Retinal vessel Deep Learning segmentation comparison

Daniela Herrera-Montes de Oca [1], Gilberto Ochoa-Ruiz[1], Miguel González-Mendoza[1], Christian Mata[2]

[2]Tecnologico de Monterrey, School of Engineering and Sciences, Mexico
[2]Universitat Politècnica de Catalunya, Spain

Abstract

The retinal vessel network studied through fundus images contributes to the diagnosis of multiple diseases not only found in the eye. The segmentation of this system may help the specialized task of analyzing these images by assisting the quantification of morphological characteristics. Due to its relevance, several Deep Learning-based architectures have been tested for addressing this problem automatically. However, the impact of the loss function selection on the segmentation of the intricate retinal blood vessel system hasn't been evaluated. In this work, we present the comparison of the loss functions Binary Cross Entropy, Dice, Tversky, and Combo loss using the SA-UNet with the DRIVE dataset. Their performance is assessed using four metrics: AUC, mean squared error, dice score, and Hausdorff distance. The second part, it is also compared four segmentation architectures: UNet, SA-UNet, Attention UNet, and Nested UNets.

1. Introduction

Utilizing eye fundus images is relevant for the identification of not only eye diseases but also systemic diseases since the retina is susceptible to changes in the circulation of the brain [1]. The study of retinal vessel structure through non-invasive techniques, assists the identification of cardiovascular diseases, hypertension, stroke, and retinopathies [7]. Diabetic retinopathy is present in have 80% to 85% of the patients that have diabetes for more than 10 years [11] and the gold standard for detecting it is the fundus imaging [5].

The examination of the image and study of morphological changes in this structure is a specialized task. As a consequence of the process of projecting the 3-D semi-transparent retinal tissue into a 2-D imaging plane [1], the evaluation of the image faces challenges. Quantification of the structure analysis (symmetry, width, length), to understand the pathological changes produced, is made through image processing and segmentation of the vessels [2].

Segmentation of the retinal vascular network has been addressed using deep learning with several architectures. There are multiple based on the UNet structure: SA-UNet [2], Residual UNet, Recurrent UNet, Attention-UNet [8], Generative Adversarial Networks [12], between others. Attention modules have been added to the principal architectures since this mechanism tells where to focus [14], which may help in the segmentation of intricate patterns such as the vessel network.

Besides the variety of architectures that exist for segmentation, the learning algorithm is instigated by the loss function, which should be selected depending on the objective [3]. For evaluating segmentation results there also exist multiple metrics which need to be selected considering the purpose and sensitivity of each one. The impact in the segmentation and evaluation of the quality by choosing loss functions and metrics is a challenge in deep learning segmentation.

Although comparisons had been made between the performance of architectures in retinal vessel segmentation [4, 12], the effect of loss function selection and metric hasn't been reported to our knowledge. In this study, we compare the loss function using different metrics using the SA-UNet which was designed for this task and has as a base the UNet. Also, some of the most relevant architectures are compared.

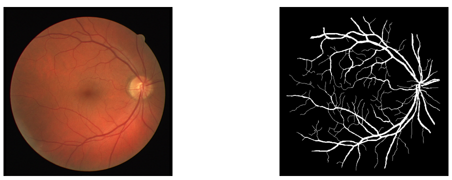
The paper contains the following order of sections. The section 2 contains the description of the dataset, loss functions, metrics, and deep learning structures used for each experiment. The results are presented in section 3. The conclusion and future work are presented in the section 4.

2. Data and methods

The technical contribution of this work is the evaluation of the impact of four loss functions with four metrics on the retinal vessel segmentation using SA-UNet. Al so the comparison of four deep learning segmentation architectures using the same loss function.

2.1. Dataset

The images used in this work come from a retinal vessel segmentation dataset: Digital Retinal Images for Vessel Extraction (DRIVE). Contains 40 images, of which 7 are abnormal pathology cases. It was used the version with data augmentation that includes random rotation, Gaussian noise, color jittering, and flips (horizontal, vertical, and diagonal) from the SA-UNet paper [2]. This incremented the number of training images from 20 to 256 images. An example of the DRIVE dataset is displayed in figure 1, there is the fundus image and the binary ground truth with the labels of the blood vessels.



(a) Test image (b) Ground truth

Figure 1: Sample images for the DRIVE dataset

2.2. Segmentation metrics

For evaluating the quality of the segmentation, there are multiple metrics that can be used depending on the data and the segmentation task [13]. For the first part of loss function comparison, four metrics are used: the Dice coefficient, the area under the ROC curve (AUC), means square error, and Hausdorff distance (HD). The Dice coefficient computes the pair-wise overlap between the segmentation and ground truth: $DICE = \frac{2|\text{Segmentation} \cap \text{Ground Truth}|}{|\text{Segmentation}|+|\text{Ground Truth}|}$. The ROC curve is the plot of the true positive rate (TPR) and the false positive rate (FPR), the AUC was designed as a measure of accuracy. The mean square error (MSE) is a probabilistic measure. The HD is a spatial-based metric measured in voxel size and measures the distance between the ground truth and the segmentation.

2.3. Loss functions

The loss functions help us in the mathematical representation of our segmentation objectives in Deep Learning for making it accurate and faster [3]. Four loss functions were evaluated: Dice, Tversky, Binary Cross Entropy, and Combo loss. The Dice loss function is based on the dice coefficient, explained in the previous section, and minimizes the similarity between the ground truth and the segmentation. The Tversky loss is based on the Dice loss and achieves a better balance between precision and recall, emphasizing the false negatives [6]. The Binary Cross Entropy loss measures the dissimilarity between two probability distributions. The Combo loss is a weighted summation between the Dice loss and a variation of the cross-entropy; this brings the advantages from both losses [3].

2.4. Deep Learning architectures

For the first part of loss functions comparison with multiple metrics, we selected Spatial Attention UNet (SA-UNet) proposed by Guo et al. [2]. This adds a spatial attention module between the encoder and decoder. It also has dropout convolutional blocks. For the second part four deep learning models for comparison. Besides SA-UNet it was also selected Attention-UNet, Nested UNet, and the base architecture which is the UNet proposed by Ronneberger et al. [10]. Attention UNet was proposed by Oktay et al. [9], it adds an Attention Gate between the union of the skip connection and the decoder. The UNet++ is a nested UNet architecture where the encoder and decoder are connected through a series of nested dense convolutional blocks [15]. For the experiments 100, epochs with early stopping were used.

3. Results

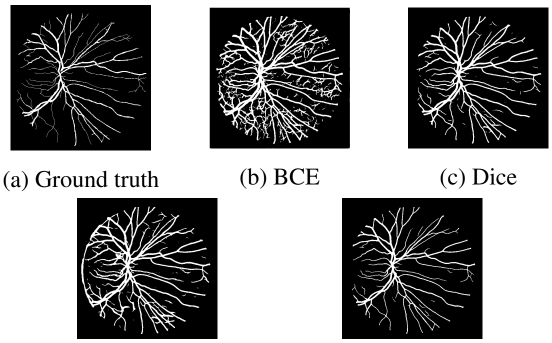
The first part consists of the comparison of the SA-UNet using four loss functions and four metrics. The summary of the experiments can be found on the table 1, it has the average metric for each experiment. The arrow $\uparrow$ represents that the metric is better when is bigger, while the arrow $\downarrow$ means that is better when is close to 0. We can see that although they were trained for the same number of epochs, the metrics are different depending on the loss function.
If we would base the performance on only one metric, we could ignore the overall performance. For example, in the case of AUC, the loss function that had the best performance was the Tversky loss however, looking at the other metrics the Combo loss had the best one with respect to the others.

Table 1: Results of SA-UNet using multiple loss functions and metrics (average).

Loss functions	**AUC** $\uparrow$	**MSE** $\downarrow$	**Hausdorff distance** $\downarrow$	**Dice score** $\uparrow$
Dice	0.9431	0.0565	9.3039	0.7327
Tversky	0.9442	0.0692	10.1913	0.6951
Binary Cross Entropy	0.9197	0.1339	21.0970	0.5470
Combo	0.9335	0.0355	7.9129	0.8059

The visual results can be seen in figure 2, the same image is compared against the four loss functions. The image 2a shows the ground truth and the rest of the images are the segmentation results with a different loss function.

For the second part, four segmentation architectures were evaluated (UNet, SA-UNet, Attention UNet, and



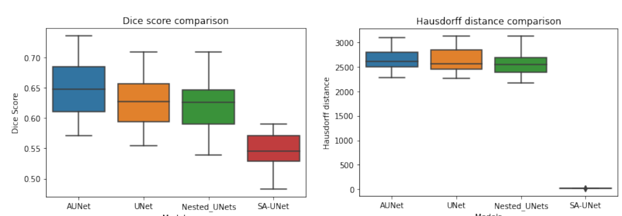
(a) Ground truth (b) BCE (c) Dice

(d) Tversky (e) Combo

Figure 2: Comparison of loss functions performance using SA-UNet.

Nested UNets). The results were evaluated using the Hausdorff distance ↓ and the Dice score ↑. The summary of the results can be seen in the boxplots from figure 3. Regarding the Attention UNet had the best average and distribution for the dice score results. The results for the Hausdorff distance are more compact meaning that they are less variable. The SA-UNet had the smaller average from the four architectures, which means that the distance between the segmentation and the ground truth was smaller and had better performance. The other three architectures have similar results.

For determining if the results were statistically different, first a Friedman test was made between the segmentation results of the four architectures. This showed that they were. For determining between which ones, a Wilcoxon test was made between each architecture. It showed that they were all significantly different from each other. The results obtained for these tests are shown in table 2.



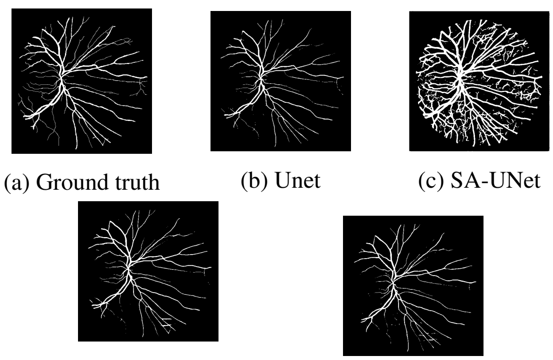
(a) Dice score (b) Hausdorff distance

Figure 3: Results from Deep Leaning based architectures comparison

The visual segmentation results for each Deep Learning-based architecture are shown in figure 4. The ground truth is compared against the segmentation prediction. The same case is used for comparison. For these experiments it can be seen that in all models fine blood vessels are still missing, this could have been improved by adding more epochs.

Table 2: Results of Friedman and Wilcoxon statistical tests for the segmentation results against each architecture.

Test	Statistic	pvalue
Friedman	32.5	8.764248e-08
AUNet vs UNet	0.0	8.857458e-05
AUNet vs SA-UNet	1.0	1.033465e-04
AUNet vs Nested UNets	0.0	8.857458e-05
UNet vs Nested UNets	45.0	2.509351e-02
UNet vs SA-UNet	4.0	1.628558e-04
Nested UNets vs SA-UNet	9.0	3.384547e-04



(a) Ground truth (b) Unet (c) SA-UNet

(d) Attention UNet (e) Nested UNets

Figure 4: Comparison of architectures segmentation performance.

4. Conclusion and future work

We observed that there was a significant impact in the selection of the loss function, this was reflected in the average of each metric. Considering the overall performance of the metrics, the best loss function was the Combo. The conclusion could have been different if it had only been determined by one metric. Therefore the combination of the loss function and metric needs to be considered. For the Deep Leaning based architecture comparison, although they all were UNet based they had different results which were also reflected in the statistical tests. For the Dice score, it was the Attention Unet that had better performance, however in the Hausdorff distance it was better than the SA-Unet, this shows the sensitivity of the metrics and how it should be globally evaluated.

For future work and a better understanding of the impact on retinal vessel segmentation, a comparison between the four loss functions and four architectures should be done. The use of other datasets could also bring more comprehensive studies.

References

[1] Michael D. Abramoff, Mona K. Garvin, and Milan Sonka. Retinal imaging and image analysis. *IEEE Reviews in Biomedical Engineering*, 3:169–208, 2010. 1

[2] Changlu Guo, Marton Szemenyei, Yugen Yi, Wenle Wang, Buer Chen, and Changqi Fan. Sa-unet: Spatial attention u-net for retinal vessel segmentation. *ArXiv*, abs/2004.03696, 2020. 1, 2

[3] Shruti Jadon. A survey of loss functions for semantic segmentation. In *2020 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB)*, page 1–7. IEEE, Oct 2020. 1, 2

[4] Aashis Khanal and Rolando Estrada. Dynamic deep networks for retinal vessel segmentation. *Frontiers in Computer Science*, 2:35, Aug 2020. 1

[5] Somya Kumari, Pradeep Venkatesh, Nikhil Tandon, Rohan Chawla, Brijesh Takkar, and Atul Kumar. Selfie fundus imaging for diabetic retinopathy screening. *Eye*, Oct 2021. 1

[6] Jun Ma, Jianan Chen, Matthew Ng, Rui Huang, Yu Li, Chen Li, Xiaoping Yang, and Anne L. Martel. Loss odyssey in medical image segmentation. *Medical Image Analysis*, 71:102035, Jul 2021. 2

[7] Maliheh Miri, Zahra Amini, Hossein Rabbani, and Raheleh Kafieh. A comprehensive study of retinal vessel classification methods in fundus images. *Journal of Medical Signals and Sensors*, 7(2):59–70, Jun 2017. 1

[8] Ozan Oktay, Jo Schlemper, Loïc Le Folgoc, Matthew C. H. Lee, Mattias P. Heinrich, Kazunari Misawa, Kensaku Mori, Steven G. McDonagh, Nils Y. Hammerla, Bernhard Kainz, Ben Glocker, and Daniel Rueckert. Attention u-net: Learning where to look for the pancreas. *CoRR*, abs/1804.03999, 2018. 1

[9] Ozan Oktay, Jo Schlemper, Loïc Le Folgoc, Matthew C. H. Lee, Mattias P. Heinrich, Kazunari Misawa, Kensaku Mori, Steven G. McDonagh, Nils Y. Hammerla, Bernhard Kainz, Ben Glocker, and Daniel Rueckert. Attention u-net: Learning where to look for the pancreas. *CoRR*, abs/1804.03999, 2018. 2

[10] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. *U-Net: Convolutional Networks for Biomedical Image Segmentation*, volume 9351, page 234–241. Springer International Publishing, 2015. 2

[11] Dilip Singh Sisodia, Shruti Nair, and Pooja Khobragade. Diabetic retinal fundus images: Preprocessing and feature extraction for early detection of diabetic retinopathy. *Biomedical and Pharmacology Journal*, 10(2):615–626, Jun 2017. 1

[12] Jaemin Son, Sang Jun Park, and Kyu-Hwan Jung. Towards accurate segmentation of retinal vessels and the optic disc in fundoscopic images with generative adversarial networks. *Journal of Digital Imaging*, 32(3):499–512, Jun 2019. 1

[13] Abdel Aziz Taha and Allan Hanbury. Metrics for evaluating 3d medical image segmentation: analysis, selection, and tool. *BMC Medical Imaging*, 15(1):29, Dec 2015. 2

[14] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. *CBAM: Convolutional Block Attention Module*, volume 11211, page 3–19. Springer International Publishing, 2018. 1

[15] Zongwei Zhou, Md Mahfuzur Rahman Siddiquee, Nima Tajbakhsh, and Jianming Liang. *UNet++: A Nested U-Net Architecture for Medical Image Segmentation*, volume 11045, page 3–11. Springer International Publishing, 2018. 2