Diabetic Retinopathy Detection through image analysis using Deep Convolutional Neural Networks

Jordi DE LA TORRE, Aida VALLS, Domenec PUIG
Departament d'Enginyeria Informàtica i Matemàtiques
Universitat Rovira i Virgili, Tarragona
e-mail: jordi.delatorre@gmail.com, aida.valls@urv.cat, domenec.puig@urv.cat

Abstract. Diabetic Retinopathy is one of the main causes of blindness and visual impairment for diabetic population. The detection and diagnosis of the disease is usually done with the help of retinal images taken with a mydriatric camera. In this paper we propose an automatic retina image classifier that using supervised deep learning techniques is able to classify retinal images in five standard levels of severity. In each level different irregularities appear on the image, due to micro-aneurisms, hemorrages, exudates and edemas. This problem has been approached before using traditional computer vision techniques based on manual feature extraction. Differently, we explore the use of the recent machine learning approach of deep convolutional neural networks, which has given good results in other image classification problems. From a training dataset of around 35000 human classified images, different convolutional neural networks with different input size images are tested in order to find the model that performs the best over a test set of around 53000 images. Results show that it is possible to achieve a quadratic weighted kappa classification score over 0.75 not far from human expert reported scores of 0.80.

Keywords. deep learning, convolutional neural networks, supervised learning, computer vision, diabetic retinopathy

1. Introduction

Diabetic Retinopathy (DR) is a leading disabling chronic disease and one of the main causes of blindness and visual impairment in developed countries for diabetic patients. Studies reported that 90% of the cases can be prevented through early detection and treatment [1]. Eye screening through retinal images is used by physicians to detect the lesions related with this disease. Due to the increasing number of diabetic people, the amount of images to be manually analyzed is becoming unaffordable. Moreover, training new personnel for this type of image-based diagnosis is long, because it requires to acquire expertise by daily practice.

In 2003 the medical community established a standardized classification based on four severity stages [2] determined by the type and number of lesions (as micro-aneurysms, hemorrhages and exudates): class 0 referring to no apparent retinopathy, class

1 as a Mild Non-Proliferative Diabetic Retinopathy (NPDR), class 2 as Moderate NPDR, class 3 as a Severe NPDR and class 4 as a Proliferative DR.

In this paper we propose an automated classification of retinal images into these 4 classes using Deep Learning techniques. The paper is organized as follows: we present the background of this work, next we explain the characteristics of the available data and why deep learning techniques can be applied over them, next we explain the methodology used for solving the problem, we show the obtained results and finally we expose the conclusions and further steps for improving the results.

2. Background

In the last decade, several attempts to automatize the DR diagnosis through images of the eye fundus have been tested. Those models have been based on extracting hand-crafted fixed engineered features or fixed kernels from the image and, then, using a trainable classifier on top of those features to get the final classification. Using this model, the problem of the DR detection has been tackled by feature extraction using on hand models targeted to the detection of microaneurisms, haemorrhages and exudates in retinal images (e.g. [3], [1]). This type of approach requires a good understanding of the mechanism of the disease to be able to find the important features present in the image. This knowledge is specific of the problem to be solved, requires a lot of labor time and is task specific and thereby not reusable for other different classification problems.

Deep learning techniques are focused on learning multiple levels of representation and abstraction that help to make sense of the hidden information in data such as images. In this way, having a complete set of correctly classified images and without any a priori understanding of the features, the system is able to learn the properties of the image that minimize a defined cost function that is direct or indirectly related with the classification score index that has to be optimized. DCNN have proved to be the best available method for solving the biggest classification challenges, like object recognition in IMAGENET[4]. In the case of DR classification, all images are very similar and very small and vague lesions have to be detected at different locations in order to find the correct severity category of DR. This paper, thus, wants to study the performance of DCNNs for this kind of medical image analysis.

3. Data

The dataset used in this work consists of two independent high resolution image sets (train and test). For every patient right and left eye images are reported. All the images are classified by ophthalmologists according to the standard severity scale presented before [2]. The images are taken in variable conditions: by different cameras, illumination conditions and resolutions.

The training set contains a total of 35.126 images; 25.810 of class 0, 2.443 of class 1, 5.292 of class 3, 873 of class 3 and 708 of class 4. The test set contains a total of 53.576 images; 39.533 of class 0, 3.762 of class 1, 7.861 of class 2, 1.214 of class 3 and 1.206 of class 4. Notice that it is highly imbalanced.

4. Methodology for retinal image classification

Next we explain the steps of the DCNN construction for performing diabetic retinopathy detection based on the available data: evaluation function, data pre-processing and data augmentation, DCNN model, training, testing and the probabilistic combination of the models of both eyes.

4.1. Evaluation function

The performance of the classification model is measured using the quadratic weighted kappa index (κ). It is known as Cohen's Kappa and measures inter-rater agreement for categorical items in multi-class classification problems[5], penalizing the discrepancy quadratically with the distance from the correct class. It is generally thought to be a more robust measure than simple percent agreement calculation, since κ takes into account the agreement occurring by chance. This metric typically varies from 0 (random agreement) to 1 (complete agreement). Negative values are also possible, the maximum possible negative value (-1) indicates a complete disagreement between classes.

The κ coefficient is used in the training and testing stages. We aim at achieving a value close to human performance, which reports values of κ of about 0.80 in the prediction of the correct class in the DR disease. This value is considered excellent because small discrepancies between the class prediction does not affect the treatment of the disease.

4.2. Data pre-processing and data augmentation

DCNNs require large data-sets in order to avoid overfitting. A class balanced data-set is also desirable as well [6]. Thus, a data augmentation [7] scheme has been applied in two stages: first a copy of the training examples of the small classes is done until they have the same number of images as the biggest class. This generates an equilibrated training set. After this first step, the next transformations are applied: cropping, rotation, X and Y mirroring, brightness and contrast correction.

These transformations are applied to every image of the balanced training set and redone for every training epoch. With this scheme we try to make the final prediction invariant to rotation, brightness and contrast over the training set. A channel-wise normalization is also applied to have 0 mean and standard deviation equal to 1 in the input data.

4.3. DCNN Model

We use a set of convolutional layers with dimensional down-sizing blocks between them, also known as pooling layers, to extract the statistical information from the data (feature extraction) , which is passed to the posterior layers to construct more elaborate abstractions (features of features) that are useful for the classification. As a final stage a fully connected set of layers perform the classification based on the information coming from the last layers of the convolutional network. This kind of structure provides an end to end learning process, where either the classes or the features are learned from data with no human intervention.

4.4. Training procedure

As a multi-class classification problem, a log-loss function is used to perform the optimization in the learning stage. The original training set is splitted in two random subsets: one with 90% of the data and other with 10%. The last one is used as a cross validation set for hyperparameter selection. The value of κ is calculated every epoch either for the training or the test set. The model chosen is the one that maximizes the κ over the cross validation set. In all models a leaky ReLU[8] activation function is used. In all layers a batch normalization [9] is applied before the activation function in order to reduce the gradient vanishing problem that occurs in deep networks, reduce the internal covariance shift and improve the regularization. As an additional regularizer a dropout [10] (p=0.5) is performed before the final classification layer. L2 regularization has been tested with no significative improvements in the final classification results. A random initialization based in the Kaiming&He approach[11] is used for all the networks. All models are optimized using stochastic gradient descent with Nesterov momentum.

4.5. Testing procedure

The network is trained with a data augmented scheme that include rotations. Using a trained network, different ensemble[12] schemes are tested over the cross validation set to identify the one that maximizes the classification accuracy. Finally an ensemble of the average of five 72° rotated evaluations is chosen as a testing procedure due to its good compromise between computation costs and performance, achieving on average 0.03 points of κ more than an unique evaluation.

4.6. Probabilistic combination of the models of both eyes

Diabetic Retinopathy usually affects both eyes, specially when the illness is in high severity stages. We propose to take into account the frequency of co-ocurrence of the five DR categories in the images of our dataset. The dataset is big enough to infer from the frequencies of co-occurrence of the classes, the conditional probabilities of having one class in one eye given another class in the other (see table 1). Using the frequentist interpretation that defines an event's probability as the limit of its relative frequency in a large number of trials, we use these frequencies as an estimation for the calculation of the conditional probabilities $P(Left|Right)$ and $P(Right|Left)$. Being $P(Left)$ and $P(Right)$, the probability distributions obtained by our predictive model with the left image and the right image, respectively, we can estimate P_L and P_R using $P_L = P(Left|Right)P(Right)$ and $P_R = P(Right|Left)P(Left)$. To merge the value obtained from the model with the estimation coming from the other eye, we calculate the arithmetic mean. The class with maximum value is the one selected for each eye.

5. Experiments

Different experiments have been conducted to analyze the quality of the classification with different parameters of the DCNN. First, we perform an study of the best image size. Due to the type of image classification that is done, it is crucial to choose the right size of the input images in order to detect the important features involved in the RD severity

Eyes	C0	C1	C2	C3	C4	Sum
C0	12155	407	295	3	11	12871
C1	435	600	171	2	4	1212
C2	336	222	1998	96	50	2702
C3	3	1	87	307	27	425
C4	10	1	39	40	263	353
Sum	12939	1231	2590	448	355	17563

Table 1. Frequencies of combined occurrence of classes (left eye: rows, right eye: columns)

detection. As explained before, cropping is part of the data augmentation scheme. The original size is chosen from the NN input size as $\sqrt{2}$ times the input size. In this way, due to the circular nature of the retina, we maximize the useful information of the square cropped from the center of the image. We have tested four different input sizes: 181x181 cropped to 128x128, 362x362 cropped to 256x256, 543x543 cropped to 384x384 and 724x724 cropped to 512x512 NN input.

Different number of layers have also been studied for each image size. In table 2 are presented the higher classification rates obtained from the best models obtained for different input sizes. The number of layers of the best models are also shown. As the input size and the complexity of the network is increased, the results obtained become better. Greater input sizes allow a better definition of the underlying features present in the image. Increasing the number of layers, allows the construction of more elaborate abstractions in the form of features of features that improve the classification score.

The architecture that performs better is a very deep network of 16 layers: 14 3x3 convolutional layers halving the map size every two layers, one fully connected classification layer and a final softmax output layer. The required deepness indicates that the distinction of the severity categories of DR in images in not an easy classification problem. With the 512x512 model we achieve a $\kappa_{test} = 0.725$ using only the information contained in the examined retine image.

Next, we study if the inclusion of the co-ocurrence information of both eyes is able to improve the quality of the classifier. Table 2 shows the final accuracy obtained combining both eyes for all the sizes. In the 512x512 model the score is increased in about 0.03 κ points, a value that makes our model perform near human level expertise with a final $\kappa_{test} = 0.752$.

		One eye information			Two eyes information		
Layers	Input size	κ_{test}	FN	FP	κ_{test}	FN	FP
12	(3,128,128)	0.488	11.6%	11.5%	0.555	11.2%	12.9%
14	(3,256,256)	0.636	4.4%	28.7%	0.661	4.4%	28.7%
16	(3,384,384)	0.668	7.9%	14.9%	0.722	11.2%	4.0%
16	(3,512,512)	0.725	5.0%	11.9%	0.752	6.5%	7.0%

Table 2. Best classification results achieved with one eye and two eyes information

6. Conclusions and Future Work

In this paper is shown that deep learning techniques are a promising technique for solving medical imaging problems like the diabetic retinopathy detection. Having enough data this method is able to perform near human level expertise achieving κ values of 0.752 not far from the κ achieved by human experts, around 0.80.

Future work will be centered on testing higher resolution input images, newer schemes such as residual networks, the use of alternative cost functions that encode the prior information of the ordering of the classes and other more elaborated methods for combining the information coming from both eyes.

Acknowledgments

This work is supported by the URV grant 2014PFR-URV-B2-60 and the Spanish research projects PI15/01150 and PI12/01535 (Instituto de Salud Carlos III). The authors would like to thank to the California Healthcare Foundation and EyePACS for providing the images used in this study.

References

[1] Jordina Torrents-Barrena, Jaime Melendez, Aida Valls, Pere Romero, and Domenec Puig. Screening for diabetic retinopathy through retinal colour fundus images using convolutional neural networks. In *Artificial Intelligence Research and Development - Proceedings of the 18th International Conference of the Catalan Association for Artificial Intelligence. IOSPress*, pages 259–262, 2015.

[2] Wilkinson CP, Ferris FL 3rd, Klein RE, Lee PP, Agardh CD, Davis M, Dills D, Kampik A, Pararajasegaram R, and Verdaguer JT. Proposed international clinical diabetic retinopathy and diabetic macular edema disease severity scales. *Ophthalmology*, 110(9):1677–82, 2003.

[3] L. R. Sudha and S. Thirupurasundari. Analysis and detection of haemorrhages and exhudates in retinal images. *International Journal of Scientific and Research Publications*, 4:1–5, 2014.

[4] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015.

[5] Joseph L. Fleiss, Jacob Cohen, and B. S. Everitt. Large sample standard errors of kappa and weighted kappa. *Psychological Bulletin*, 72(5):323–327, 1969.

[6] Mazurowski MA, Habas PA, Zurada JM, Lo JY, Baker JA, and Tourassi GD. Training neural network classifiers for medical decision making: The effects of imbalanced datasets on classification performance. *Neural Networks*, 21(2-3):427–436, 2008.

[7] 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.

[8] George E. Dahl, Tara N. Sainath, and Geoffrey E. Hinton. Improving deep neural networks for LVCSR using rectified linear units and dropout. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8609–8613, 2013.

[9] S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *Deep Learning Symposium (NIPS 2015)*, 2015.

[10] P. Baldi and P. Sadowski. Understanding dropout. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems (NIPS)*, volume 26, pages 2814–2822, 2013.

[11] K. He, X. Zhang, S. Ren, and J. Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 1026–1034, 2015.

[12] Zhi-Hua Zhou, Jianxin Wu, and Wei Tang. Ensembling neural networks: Many could be better than all. *Artificial Intelligence*, 137:239–263, 2002.