The final publication is available at

http://dx.doi.org/10.1007/978-3-030-03493-1_24

Additional Information

Retinal Image Synthesis for Glaucoma Assessment using DCGAN and VAE Models

Andres Diaz-Pinto$^{(\boxtimes)1}$, Adrián Colomer^1, Valery Naranjo^1, Sandra Morales^1, Yanwu Xu^2, and Alejandro F Frangi^3

^{1} Instituto de Investigación e Innovación en Bioingeniería, I3B, Universitat Politècnica de València, Camino de Vera s/n, 46022 Valencia, Spain.
andiapin@upv.es
^{2} Guangzhou Shiyuan Electronics Co., Ltd. (CVTE), Guangzhou 510670, China.
^{3} CISTIB, University of Sheffield, Sheffield S1 3JD, UK.

Abstract. The performance of a Glaucoma assessment system is highly affected by the number of labelled images used during the training stage. However, labelled images are often scarce or costly to obtain. In this paper, we address the problem of synthesising retinal fundus images by training a Variational Autoencoder and an adversarial model on 2357 retinal images. The innovation of this approach is in synthesising retinal images without using previous vessel segmentation from a separate method, which makes this system completely independent. The obtained models are image synthesizers capable of generating any amount of cropped retinal images from a simple normal distribution. Furthermore, more images were used for training than any other work in the literature. Synthetic images were qualitatively evaluated by 10 clinical experts and their consistency were estimated by measuring the proportion of pixels corresponding to the anatomical structures around the optic disc. Moreover, we calculated the mean-squared error between the average 2D-histogram of synthetic and real images, obtaining a small difference of 3×10^{-4}. Further analysis of the latent space and cup size of the images was performed by measuring the Cup/Disc ratio of synthetic images using a state-of-the-art method. The results obtained from this analysis and the qualitative and quantitative evaluation demonstrate that the synthesised images are anatomically consistent and the system is a promising step towards a model capable of generating labelled images.

Keywords: Medical imaging, Retinal Image Synthesis, Fundus Images, DCGAN, VAE

1 Introduction

Glaucoma is an irreversible eye disease mainly characterised by optic nerve fibre loss. This loss is given by the increased intraocular pressure (IOP) and/or loss of blood flow to the optic nerve. In a fundus image, the optic nerve head or optic disc can be visually separated into two zones, a bright and central zone called optic cup and a peripheral part called neuro-retinal rim. See Fig. 1(a).
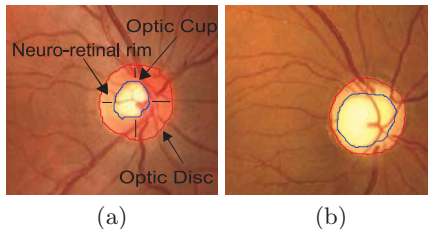
Fig. 1. Digital fundus images cropped around optic disc. (a) Main structures of a healthy optic disc and (b) glaucomatous optic disc.

While the optic disc (OD) and cup are present in all individuals, an abnormal size of the cup with respect to the optic disc is a characteristic of a glaucomatous eye, as it is shown in Fig. 1(b). A deep understanding of the anatomy of the optic disc is crucial for glaucoma understanding. For that reason, different approaches have been developed towards optic disc analysis for Glaucoma assessment using retinal images. For instance, in a state-of-the-art method developed by Chen et al. [1], they used cropped images to train and evaluate their system and obtaining an area under the ROC curve of 0.831 on a database of 650 images.

However, the amount of available images is a huge problem when trying to generalise. For that reason, retinal image synthesizer has been a focus of the scientific community. For instance, in [2] Fiorini et al. used a system that generated the retinal background and the fovea and another system to generate the optic disc by using a large dictionary of patches with no vessels that are later registered. After that, the authors developed a complementary work that is mainly focused on vessel generation [3]. Although their method allows the generation of high-quality and large resolution images, the process of concatenating the generation of the main parts of the images is a considerable complex computational algorithm that relies on how well the images are registered.

Another approach in retinal image synthesis is the one developed by Costa et al. [4]. In their work, they used a method trained on vessel networks and their corresponding retinal fundus images. In other words, they learn a transformation between the vessel trees and the retinal fundus. The main limitation of their method is the dependency of an independent algorithm to segment the vessels.

In another paper, Costa et al. present a method which improves their previous work. Instead of learning a transformation between the vessel trees and the corresponding retinal image, the authors used the original vessel trees to train an autoencoder. Then, the synthetic vessel trees are used as input to the retinal image synthesizer [5].

Although the latter system proposed by Costa et al. is a substantial improvement in their previous work, both methods are dependent on how well the independent method extracts the vessels. The quality of the segmented vessel tree will affect the synthetic vessel trees and then, the final retinal image.

In this paper, we are mainly focused on developing image synthesizers of retinal fundus images. In contrast to previous works, this novel approach does not

need the vessel masks and used more images in the training stage. We trained two well-known image generators: The Variational Autoencoder (VAE) [6] and the Deep Convolutional Generative Adversarial Networks (DCGAN) [7] using 2357 images cropped around the optic disc. After that, we used these models to generate synthetic retinal samples to finally evaluate them. Ten clinical experts checked the quality and global consistency of the generated images. Moreover, we compared the structural properties of synthetic and real images by measuring the proportions of the area occupied by the vessel network and optic disc. The consistency in colour terms between the synthetic and real images is also measured by extracting the 2D-histogram (or chromaticity diagram) and computing the mean-squared error.

2 Materials and Methods

2.1 Materials

A total of 2357 images from five public glaucoma-labelled databases: HRF [8] (45 images), Drishti-GS1 [9] (101 images), ORIGA-light [10] (650 images) RIM-ONE [11] (455 images) and sjchoi86-HRF [12] database (401 images) and a private database, ACRIMA (705 images), were used to train the generative models used in this work. All images were manually cropped around the optic disc by an expert, with the exception of RIM-ONE images that are already cropped.

For all the experiments carried out in this work, the open source Deep Learning library Keras [13] and NVIDIA Titan Xp GPU were used.

2.2 Variational Autoencoder

The Variational Autoencoder is composed by two neural networks: the approximate inference network (or encoder), that maps a training example to a latent (hidden) space, and the decoder network that maps from the latent space to a synthetic sample. In this work, we used the architecture proposed in [6], in which the prior over the latent space is a centred isotropic multivariate Gaussian, and the encoder and decoder are fully-connected neural networks with a single hidden layer.

During training or learning phase, the encoder obtains the latent variables z from the input data and the decoder draws those variables to generate a sample. After that, during the generation phase, VAE draws samples from the latent space that run through the decoder to finally obtain a synthetic sample. The VAE architecture can be seen from the Fig. 2(a).

2.3 Generative Adversarial Network

Generative Adversarial Networks, or GAN, are deep neural net architectures comprised of two nets. One is called the generator and the other (the adversary) is called the discriminator.

A class of CNN called Deep Convolutional Generative Adversarial Networks (DCGAN) that are based on the adversarial strategy was used for this work. This architecture was a major improvement on the first GAN, generating better quality images and more stability during the training stage. Following the guidelines to construct the generator and discriminator, described in the paper written by Radford et al. [7], we implemented and trained them on cropped retinal images using the original discriminator and generator cost functions.

In the same way, as in the VAE approach, synthetic image generation using the DCGAN mainly consists of two phases: a learning phase and generation phase. For the training phase, the generator draws samples from an N-dimension normal distribution that run through the generator to obtain a synthetic sample and the discriminator attempts to distinguish between images drawn from the generator and images from the training set. A figure of a DCGAN architecture can be seen from Fig. 2(b).
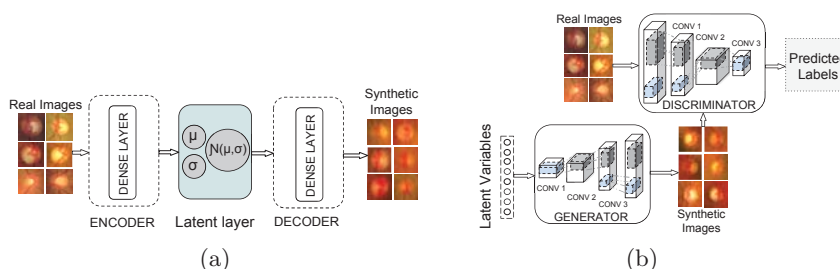

Fig. 2. Schema of the VAE and DCGAN architecture. (a) VAE architecture and (b) DCGAN architecture.

3   Experimental Evaluation

Although a great effort to develop objective metrics that correlate with perceived quality measurement has been made in recent years, it is still a challenging task. In the case of quality evaluation of synthetic images, it should be specific for each application [14]. For that reason, we created a database composed of 200 images: 100 synthetic images and 100 real images (randomly selected from the training set) to perform a qualitative and quantitative evaluation of our methods. This database was analysed by clinical experts with the aim of evaluating the anatomical consistency and plausibility of the synthetic images. The percentage of pixels that composed the vessels and the percentage of pixels that composed the optic disc were also compared between the synthetic and real images. To obtain these percentages, optic disc masks were manually segmented by clinical experts and the vessel masks were automatically segmented using the method proposed in [15]. Moreover, an averaged chromaticity diagram per class (real and synthetic) was computed with the aim of evaluating the colour properties of the images. The mean-squared error between the averaged histograms and the individual chromaticity diagram of each sample was measured.

4 Results and Discussion

We trained the VAE and DCGAN architectures on cropped retinal images from six different databases without using data augmentation. In order to keep a trade-off between performance and system complexity, the images were automatically re-scaled to the following resolutions: 28×28 pix, 56×56 pix, 112×112 pix and 224×224 pix. For each image size, we tested a range of N-dimensional latent spaces from 32 to 100 latent variables. Each latent space was explored in order to check that the systems do not memorise the training database and, at the same time, it generates plausible retinal images. To do that, we used spherical interpolation to evaluate intermediate latent representation points [16].

For training the VAE model, we ran several tests and found out that the best results are obtained when using a 100-dimension latent space and image resolution of 28×28 and 56×56 pix. Running for 500 epochs and a small batch size of 64, we obtained the synthetic images presented in Fig. 3
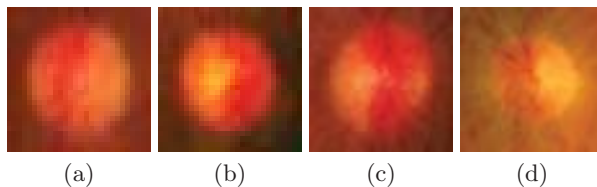

(a) (b) (c) (d)

Fig. 3. Examples of images synthesised by the VAE architecture. (a-b) Images of 28×28 pix and (c-d) images of 56×56 pix.

Although the texture of the synthetic images obtained from VAE is similar to the real images, they are blurry and do not have the expected features in a fundus image. For that reason, we only trained on the resolution 28×28 and 56×56 pix.

Regarding the DCGAN architecture, we found that realistic images were obtained when using an image size of 224×224 pix, a small batch size of 32 and 35000 steps. Examples of them are shown in Fig. 4(d-f).
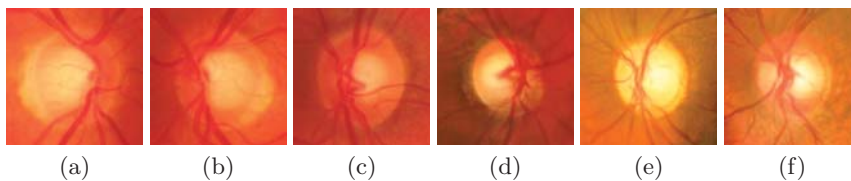

(a) (b) (c) (d) (e) (f)

Fig. 4. Examples of real images and examples of synthetic images generated by the DCGAN architecture. (a - c) Real images and (d - f) Synthetic images.

The main advantage of using the DCGAN architecture is that synthetic images are sharper than the ones synthesised by the VAE approach. We can see from Fig. 4(d-f) well-defined optic disc shapes, how blood vessels converge into

the optic disc and the right and left eye symmetry. For this reason, we continued with the evaluation of only the images synthesised by the DCGAN architecture.

Qualitative evaluation of the database described in Section 3 (100 synthetic images and 100 real images) was carried out by ten experts with 3 to 10 years of experience. For each expert, we calculated the Cohen's kappa coefficient using the ground-truth labels (Fake - Real) and the labels given by each expert. The Cohen's kappa coefficient ranges from -1 to $+1$, where 0 represents the amount of agreement that can be expected from random chance, and 1 represents a perfect agreement between the ground-truth and the expert. The obtained results are presented in Fig. 5.
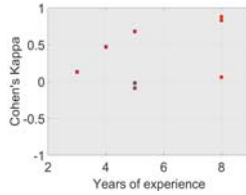

Fig. 5. Qualitative evaluation using Cohen's Kappa coefficient and years of experience

It can be seen from Fig. 5 that although the Cohen's Kappa coefficient is high for two experts with high expertise, most of them were fooled when evaluating synthetic images.

Quantitative evaluation was carried out by measuring the average proportion of pixels belonging to the vessel and optic disc structures. Table 1 shows the obtained results.

Table 1. Mean and standard deviation of pixel proportion occupied by the vessels, optic disc and background on the evaluation database.

	Synthetic Images	Real Images
Vessel proportion	0.1431 ± 0.0306	0.1519 ± 0.0306
Optic Disc proportion	0.1776 ± 0.0339	0.2456 ± 0.0722
Background	0.6792 ± 0.0428	0.6025 ± 0.0795

It is possible to observe from Table 1 that the mean proportions between synthetic and real images are very similar. The small difference between the mean proportion of the synthetic and real optic discs depends on the normal variation of the optic disc size among real fundus images.

To support the quantitative evaluation and to analyse the similarity between synthetic and real images, we also obtained the average 2D-histogram of real and synthetic images. These 2D-histograms were constructed using the Red and Green channels normalized by the luminance (See Fig. 6).

Moreover, we calculated the mean-squared error between the average 2D-histograms and the chromaticity diagram of each of the 200 images of the database. The obtained results are presented in Table 2
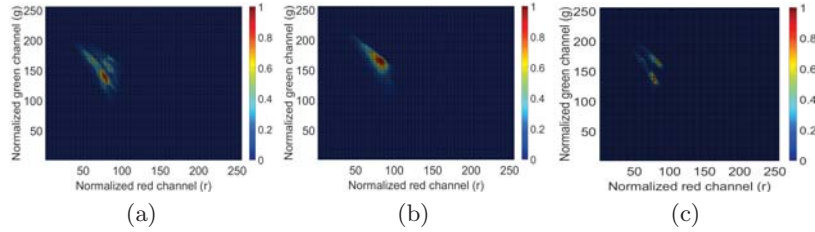
Fig. 6. Average 2D-histograms of the synthetic and real images. (a) Average 2D-histogram of real images, (b) Average 2D-histogram of synthetic images and (c) Mean-squared error between synthetic and real 2D-histogram.

Table 2. Average and standard deviation of the mean-squared error between the average 2D-histograms and all images.

Average 2D-histogram	Real Images	Synthetic Images
Real	$0.0028 \pm 3.25 \times 10^{-4}$	$0.0036 \pm 5.43 \times 10^{-4}$
Synthetic	$\mathbf{0.0031 \pm 4.61 \times 10^{-4}}$	$0.0022 \pm 5.62 \times 10^{-4}$

The obtained results of this evaluation show that the mean-squared error between synthetic and real images is smaller than the resulting standard deviation among real images (3.25×10^{-4}).

An additional experiment to further analyse the latent space and cup size of the images was performed. We automatically measured the Cup/Disc ratio (CDR) to 1500 synthetic images using the method proposed by Fu et al. [17]. Based on the CDR value, we obtained 743 glaucomatous images when setting the CDR threshold to 0.6 and 344 glaucomatous images when setting the CDR threshold to 0.7.

5   Conclusion

In this paper, two generative models based on the VAE and DCGAN architecture were trained on cropped retinal images from one private and five public databases (2357 retinal images). In contrast to previous approaches that are based on the vessel masks to train their system, the models presented here do not need the vessel masks to synthesise images. Using the DCGAN model, high plausible cropped retinal images were generated and evaluated by clinical experts. Results from this evaluation prove that this initial system is a promising step towards a model capable of generating labelled cropped images.

Acknowledgments

References

1. Chen, X., Xu, Y., Yan, S., Wong, D.W.K., Wong, T.Y., Liu, J.: Automatic Feature Learning for Glaucoma Detection Based on Deep Learning. In: Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015, Cham, Springer International Publishing (2015) 669–677
2. Fiorini, S., Biasi, M.D., Ballerini, L., Trucco, E., Ruggeri, A.: Automatic Generation of Synthetic Retinal Fundus Images. In: Smart Tools and Apps for Graphics - Eurographics Italian Chapter Conference, The Eurographics Association (2014)
3. Bonaldi, L., Menti, E., Ballerini, L., Ruggeri, A., Trucco, E.: Automatic Generation of Synthetic Retinal Fundus Images: Vascular Network. Procedia Computer Science **90**(Supplement C) (2016) 54 – 60
4. Costa, P., Galdran, A., Meyer, M.I., Niemeijer, M., Abràmoff, M., Mendona, A.M., Campilho, A.: End-to-end Adversarial Retinal Image Synthesis. IEEE Transactions on Medical Imaging **PP**(99) (2017) 1–1
5. Costa, P., Galdran, A., Inês Meyer, M., Abràmoff, M.D., Niemeijer, M., Mendonça, A.M., Campilho, A.: Towards Adversarial Retinal Image Synthesis. arXiv: 1701.08974 (January 2017)
6. Kingma, D.P., Welling, M.: Auto-Encoding Variational Bayes. arXiv: 1312.6114 (December 2013)
7. Radford, A., Metz, L., Chintala, S.: Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks. arXiv: 1511.06434 (November 2015)
8. Köhler, T., Budai, A., Kraus, M.F., Odstrčilik, J., Michelson, G., Hornegger., J.: Automatic no-reference quality assessment for retinal fundus images using vessel segmentation. In: Proceedings of the 26th IEEE International Symposium on Computer-Based Medical Systems. (2013) 95–100
9. Sivaswamy, J., Krishnadas, S., Joshi, G.D., Jain, M., Ujjwal, A., S.T.: Drishti-GS: Retinal image dataset for optic nerve head (ONH) segmentation. In: 2014 IEEE 11th International Symposium on Biomedical Imaging (ISBI). (2014) 53–56.
10. Zhang, Z., Yin, F.S., Liu, J., Wong, W.K., Tan, N.M., Lee, B.H., Cheng, J., Wong, T.Y.: ORIGA-light: An online retinal fundus image database for glaucoma analysis and research. In: 2010 Annual International Conference of the IEEE Engineering in Medicine and Biology. (Aug 2010) 3065–3068
11. Medina-Mesa, E., Gonzalez-Hernandez, M., Sigut, J., Fumero-Batista, F., Pena-Betancor, C., Alayon, S., de la Rosa., M.G.: Estimating the amount of hemoglobin in the neuroretinal rim using color images and OCT. Current Eye Research. **41**(6) (2015) 798–805
12. sjchoi86: sjchoi86-HRF Database. `https://github.com/sjchoi86/retina_dataset/tree/master/dataset` (2017) Accessed: 2017-07-02.
13. Chollet, F., et al.: Keras. `https://github.com/fchollet/keras` (2015) Accessed: 2017-05-21.
14. Theis, L., van den Oord, A., Bethge, M.: A note on the evaluation of generative models. In: International Conference on Learning Representations. (Apr 2016)
15. Morales, S., Naranjo, V., Navea, A., Alcañiz, M.: Computer-Aided Diagnosis Software for Hypertensive Risk Determination Through Fundus Image Processing. IEEE Journal of Biomedical and Health Informatics **18**(6) (Nov 2014) 1757–1763
16. White, T.: Sampling Generative Networks. arXiv: 1609.04468 (September 2016)
17. Fu, H., Cheng, J., Xu, Y., Wong, D.W.K., Liu, J., Cao, X.: Joint Optic Disc and Cup Segmentation Based on Multi-label Deep Network and Polar Transformation. IEEE Transactions on Medical Imaging **PP**(99) (Jan 2018) 1–1