Improving the Accuracy of Diabetic Retinopathy Severity Classification with Transfer Learning

Narayana Bhagirath Thota
Electrical Engineering & Computer Science
Oregon State University
Corvallis, OR, USA
thotan@oregonstate.edu

Doshna Umma Reddy
Electrical Engineering & Computer Science
Oregon State University
Corvallis, OR, USA
doshna.ummareddy@oregonstate.edu

***Abstract*—Diabetic Retinopathy (DR) is a major cause of blindness in Diabetic patients, and its early detection benefits diagnosis and subsequent treatment methods. In this work, a convolutional neural network uses the VGG-16 model as a pre-trained neural network for fine-tuning, and, thereby classifying the severity of DR. The model also uses efficient deep learning techniques including data augmentation, batch normalization, dropout layers and learn-rate scheduling on high resolution images to achieve higher levels of accuracy. An average class accuracy (ACA) of 74%, sensitivity of 80% at a specificity of 65% and area under the curve (AUC) of 0.80 have been achieved, which are higher than previously reported results obtained using other pre-trained networks or models.**

Keywords—Diabetic Retinopathy, VGG-16, transfer learning, Kaggle competition, Inception-V3, convolutional neural network

I. Introduction

Diabetic Retinopathy is a major cause of blindness in both type-1 and type-2 Diabetics [1]. It is caused by damage to the blood vessels in the retina because higher than normal levels of glucose restrict blood flow to the retina. The U.S. Centers for Disease Control and Prevention estimates that > 30 million people in the US have diabetes [2], and > 93 million are affected by DR.

The severity of diabetic retinopathy can be classified into five levels: 0 = No, 1 = Mild, 2 = Moderate, 3 = Severe, and 4 = Proliferative DR. As depicted in Fig. 1, the prominent visible features of a retina affected with DR include: exudate (mass of cells and fluid that have seeped out of a blood vessel), hemorrhage (an escape of blood from a ruptured blood vessel, especially when profuse), aneurysm (excessive localized enlargement of a blood vessel caused by a weakening of its wall), cotton wool spot (fluffy white patch on the retina due to damage to nerve fibers resulting from accumulations of axoplasmic material within the nerve fiber layer) and neovascularization (the growth of new, but defective blood vessels).

Early detection of these symptoms or features significantly increases the effectiveness of subsequent treatments. These 5 classes can be grouped into 2 categories: Non-Proliferative DR (NPDR, comprising classes 0,1,2 and 3) and Proliferative DR (PDR, class 4). In the case of NPDR, for example, the growth of DR is slowed significantly if the patient's sugar levels are controlled, whereas with PDR, the patient usually requires immediate laser surgery or a vitrectomy (a surgical procedure wherein vitreous humor gel that fills the eye cavity is removed to provide better access to the retina).

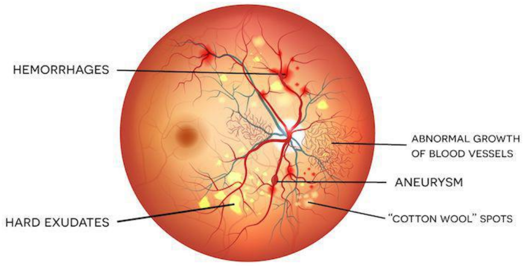

Fig. 1. Various features of DR in a retinal image of an eye [3].

The process of identifying the severity of DR is currently manual which requires a significant amount of time from a trained clinician to examine fundus images of the retina. In areas where the incidence of diabetes in local populations is high and DR diagnosis is most urgent, trained clinicians and diagnostic equipment are often lacking. Thus, it has long been recognized that new tools and techniques are needed to enable widespread inexpensive and robust automated DR screening. Recent research using pattern recognition and deep learning techniques has advanced this goal. Several techniques described in this paper provide further advances.

First, a summary of related state-of-the-art publications is presented. A deep learning algorithm for the automated detection of Diabetic Retinopathy and Diabetic macular edema (i.e., macular swelling) in retinal fundus photographs with high sensitivity and specificity was described by Gulshan, et al. [4]. Venugopal, et al. [5] employed image preprocessing of a fundus dataset to develop a cellular neural network model using Google's GoogLeNet (aka Inception V1). Wei, et al. [6] used transfer learning in a Convolutional Neural Network (CNN) on an expert-labeled laser-scar dataset. The architecture employed only convolutional and pooling layers (i.e., no fully connected layers), which reduced the number of trainable parameters and provided for better interpretability of the network. Zeng et al. [7] described a highly accurate binocular Siamese-like model wherein the five levels of severity of DR were classified into two-groups: (0,1) and (2,3,4). A shortcoming of the model is that it can diagnose DR only in those patients with equal levels of severity in both eyes. Another concern is the classification of patients with a severity level of 1 as non-referable DR, which does indeed correspond to mild level of DR. At the time, it is acceptable in most cases at the time to classify this level of severity (caused by swelling of the tiny blood vessels in the retina) as Non-referable DR. Over a relatively short time, however, the DR can worsen to higher levels of severity if proper care is not taken right away. Many other works have concentrated on binary

978-1-7281-8058-8/20/$31.00 ©2020 IEEE

severity-level classification; e.g., (0) vs. (1,2,3,4) [8][9]. While this approach is more reasonable in terms of the computational requirements, it is obviously advantageous for the patient to know the exact level of severity at any given time. Bravo, et al. [10] performed severity classification using Google's Inception V4 as a pre-trained network along with several other pre-processing techniques to obtain an average class accuracy of 50% on the test data. Zhao, et al. [11] performed severity classification using a BiRA-Net, which combined an attention model for feature extraction and a bilinear model for fine-grained classification.

In this paper, we propose a deep learning model which uses VGG-16 as pre-trained network for fine tuning to classify the fundus images of the retina into the five levels of severity. Final classification results evaluated using the receiver operating characteristic (ROC) show that the proposed model outperforms existing models.

II. Methodology

A. *Image Preprocessing and Augmentation*

The dataset used herein from the Kaggle competition contains 35,126 images of both the left and right eyes [12]. The dataset was obtained from *EyePACS* which is a free platform for screening DR. The distribution of these images classified into the five different severity levels is given in Table 1. Each image is first preprocessed by normalizing, centering, and cropping it to a size of 512 x 512 pixels; the higher resolution (i.e., better quality images) enables a higher classification accuracy. Unfortunately, the huge class imbalances in Table I (e.g., 25810 Level = 0 vs. 708 Level = 4 images) can cause the model to become over fitted and unstable.

To mitigate the problem in this work, the dataset is augmented by randomly rotating images between 0 to 360 degrees, flipping them horizontally or vertically, and changing their brightness, contrast, hue, and saturation levels. Before the augmentation step, however, the dataset is split into training and test sets; each test set has 200 images for each class. The remaining 34,126 images are again split into training and validation sets with 100 images for each class. In the end, augmentation is done on the remaining images to create ~100,000 images, wherein the classes have approximately equal numbers of training examples.

TABLE I. Class Distribution of Images in the Dataset

Class	Level of Severity	No. of images
0	No DR	25810
1	Mild DR	2443
2	Moderate DR	5292
3	Severe DR	873
4	Proliferative DR	708

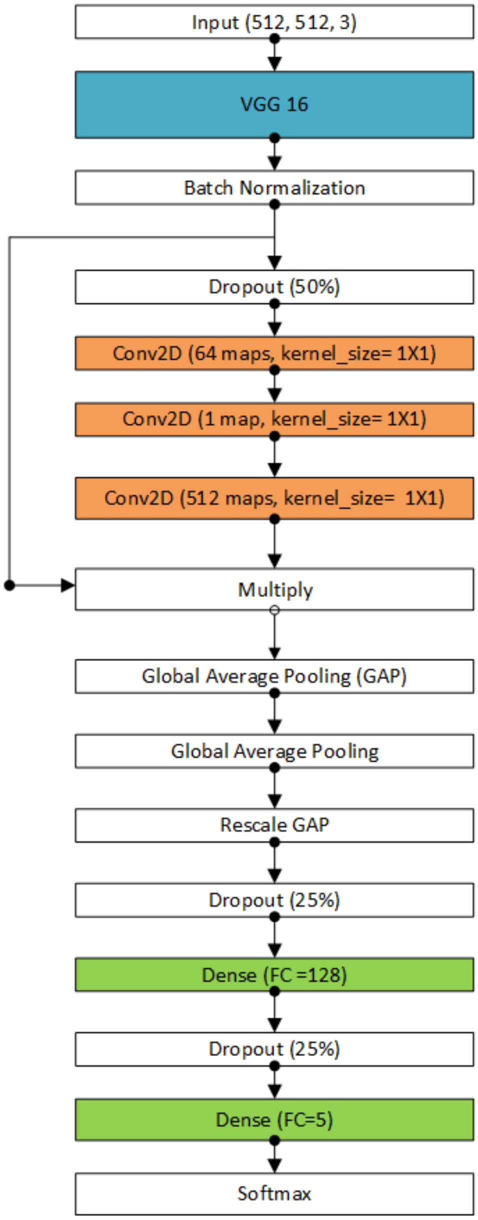

Fig. 2. Architecture of the proposed CNN.

B. *Architecture of the Convolutional Neural Network*

A block diagram of the deep learning CNN model presented in this paper is shown in Fig. 2. It accepts fundus images corresponding to either the left or the right eye as inputs and transmits them into the subsequent layers of the CNN.

The network uses the VGG-16 architecture (Fig. 3) as a pre-trained model for fine-tuning. VGG16, a CNN model proposed by Simonyan and Zisserman [13], achieved a 92.7% top-5 test accuracy in ImageNet (a dataset of over 14 million images belonging to 1000 classes). It improves on AlexNet [14] by replacing large kernel-sized filters with multiple 3x3 kernel-sized filters one after the other. The three convolutional layers in last block (Block-5) are made trainable in this design. The three Dense layers in the VGG-16 model are removed and replaced by a Dense layer with 128 neurons and a rectified linear unit (RELU) activation function. It is a piecewise linear function that outputs the input directly if it is positive, and zero otherwise.

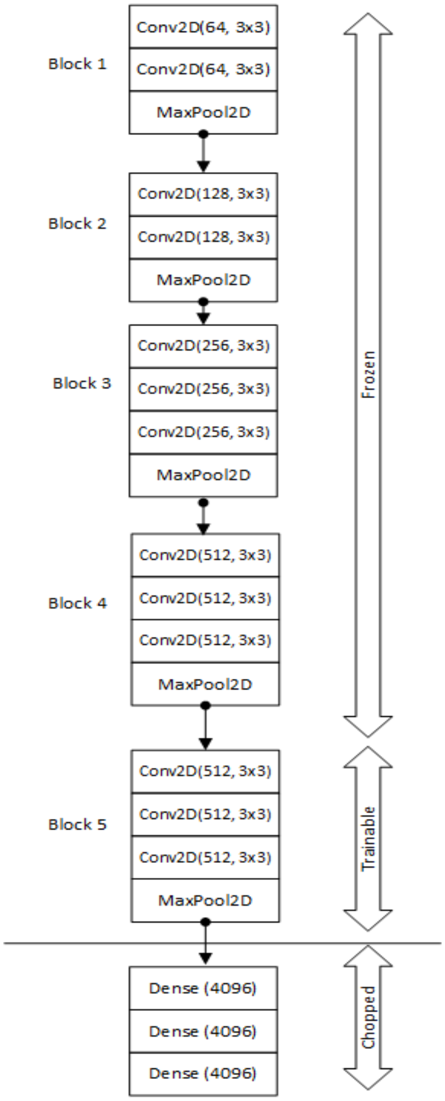

Fig. 3. Architecture of the VGG-16 (pre-trained) model in which the last block, Block 5, is made trainable.

The output layer has 5 neurons and dropout layers are also used at different levels to prevent any overfitting. Batch normalization [15] has been used to speed up the training process, and Global Average Pooling (GAP) [16] layers to minimize overfitting by reducing the total number of parameters in the model. The training model parameters are summarized in Table 2.

Categorical cross entropy has been chosen as the loss function to be optimized. It compares the distribution of the predictions (the activations in the output layer, one for each class) to the true distribution where the probability of the true class is set to 1 and to 0 for the other classes. The categorical cross entropy loss function is given by:

$$L(y,\hat{y}) = -\sum_{j=0}^{M}\sum_{i=0}^{N}(y_{ij} * \log(\hat{y}_{ij})) \qquad (1)$$

where $\hat{y}_{ij}$ is the predicted value and y_{ij} is the actual one-hot encoded value for a given class. A separate loss for each class label per example is calculated and results are summed for N classes (for one example). Similarly, the loss is computed for each of the M examples and all the losses are summed to give the total loss for the training and validation sets.

TABLE 2. PARAMETERS FOR TRAINING THE MODEL

Parameter	Chosen value(s)
Learning Rate	0.001(Reduces by a factor of 10, every 3 epochs, if validation loss remains unchanged)
Optimizer	ADAM (Adaptive Moment Estimation)
Loss function	Categorical Cross Entropy
Early Stops/ Epochs	Waits for 8 epochs (once validation loss stops decreasing) and stops. Otherwise, it continues for 30 epochs
Batch Size	100

III. RESULTS AND DISCUSSION

The training process described above has been implemented on Nvidia Tesla V100 High Performance GPU with 16 GB Memory, 61 GB RAM and 100 GB SSD. It runs either for 8 hours (30 epochs) or for 8 consecutive epochs wherein no improvement in the validation loss invokes the Early Stopping mechanism. Learning Rate Scheduling (reducing Learn Rate by a factor of 10, for 3 epochs when there is no reduction in validation loss), has been found to immensely help in this training process.

To evaluate the performance of the model, the values of precision, recall (i.e., sensitivity), F1-score, specificity, and false positive rate are computed using the following equations:

$$Precision = \frac{TP}{FP+TP} \qquad (2)$$

$$Recall = Sensitivity = \frac{TP}{TP+FN} \qquad (3)$$

$$F1\text{-}score = \frac{2*P*R}{P+R} \qquad (4)$$

$$Specificity = \frac{TN}{FP+TN} \qquad (5)$$

$$False\ Positive\ Rate = 1 - Specificity \qquad (6)$$

where TP = No. of True Positives, FP = No. of False Positives, TN = No. of True Negatives, and FN = No. of False Negatives with P = Precision and R = Recall.

A macro-F1 score is obtained by taking an average of the F1-scores for all 5 classes, whereas a Micro-F1 score is obtained by computing the harmonic mean between the micro-precision and micro-recall values. The micro-precision value is determined across all the classes by taking the ratio of the sum of all True Positives to the sum of all Predicted Positives. Similarly, a micro-recall value is obtained across all the classes by taking the ratio of the sum of all True Positives to the sum of all Actual Positives. The results are also used to plot the Receiver Operating Characteristic (ROC) curve as shown in Fig. 4. Very good values are achieved: The area under the curve (AUC) is 0.80, sensitivity is 80%, and the specificity is about 65%.

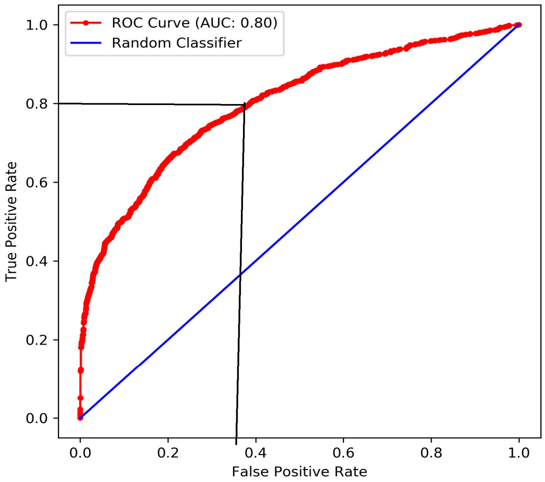

Fig. 4. Receiver Operating Characteristic (ROC) curve (AUC = 0.8) for the 5-level severity classification.

TABLE 3. COMPARISON OF SEVERITY CLASSIFICATION RESULTS

	ACA	Macro-F1	Micro-F1
Bravo, et al. [9]	0.5051	0.5081	0.5052
ResNet-50 [10]	0.4689	0.4753	0.4689
Bi-ResNet [10]	0.4889	0.5503	0.4897
RA-Net [10]	0.4717	0.5268	0.4724
BiRA-Net [10]	0.5431	0.5725	0.5436
This work	**0.74**	**0.60**	**0.68**

The 5-level severity classification accuracy levels are, as expected, not as high as for the two-level classification (Non-Referable DR (0 1) vs. Referable DR (2 3 4) or the Healthy (0) vs Sick (1 2 3 4) results. However, the model proposed herein achieves 74% accuracy, which is much higher than for the previous approaches as listed in Table 3 [11]. Average Class Accuracies (ACA) as well as Macro-F1 and Micro-F1 scores on unseen test data are shown.

As the retinal images provided here (in the Kaggle) may have been obtained through different hardware equipment (and by operators with different levels of expertise), they can have a large variation in the quality of images. For example, due to low contrast, blurred or lower quality images, an early stage or signs of DR can be easily obscured. Therefore, image quality is a very important factor in the sensitivity of the DR Detection process [8]. Other possible causes for the accuracy levels to be less than 80% could be because of the very subtle differences between Level-0 and Level-1 DR. This can be inferred from some earlier work [9-10] wherein the Non-Referable vs. Referable DR classification always seems to give higher accuracies than the Healthy vs. Sick eye classifications. Although beyond the scope of this paper, this issue needs to be investigated more deeply in future work. Also, better methods are needed to distinguish the subtle differences between Level-0 and Level-1 images.

IV. CONCLUSIONS

A CNN model employing VGG-16 as a pre-trained neural network to train on fundus images to predict the level of severity of DR for each input image is presented in this paper. The model uses efficient deep learning methods including data augmentation, batch normalization, fine-tuning the last block (Block-5) of VGG-16, learn-rate scheduling, etc., on high resolution images to achieve higher classification accuracies.

The evaluation results obtained using Nvidia Tesla V100 High Performance GPU show that the proposed model achieved an AUC = 0.80 for 5-level classification, which is better than for existing methods using Inception-V3 or ResNet50. Moreover, the results also demonstrate the potential for early detection and identification of the severity of DR. To achieve further increases in accuracy, more training data is needed, hopefully with an equal number of images in each class.

REFERENCES

[1] S. Das and C. Malathy, “Survey on the diagnosis of diseases from retinal images,” *J. of Physics: Conference Series, National Conf. on Mathematical Techniques and its Applications*, 2018, pp. 1-11.

[2] National Diabetes Statistics Report, Centers for Disease Control and Prevention (CDC), US Dept. of HHS, “Estimates of Diabetes and its Burden in the United States,” pp. 1-20, 2017.

[3] “Treating Diabetic Retinopathy,”
https://www.eyeops.com/contents/our-services/eye-diseases/diabetic-retinopathy

[4] V. Gulshan, et al., “Development and validation of a deep learning algorithm for detection of Diabetic Retinopathy in retinal fundus photographs,” *J. of the American Medical Association*, vol. 316, no. 22, pp. 2402-2410, Dec. 13, 2016.

[5] G. Venugopal, R. Vishwanathan, and R. Joseph, “How AI enhances and accelerates Diabetic Retinopathy detection,” *Cognizant Global Technology Office*, pp. 1-16, Feb. 2018.

[6] Q. Wei, X. Li, H. Wang, D. Ding, W. Yu, and Y. Chen, “Laser scar detection in fundus images using convolutional neural networks,” *Asian Conf. on Computer Vision*, Dec. 2018, pp. 191-206.

[7] X. Zeng, H. Chen, Y. Luo, and W. Ye, “Automated detection of Diabetic Retinopathy using a binocular Siamese-like convolutional network,” *IEEE Intl. Symp. on Circuits and Systems*, May 2019, pp. 1-5.

[8] A. Rakhlin, “Diabetic Retinopathy detection through integration of deep learning classification framework,” *bioRXiv*, pp. 1-11, Feb. 2017.

[9] S. Islam, Md. Hasan, and S. Abdullah, “Deep learning based early detection and grading of Diabetic Retinopathy usingretinal fundus images,” *arXiv:1812.10595*, pp. 1-12, Dec. 2018.

[10] M.A. Bravo and P.A. Arbelaez, “Automatic Diabetic Retinopathy classification,” *Proc. SPIE 10572, Intl. Conf. on Medical Information Processing and Analysis*, Nov. 2017, pp. 1-10.

[11] Z. Zhao, et al., “BiRA-Net: Bilinear attention net for Diabetic Retinopathy grading,” *arXiv:1905.06312*, pp. 1-5, July 2019.

[12] “Kaggle: Diabetic Retinopathy Detection,” https://www.kaggle.com/c/Diabetic-Retinopathydetection

[13] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv:1409.1556*, pp. 1-14, Apr. 2015.

[14] A Krizhevsky, I Sutskever, GE Hinton, “Imagenet classification with deep Convolutional Neural Networks,” Advances in Neural Information Processing Systems, 1097-1105, 2012

[15] M S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," *arXiv:1502.03167*, pp. 1-11, Mar. 2015.

[16] M. Lin, et al., "Network In network," *arXiv:1312.4400v3*, pp. 1-10, Mar. 2014.