset, we used less than three percent of images came from the 4th and 5th class, meaning changes had to be made in our network to ensure it could still learn the features of these images.

In this paper, we introduce a deep learning-based CNN method for the problem of classifying DR in fundus imagery. This is a medical imaging task with increasing diagnostic relevance, discussed earlier, and one that has been subject to many studies in the past. As far as we are aware, this is the first paper discussing the five class classification of DR using a CNN approach. Several new methods are introduced to adapt the CNN to our large dataset. We then analyse the performance and dissect the capabilities of our network.

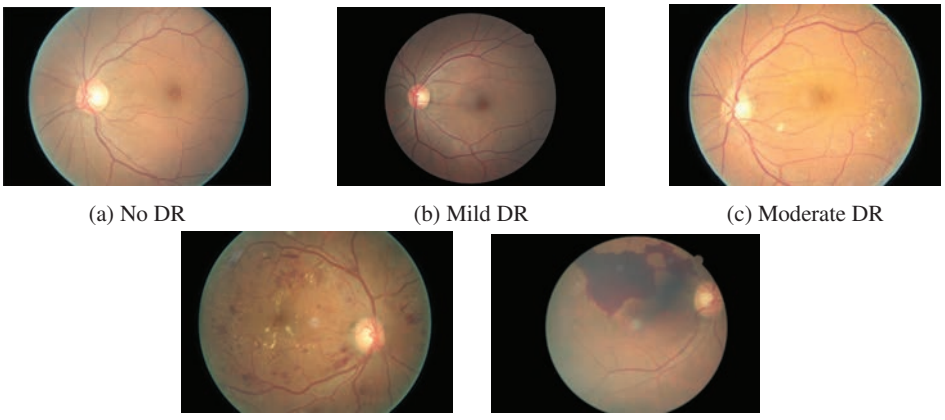
The remainder of this paper is organised as follows. Section 2 presents an overview of related work, section 3 describes the architecture of the CNN and the training methods used in this work, section 4 presents the results from our experiments, section 5 concludes the paper with discussion on the results and future work.

2. Related Work

Extensive research has been carried out on methods for a binary classification of DR with encouraging results. Gardner et al used Neural Networks and pixel intensity values to achieve sensitivity and specificity results of 88.4% and 83.5% respectively for yes or no classification of DR[16]. They used a small dataset of around 200 images and split each image in to patches and then required a clinician to classify the patches for features before SVM implementation.

Neural Networks have also been used in three-class classification of DR. Nayak et al[17] used features such as the area of exudates and the area of blood vessels together with texture parameters. Features are entered into the neural network to classify images into normal, non-proliferative retinopathy and proliferative retinopathy. The neural network used these features as input for classification. The detection results were validated by comparing with grading from expert ophthalmologists. They demonstrated a classification accuracy of 93%, sensitivity of 90% and specificity of 100%. This was carried out on a dataset of 140 images and feature extraction was required on all images in both training and testing which can be time consuming.

The vast majority of research on the five-class classification that has been carried out has used support vector machines (SVMs). Acharya et al[18] have created an automated method for identifying the five classes. Features, which are extracted from the raw data using a higher-order spectra method, are fed in to the SVM classifier and capture the



(a) No DR (b) Mild DR (c) Moderate DR

(d) Severe DR (e) Proliferative DR

Fig 1: Stages of diabetic retinopathy (DR) with increasing severity

variation in the shapes and contours in the images.

This SVM method reported an average accuracy of 82%, sensitivity of 82% and specificity of 88%. Acharya et al[19] also created a five-class classification method by calculating the areas of several features such as haemorrhages, micro-aneurysms, exudate and blood vessel. The features determined to be the most crucial; blood vessels, micro-aneurysms, exudates, and haemorrhages, were extracted from the raw images using image processing techniques. These were then fed to the SVM for classification. A sensitivity of 82%, specificity of 86% and accuracy of 85.9% was achieved using this system. These methods were performed on relatively small datasets and the drop in sensitivity and specificity was likely due to to the complex nature of the five class problem.

Adarsh et al[20] also used image processing techniques to produce an automated diagnosis for DR through the detection of retinal blood vessels, exudate, micro-aneurysms and texture features. The area of lesions and texture features were used to construct the feature vector for the multi-class SVM. This achieved accuracies of 96% and 94.6% on the public 89 and 130 image databases DIARETDB0 and DIARETDB1 respectively.

Each of the previous five class methods required feature extraction from the images before being input to an SVM classifier and have only been validated on small test sets of approximately 100 images. These methods are less real-time applicable than a CNN.

3. Method and Structure

The structure of our neural network, shown in Fig 1, was decided after studying the literature for other image recognition tasks. Increased convolution layers are perceived to allow the network to learn deeper features. For example, whereas

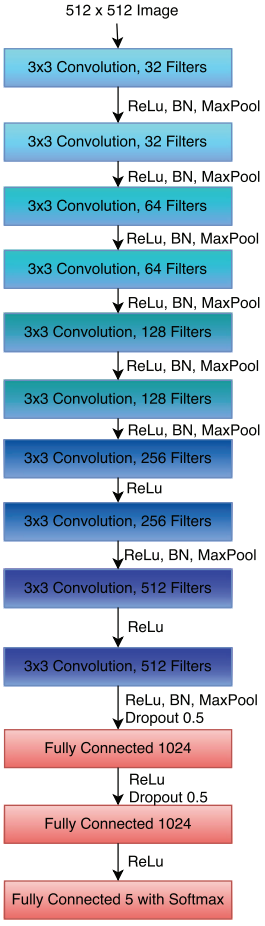
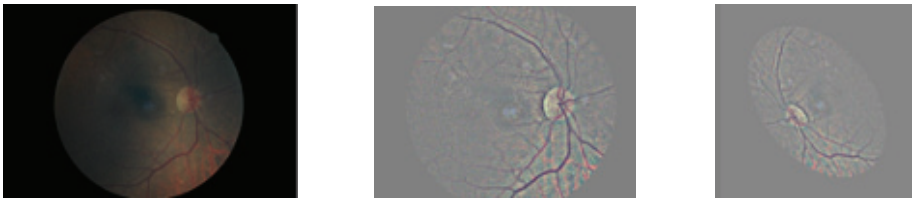


Fig 2: Network architecture



(a) Original image (b) Preprocessed image (c) Augmented image

Fig 3: Illustration of the preprocessing and augmentation processes

the first layer learns edges the deepest layer of the network, the last convolutional layer, should learn the features of classification of DR such as hard exudate. The network starts with convolution blocks with activation and then batch normalisation after each convolution layer. As the number of feature maps increases we move to one batch normalisation per block.

All maxpooling is performed with kernel size 3x3 and 2x2 strides. After the final convolutional block the network is flattened to one dimension. To avoid overfitting we use weighted class weights relative to the amount of images in each class. Likewise, we perform dropout on dense layers, to reduce overfitting, until we reach the dense five node classification layer which uses a softmax activation function to predict our classification. The leaky rectified linear unit [13] activation function was used, applied with a value of 0.01, to stop over reliance on certain nodes in the network. Similarly, in the convolution layers, L2 regularisation was used for weight and biases. The network was also initialised with Gaussian initialisation to reduce initial training time. The loss function used to optimise was the widely used categorical cross-entropy function.

3.1. Dataset, Hardware and Software

The dataset used for testing was provided by the Kaggle coding website (https://www.kaggle.com) and contains over 80,000 images, of approximately 6M pixels per image and scales of retinopathy. Resizing these images and running our CNN on a high-end GPU, the NVIDIA K40c, meant we were able to train on the whole dataset. The NVIDIA K40c contains 2880 CUDA cores and comes with the NVIDIA CUDA Deep Neural Network library (cuDNN) for GPU learning. Through using this package around 15,000 images were uploaded on the GPU memory at any one time. The deep learning package Keras (http://keras.io/) was used with the Theano (http://deeplearning.net/software/theano/) machine learning back end. This was chosen due to good documentation and short calculation time. An image can be classified in 0.04 seconds meaning real-time feedback for the patient is possible.

3.2. Preprocessing

The dataset contained images from patients of varying ethnicity, age groups and extremely varied levels of lighting in the fundus photography. This affects the pixel intensity values within the images and and creates unnecessary variation unrelated to classification levels. To counteract this, colour normalisation was implemented on the images using the OpenCV (http://opencv.org/) package. The result of this can be seen in Fig 3 (b). The images were also high resolution and therefore of significant memory size. The dataset was resized to 512x512 pixels which retained the intricate features we wished to identify but reduced the dataset to a memory size the NVIDIA K40c could handle.

3.3. Training

The CNN was initially pre-trained on 10,290 images until it reached a significant level. This was needed to achieve a relatively quick classification result without wasting substantial training time. After 120 epochs of training on the initial images the network was then trained on the full 78,000 training images for a further 20 epochs. Neural networks suffer from severe over-fitting, especially in a dataset such as ours in which the majority of the images in the dataset are classified in one class, that showing no signs of retinopathy. To solve this issue, we implemented real-time class

weights in the network. For every batch loaded for back-propagation, the class-weights were updated with a ratio respective to how many images in the training batch were classified as having no signs of DR. This reduced the risk of over-fitting to a certain class to be greatly reduced.

The network was trained using stochastic gradient descent with Nestrov momentum. A low learning rate of 0.0001 was used for 5 epochs to stabilise the weights. This was then increased to 0.0003 for the substantial 120 epochs of training on the initial 10,290 images, taking the accuracy of the model to over 60%, this took circa 350 hours of training. The network was then trained on the full training set of images with a low learning rate. Within a couple of large epochs of the full dataset the accuracy of the network had increased to over 70%. The learning rate was then lowered by a factor of 10 every time training loss and accuracy saturated.

3.4. Augmentation

The original pre-processed images were only used for training the network once. Afterwards, real-time data-augmentation was used throughout training to improve the localisation ability of the network. During every epoch each image was randomly augmented with: random rotation 0-90 degrees, random yes or no horizontal and vertical flips and random horizontal and vertical shifts. The result of an image augmentation can be seen in Fig 3 (c).

4. Results

5,000 images from the dataset were saved for validation purposes. Running the validation images on the network took 188 seconds. For this five class problem we define specificity as the number of patients correctly identified as not having DR out of the true total amount not having DR and sensitivity as the number of patients correctly identified as having DR out of the true total amount with DR. We define accuracy as the amount of patients with a correct classification. The final trained network achieved, 95% specificity, 75% accuracy and 30% sensitivity. The classifications in the network were defined numerically as: 0 - No DR 1 - Mild DR 2 - Moderate DR 3 - Severe DR 4 - Proliferative DR.

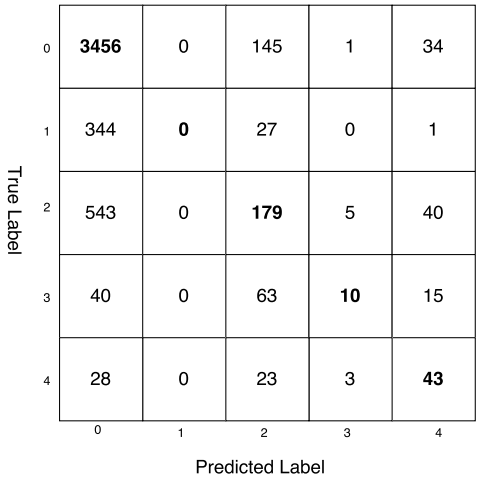


Fig 4: Confusion matrix of final classification results

5. Discussion and Conclusion

Our study has shown that the five-class problem for national screening of DR can be approached using a CNN method. Our network has shown promising signs of being able to learn the features required to classify the fundus images, accurately classifying the majority of proliferative cases and cases with no DR. As in other studies using large datasets high specificity has come with a trade off of lower sensitivity[8]. Our method produces comparable results to these previous methods without any feature-specific detection and using a much more general dataset.

The potential benefit of using our trained CNN is that it can classify thousands of images every minute allowing it to be used in real-time whenever a new image is acquired. In practice images are sent to clinicians for grading and not accurately graded when the patient is in for screening. The trained CNN makes a quick diagnosis and instant response to a patient possible. The network also achieved these results with only one image per eye.

The network has no issue learning to detect an image of a healthy eye. This is likely due to the large number of healthy eyes within the dataset. In training the learning required to classify the images at the extreme ends of the scale was significantly less. The issues came in making the network to distinguish between the mild, moderate and severe

cases of DR. The low sensitivity, mainly from the mild and moderate classes suggests the network struggled to learn deep enough features to detect some of the more intricate aspects of DR. An associated issue identified, which was certified by a clinician, was that by national UK standards around over 10% of the images in our dataset are deemed ungradable. These images were defined a class on the basis of having at least a certain level of DR. This could have severely hindered our results as the images are misclassified for both training and validation.

In future, we have plans to collect a much cleaner dataset from real UK screening settings. The ongoing developments in CNNs allow much deeper networks which could learn better the intricate features that this network struggled to learn. The results from our network are very promising from an orthodox network topology. Unlike in previous methods, nothing specifically related to the features of our fundus images have been used such as vessels, exudate etc. This makes the CNN results impressive but in future we have ideas to cater our network towards this specific task, in order to learn the more subtle classification features. We will also look to compare these networks to five class SVM methods trained on the same datasets.

To conclude, we have shown that CNNs have the potential to be trained to identify the features of Diabetic Retinopathy in fundus images. CNNs have the potential to be incredibly useful to DR clinicians in the future as the networks and the datasets continue improving and they will offer real-time classifications.

Acknowledgment: This study was funded by Fight for Sight in the form of a PhD studentship for HP (http://www.fightforsight.org.uk). We also thank NVIDIA corporation for their donation of GPU card.

References

1. Kocur, I., Resnikoff, S.. Visual impairment and blindness in europe and their prevention. *Brit J Ophthalmol* 2002;**86**(7):716–722.
2. Evans, J., Rooney, C., Ashwood, F., Dattani, N., Wormald, R.. Blindness and partial sight in England and Wales: April 1990-march 1991. *Health Trends* 1996;**28**(1):5–12.
3. Sector, S.P., et al. State of the nation 2012. *Diabetes UK* 2013;.
4. Sculpher, M., Buxton, M., Ferguson, B., Spiegelhalter, D., Kirby, A.. Screening for diabetic retinopathy: A relative cost-effectiveness analysis of alternative modalities and strategies. *Health Econ* 1992;**1**(1):39–51.
5. Benbassat, J., Polak, B.C.. Reliability of screening methods for diabetic retinopathy. *Diabetic Med* 2009;**26**(8):783–790.
6. Grading diabetic retinopathy from stereoscopic color fundus photographsan extension of the modified airlie house classification: Etdrs report number 10. *Ophthalmology* 1991;**98**(5):786–806.
7. Philip, S., Fleming, A.D., Goatman, K.A., Fonseca, S., Mcnamee, P., Scotland, G.S., et al. The efficacy of automated disease/no disease grading for diabetic retinopathy in a systematic screening programme. *Brit J Ophthalmol* 2007;**91**(11):1512–1517.
8. Fleming, A.D., Philip, S., Goatman, K.A., Prescott, G.J., Sharp, P.F., Olson, J.A.. The evidence for automated grading in diabetic retinopathy screening. *Current Diabetes Reviews* 2011;**7**:246 – 252.
9. Mookiah, M.R.K., Acharya, U.R., Chua, C.K., Lim, C.M., Ng, E., Laude, A.. Computer-aided diagnosis of diabetic retinopathy: A review. *Comput Biol Med* 2013;**43**(12):2136–2155.
10. Fukushima, K.. Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biol Cybern* 1980;**36**(4):193–202.
11. Cun, Y.L., Boser, B., Denker, J.S., Howard, R.E., Habbard, W., Jackel, L.D., et al. Advances in neural information processing systems 2. Citeseer. ISBN 1-55860-100-7; 1990, p. 396–404.
12. Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., Salakhutdinov, R.. Dropout: A simple way to prevent neural networks from overfitting. *J Mach Learn Res* 2014;**15**(1):1929–1958.
13. Nair, V., Hinton, G.E.. Rectified linear units improve restricted boltzmann machines. In: *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*. 2010, p. 807–814.
14. Ioffe, S., Szegedy, C.. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv* 2015;URL: `arXiv:1502.03167`.
15. He, K., Zhang, X., Ren, S., Sun, J.. Deep residual learning for image recognition. *arXiv* 2015;URL: `arXiv:1512.03385`.
16. Gardner, G., Keating, D., Williamson, T., Elliott, A.. Automatic detection of diabetic retinopathy using an artificial neural network: a screening tool. *Brit J Ophthalmol* 1996;**80**(11):940–944.
17. Nayak, J., Bhat, P.S., Acharya, R., Lim, C., Kagathi, M.. Automated identification of diabetic retinopathy stages using digital fundus images. *J Med Syst* 2008;**32**(2):107–115.
18. Acharya, R., Chua, C.K., Ng, E., Yu, W., Chee, C.. Application of higher order spectra for the identification of diabetes retinopathy stages. *J Med Syst* 2008;**32**(6):481–488.
19. Acharya, U., Lim, C., Ng, E., Chee, C., Tamura, T.. Computer-based detection of diabetes retinopathy stages using digital fundus images. *P I Mech Eng H* 2009;**223**(5):545–553.
20. Adarsh, P., Jeyakumari, D.. Multiclass svm-based automated diagnosis of diabetic retinopathy. In: *Communications and Signal Processing (ICCSP), 2013 International Conference on*. IEEE; 2013, p. 206–210.