Resolution Impact on Deep Learning Approaches for Diabetic Retinopathy Detection

Francis B. Moreira and Philippe O. A. Navaux
Informatics Institute
Universidade Federal do Rio Grande do Sul
Porto Alegre, Brazil
{fbmoreira, navaux}@inf.ufrgs.br

Beatriz D. Schaan, Josiane Schneiders and Mateus A. dos Reis
Universidade Federal do Rio Grande do Sul
Hospital de Clínicas de Porto Alegre
Porto Alegre, Brazil
bschaan@hcpa.edu.br

***Abstract*—Screening for diabetic retinopathy in diabetic patients is the most effective way to reduce the risk of sight loss. The demand for screening is increasing, and the currently overloaded medical doctors can not efficiently attend it. Therefore, there is a need for a methodology to automate and increase the efficiency of screening. In this study, we discuss the development of deep neural networks to detect diabetic retinopathy. We have achieved 0.93 area under the receiver operating characteristic curve.**

I. Introduction

Diabetes is a chronic disease associated with hyperglycemia due to failure in the insulin release or insulin resistance. It has an increasing prevalence, currently affecting 12% of the Brazilian population [16]. High blood glucose levels (Hyperglycemia) cause damage to the blood vessels of the retina (diabetic retinopathy), potentially leading to blindness. As a prevention method, retinographies are taken and analyzed by ophthalmologist doctors to evaluate whether a person suffers from this symptom [14]. Effective screening for diabetic retinopathy that is evaluated by ophthalmologists has been proven to reduce the risk of sight loss [1]. However, the assessment for retinopathy in our country is low (13.2% at primary health care, 11.5% at Family Health teams, 14.9% at Basic Health Care Units, and 35.9% at the tertiary health care unit) [13]. There is a significant shortfall of ophthalmologists in developing countries, and resources are limited [11].

In this context, machine learning offers a prime opportunity to perform detection on a large scale within a limited budget. A neural network model can make an initial analysis of retinographies taken by a technician. This process serves as a filter, thus reducing the number of ophthalmologists required [2], [3], [4], [5].

We perform this work in collaboration with Endocrinology Division of Hospital de Clínicas de Porto Alegre, as a part of the project "Diabetic retinopathy screening in patients with diabetes mellitus: validation of innovative method (machine learning). In this study, we show the results obtained with deep, convolutional neural networks using inception v3. We demonstrate that higher resolution increases the average neural network accuracy. We reach over 0.93 area under the curve (AUC) in the receiver operating characteristic (ROC) of our best model.

In Section II, we discuss the related work covering the used techniques and their application to diabetic retinopathy. In Section III, we show the results we obtain with our models while detailing the methodology and the effects observed with changes applied in images. In Section IV, we conclude this work and briefly discuss our future directions.

II. Background

Machine learning has gone through a revolution in the last decade. The increase in computational power has enabled artificial intelligence researchers to create much more complex models. Although Hornik et al. [6] formally proved that neural networks with a single hidden layer could be universal approximators, specialization of multiple layers has shown that these "approximators" can become accurate enough to perform human tasks.

One particularly well-known case is the deep convolutional neural network [9]. It emulates human eye behavior by filtering sub-areas of the eye through convolutions. That is, instead of connecting all layers of neurons as in a regular neural network, it constructs a layer by combining sectors of neurons from the previous layer, as seen in Figure 1. Each neuron observes only a portion of the image, much like in the human eye. Convolutions between different regions create the identification of shapes, defining features that will identify classes. Thus, deep CNNs have layers that automatically perform feature engineering. By combining multiple layers of convolution, max-pooling, rectified linear units, and fully connected layers, a convolutional neural network can accurately identify objects in an image [8].

Several works have improved the field of computer vision, with many focusing on the Imagenet dataset in the "Imagenet Large Scale Visual Recognition Challenge" (ILSVRC) [12]. The most popular and currently used architecture for a deep neural network is the Inception architecture [15]. The idea behind inception is to use features from multiple resolution levels in an efficient way, using a very deep neural network, and exploiting general-purpose graphic processing units (GPGPU) to train the neural network efficiently.

These ideas have recently been used to detect diabetic retinopathy in retinographies. In Gulshan et al.'s work [5],

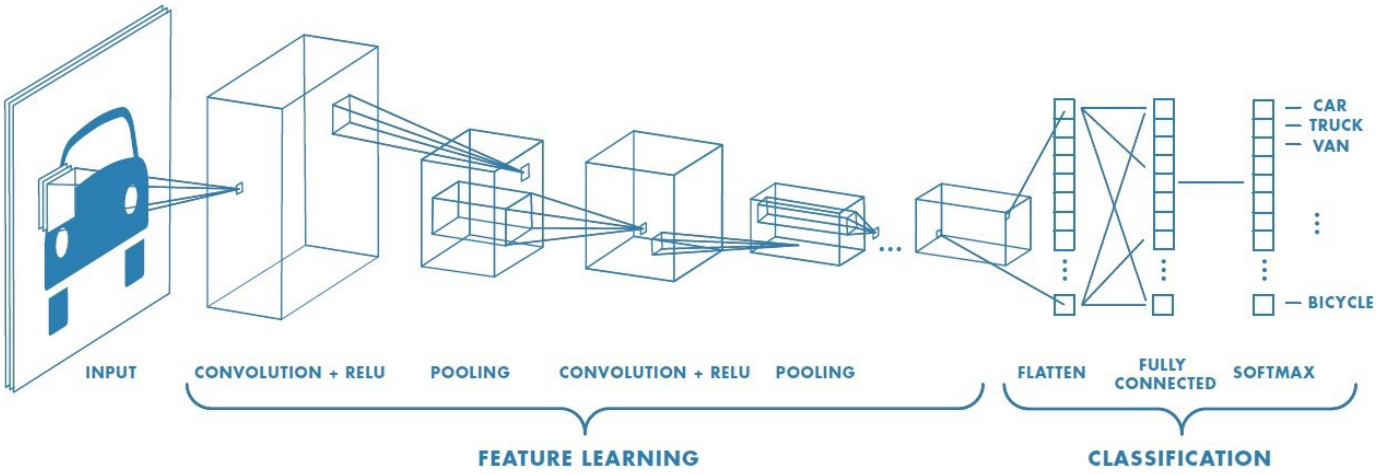

Fig. 1. Convolutional Neural Network abstraction. Source: https://www.mathworks.com/

a convolutional neural network-based in the Inception architecture version 3 is used. They feed the neural network with the data from three Indian hospitals (Aravind Eye Hospital, Sankara Nethralaya, and Narayana Nethralay) and the EyePACS dataset [2]. For validation of the network, they used additional EyePACS images and the Messidor-2 dataset [3]. An essential step in their procedure was to reevaluate all images with several graders. Their algorithm achieved an AUC of the ROC of 0.991 for EyePACS and 0.990 for Messidor-2. Thus by selecting a high sensitivity point, the algorithm achieves 97.5% sensitivity and a 93.4% specificity for the EyePACS dataset; and it achieves 96.1% sensitivity and 93.9% specificity for the Messidor-2 dataset.

These results are far superior to other works on the same task [10], [17], [19], [4]. A recent study by Voets et al. [18] attempted to reproduce these results with the EyePACS dataset available in the Kaggle competition for diabetic retinopathy [1], but it was not able to do so. After a significant change to the activation function of the final layer, the reproduction algorithm was able to achieve 0.94 AUC for EyePACS and 0.80 AUC for Messidor. The authors of the reproduction point four possible reasons for this difference; first, the reproduction had a single grade per image, whereas the original study had multiple grades as mentioned above. Second, the published list of hyper-parameters from the original study does not detail the normalization method and validation procedure used, so the reproduction algorithm is not able to achieve the same level of tuning like the original. Third, there might be errors in the original study or methodology. Fourth, the reproduction study has errors in its methodology. The study demonstrates the difficulty of reproducing a deep learning algorithm without an available model or code. The models are complex, and any lack of information will make the reproduction of a model incomplete.

Krause et al. [7] extended the original work. The focus of this study is to use adjudication to quantify errors in diabetic retinopathy grading based on individual graders and majority decision. They use the improved data to train a better neural network, using the Inception v4 architecture and increasing the resolution of the images for training. The work has found that out of the discrepancies between adjudication by retinal specialists and the majority decision of ophthalmologists, the most common were missing microaneurysms (36%), artifacts (20%), and misclassified hemorrhages (16%). They then compare the performance of the ophthalmologist majority decision (OMD) and the deep learning algorithm (DLA), using the adjudicated consensus of retinal specialists as the reference standard. For moderate or worse diabetic retinopathy, the OMD obtains 83.8% sensitivity and 98.1% specificity, whereas DLA obtains 97.1% sensitivity, 92.3% specificity, and AUC of 0.986. Thus they found that by using a small number of adjudicated consensus grades from retinal specialists as a tuning dataset and higher-resolution images as input, the algorithm improved in AUC from 0.934 to 0.986 for moderate or worse diabetic retinopathy detection.

In this study, we base our code on Voets' reproduction [18], but adapting ideas from the more recent work from Krause et al.' [7], such as higher resolution for input images.

III. Experiments

A. Algorithm Development

For an initial assessment of the viability of using machine learning as a screening tool, we trained ten neural networks based on the Inception v3 architecture. We used the same parameters available in Gulshan et al. 's work [5] and its reproduction [18]. Our initial codebase is the reproduction's codebase, available in https://github.com/mikevoets/jama16-retina-replication. All networks were initialized with imagenet weights for all layers except the fully connected layer on top, which received random weights. In Figure 2, the 42 layers and operation of the Inception v3 architecture are detailed.

As input, we used Kaggle's EyePACS dataset, redimensioning every picture to $299x299$ pixels. The dataset contains

[1]https://www.kaggle.com/c/diabetic-retinopathy-detection/data

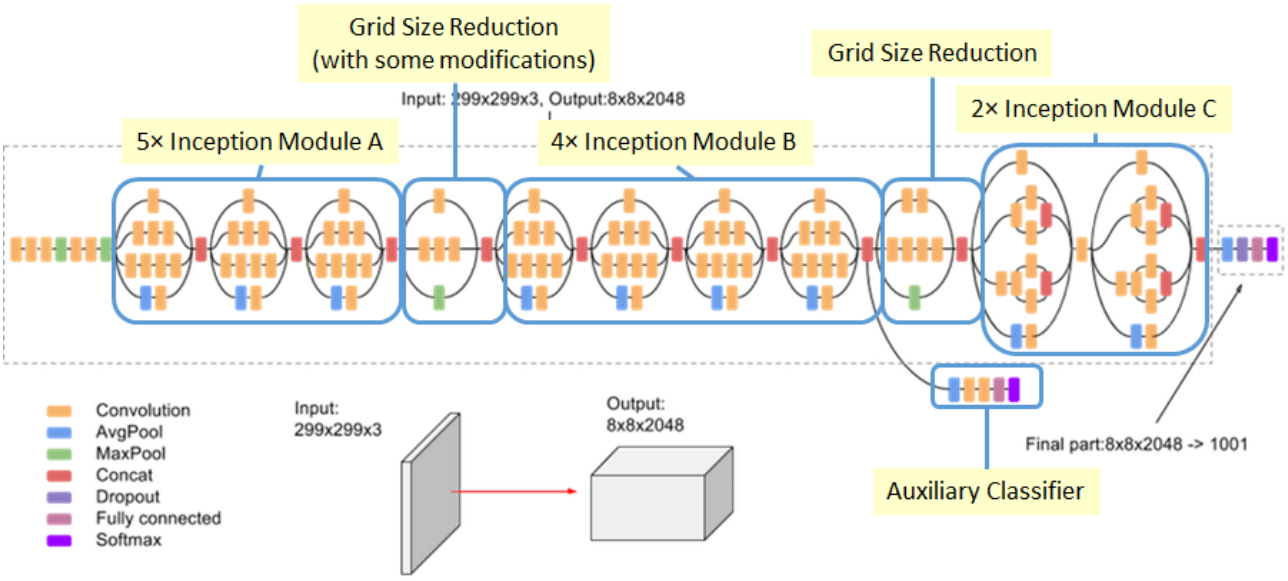

Fig. 2. Inception V3 architecture. Source: https://cloud.google.com/tpu/docs/images/inceptionv3onc–oview.png

$88,612$ images, where $65,343$ have no signs of disease; $6,205$ present mild retinopathy; $13,153$ present moderate retinopathy; $1,997$ present grave retinopathy; and $1,914$ present proliferative retinopathy. We split these images into binary training and testing datasets. Thus, we perform a binary classification: either the patient has mild or no retinopathy (class 0), or he presents moderate or worse retinopathy (class 1). We used the same data augmentation techniques to increase the set of conditions in the training dataset artificially: left to right flip, random saturation, random hue, random brightness, and random contrast. Currently, we only transform the images and do not add any new image.

The training dataset consists of $40,688$ images in class 0 and $16,458$ images in class 1. We use 80% of the training dataset to optimize the weights of the neural network and 20% of the dataset to tune hyperparameters, such as cross-entropy and AUC.

B. Receiver Operating Characteristic Curves of the Ensembles

To illustrate the performance of our models, we used receiver operating characteristic curves. These are graphical plots to illustrate the performance of a binary classification system given different discrimination thresholds. One axis represents the true positive rate, also known as sensitivity, while the other represents the false positive rate, also known as specificity. Sensitivity is a ratio between the number of images in class 1 the model correctly labeled as class 1 and all images in class 1. Specificity is a ratio between the number of images in class 0 the model correctly labeled as class 0 and all images in class 0. Each curve in our figures has 200 threshold points.

Figure 3 presents the receiver operating characteristic curve of the ensemble model composed of the ten neural networks.

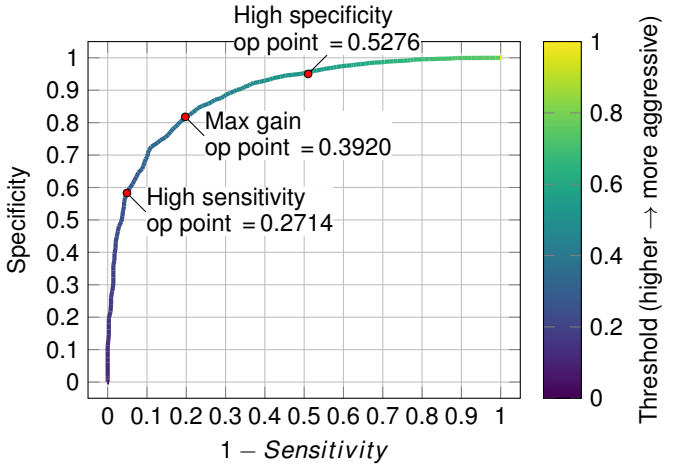

Fig. 3. Receiver Operating Characteristics of the ensemble model composed of ten neural networks with $299x299$ pixels input image resolution.

The three chosen points indicate a high specificity point, a maximum gain point, and a high sensitivity point. The high specificity point achieves 95.02% specificity, 51.05% sensitivity. The maximum gain point achieves 81.78% specificity, 80.26% sensitivity. The high sensitivity point achieves 58.30% specificity, 95.10% sensitivity. The AUC of this model's ROC is 0.89.

To improve upon this model, we have discussed with the endocrinologists from the collaborating hospital regarding the disease. We have found that the lesions and details of the disease can be represented in tiny segments of the image, which led us to increase the resolution of the input image, much like Krause et al.'s work.

Figure 4 presents the ROC curve of the ensemble model

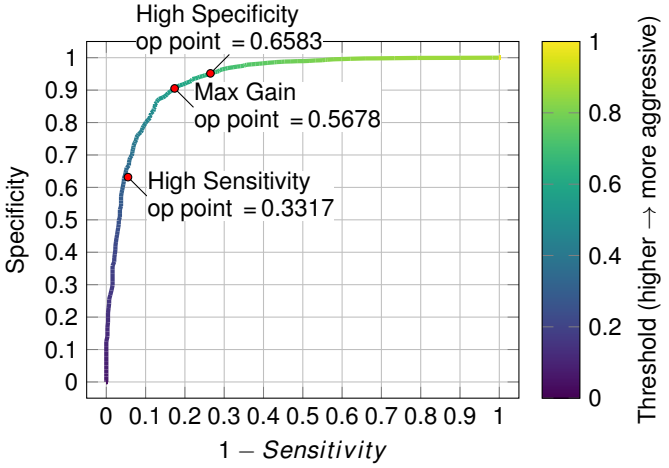

Fig. 4. Receiver Operating Characteristics of the ensemble model composed of ten neural networks with $500x500$ pixels input image resolution.

composed of the new ten neural networks. We can observe a drastic improvement, especially on the trade-off for a high sensitivity point, which now achieves over 60% accuracy. The high specificity point achieves 95.17% specificity, 73.49% sensitivity. The maximum gain point achieves 90.53% specificity, 82.85% sensitivity. The high sensitivity point achieves 64.46% specificity, 95.1% sensitivity. The AUC of this model's ROC is 0.93.

IV. Conclusions

In the healthcare context, sensitivity to screen a disease is the most critical metric of a detection mechanism. Detection of inexistent illnesses is not a serious problem when compared to an undetected illness that may require urgent treatment. However, if a predictive model aims to be efficient in reducing the load on ophthalmologists in public health, then the accuracy of the model represents how much it can save in a percentage. In this study, we have shown how a simple change to a neural network can impact its performance. Given the same sensitivity, higher resolution images provided an improvement of 6% in the accuracy of the detections. For the receiver operating characteristic curve, the area under the curve changed from 0.89 to 0.93.

We are currently working on improving the model. Our current idea is to use Inception v4, and obtain a specialized dataset to tune the model, as Krause et al. have done. Research on hyperparameters should also be considered, given the much higher area under the curve obtained in their research. Finally, the largest source of difference is the dataset available to us. Unfortunately, large datasets are not readily available. In the future, we will fuse smaller datasets to create a more extensive training dataset with varied camera configurations, which should help the network to generalize artifacts from different cameras better.

References

[1] A. D. Association et al. Standards of medical care in diabetes—2010. *Diabetes care*, 33(Supplement 1):S11–S61, 2010.

[2] J. Cuadros and G. Bresnick. Eyepacs: an adaptable telemedicine system for diabetic retinopathy screening. *Journal of diabetes science and technology*, 3(3):509–516, 2009.

[3] E. Decencière, X. Zhang, G. Cazuguel, B. Lay, B. Cochener, C. Trone, P. Gain, R. Ordonez, P. Massin, A. Erginay, et al. Feedback on a publicly distributed image database: the messidor database. *Image Analysis & Stereology*, 33(3):231–234, 2014.

[4] R. Gargeya and T. Leng. Automated identification of diabetic retinopathy using deep learning. *Ophthalmology*, 124(7):962–969, 2017.

[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] K. Hornik, M. Stinchcombe, and H. White. Multilayer feedforward networks are universal approximators. *Neural networks*, 2(5):359–366, 1989.

[7] J. Krause, V. Gulshan, E. Rahimy, P. Karth, K. Widner, G. S. Corrado, L. Peng, and D. R. Webster. Grader variability and the importance of reference standards for evaluating machine learning models for diabetic retinopathy. *Ophthalmology*, 125(8):1264–1272, 2018.

[8] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[9] Y. LeCun, Y. Bengio, et al. Convolutional networks for images, speech, and time series. *The handbook of brain theory and neural networks*, 3361(10):1995, 1995.

[10] H. Pratt, F. Coenen, D. M. Broadbent, S. P. Harding, and Y. Zheng. Convolutional neural networks for diabetic retinopathy. *Procedia Computer Science*, 90:200–205, 2016.

[11] S. Resnikoff, W. Felch, T.-M. Gauthier, and B. Spivey. The number of ophthalmologists in practice and training worldwide: a growing gap despite more than 200 000 practitioners. *British Journal of Ophthalmology*, 96(6):783–787, 2012.

[12] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015.

[13] J. Schneiders, G. H. Telo, L. G. Bottino, B. Pasinato, J. L. Neyeloff, and B. D. Schaan. Quality indicators in type 2 diabetes patient care: analysis per care-complexity level. *Diabetology & metabolic syndrome*, 11(1):34, 2019.

[14] I. M. Stratton, A. I. Adler, H. A. W. Neil, D. R. Matthews, S. E. Manley, C. A. Cull, D. Hadden, R. C. Turner, and R. R. Holman. Association of glycaemia with macrovascular and microvascular complications of type 2 diabetes (ukpds 35): prospective observational study. *Bmj*, 321(7258):405–412, 2000.

[15] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9, 2015.

[16] G. H. Telo, F. V. Cureau, M. S. de Souza, T. S. Andrade, F. Copês, and B. D. Schaan. Prevalence of diabetes in brazil over time: a systematic review with meta-analysis. *Diabetology & metabolic syndrome*, 8(1):65, 2016.

[17] A. Tufail, C. Rudisill, C. Egan, V. V. Kapetanakis, S. Salas-Vega, C. G. Owen, A. Lee, V. Louw, J. Anderson, G. Liew, et al. Automated diabetic retinopathy image assessment software: diagnostic accuracy and cost-effectiveness compared with human graders. *Ophthalmology*, 124(3):343–351, 2017.

[18] M. Voets, K. Møllersen, and L. A. Bongo. Replication study: Development and validation of deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *arXiv preprint arXiv:1803.04337*, 2018.

[19] T. Y. Wong and N. M. Bressler. Artificial intelligence with deep learning technology looks into diabetic retinopathy screening. *Jama*, 316(22):2366–2367, 2016.