2021 43rd Annual International Conference of the
IEEE Engineering in Medicine & Biology Society (EMBC)
Oct 31 - Nov 4, 2021. Virtual Conference

Segmentation in Diabetic Retinopathy using Deeply-Supervised Multiscalar Attention

Sanhita Basu[1] and Sushmita Mitra[1]

Abstract— **Diabetic Retinopathy (DR) is a progressive chronic eye disease that leads to irreversible blindness. Detection of DR at an early stage of the disease is crucial and requires proper detection of minute DR pathologies. A novel Deeply-Supervised Multiscale Attention U-Net (Mult-Attn-U-Net) is proposed for segmentation of different DR pathologies viz. Microaneurysms (MA), Hemorrhages (HE), Soft and Hard Exudates (SE and EX). A publicly available dataset (IDRiD) is used to evaluate the performance. Comparative study with four state-of-the-art models establishes its superiority. The best segmentation accuracy obtained by the model for MA, HE, SE are 0.65, 0.70, 0.72, respectively.**

I. INTRODUCTION

Long term diabetes causes Diabetic Retinopathy (DR), a progressive chronic eye disease, which leads to irreversible blindness. Although detection of DR at an early stage of the disease is crucial to prevent blindness, most of the patients become symptomatic only in the advanced stages of DR; such as non-proliferative DR (NPDR)[1] or proliferative DR (PDR)[2]. According to the "International Clinical Diabetic Retinopathy Disease Severity Scale" the severity of DR can be graded into five stages: normal, mild, moderate, severe and proliferative. Several pathologies related to DR are red lesions viz. Microaneurysms (MA), Hemorrhages (HE), and bright lesions viz. Soft and Hard Exudates (SE and EX), venous beading, neo-vascularization, etc. Information related to these pathologies are helpful in segregating DR images from normal images, as well as for grading the DR. Typically a DR detection system involves an opthalmologist manually detecting vascular abnormalities and structural changes of retina, from the color fundus images captured by fundus cameras. Due to the manual nature of DR screening methods, highly inconsistent results are often encountered from different readers. Therefore automated diagnosis of DR becomes necessary in the process of solving these problems.

State-of-the-art literature in this domain mostly deals with classification of DR stages. While earlier solutions were constrained to two class problems, involving DR and normal [1], the recent approaches [based on deep Convolutional Neural Networks (CNNs)] attempt to automatically grade DR into its different stages [2], [3], [4]. The challenges faced by researchers in implementing such deep network architectures include the need for considerable reduction in input image size during training; which, again, results in complete loss of most of the relevant minute pathological information necessary for grading.

As far as our knowledge goes, literature on segmentation of retinal lesions – particularly, the different DR pathologies – is scarce. One of the reasons for this is the lack of annotated datasets, mainly due to the difficulty in acquiring pixel-level annotation. Tan *et al.* [5] segmented EX, HE and MA automatically, using a single CNN, but obtained a very low sensitivity. Recently a Diabetic Retinopathy Image Dataset (IDRiD)[3] [6] has been released, in conjunction with a challenge (Diabetic Retinopathy: Segmentation and Grading Challenge) organized at the IEEE International Symposium on Biomedical Imaging (ISBI 2018). A total of 22 teams participated in the segmentation challenge, where the task was to segment the four pathology classes MA, HE, SE and EX. Most of the top performing methods used variants of the popular U-Net [7], which has been extensively employed in medical image segmentation in recent years. While for bigger lesions like EX and SE a moderate segmentation accuracy was achieved, the smaller lesions like MA and HE resulted in considerably poor segmentation. Incidentally, the highest scoring method achieved 0.5017 AUPR (Area Under the Precision and Recall curve) for the MA segmentation task.

Pathologies such as MA and HE are usually scarce, and their dimensions are negligible as compared to the entire image. Standard segmentation architectures like U-Net and Fully Convolutional Networks [8] do not yield satisfactory results, mainly because of the progressive use of max pooling operation in the encoding path to achieve translational invariance over small spatial shifts in the input image. While extracting dense semantic features, they tend to lose the spatial information (boundary details) of these tiny pathologies; which is not beneficial for tasks where delineation of the boundary is vital. This necessitates the design of a specialized architecture capable of passing relevant spatial information of such pathologies (before each max pooling step) to the corresponding stage of the decoding path, for correctly delineating them at the output. In order to address the problem of segmenting such DR pathologies, an attention mechanism [9] is introduced at different levels. Here Attention Gates (AGs) are integrated with the U-Net architecture, to generate attention maps, before concatenating them to the decoder path. Besides, for an improved representation of intermediate features, a

This work was supported by the J. C. Bose National Fellowship, sanction no. JCB/2020/000033

[1]Sanhita Basu and Sushmita Mitra are with Machine Intelligence Unit, Indian Statistical Institute, Kolkata, India tania.sanhita@gmail.com, sushmita@isical.ac.in

[1]occurs when the blood vessels leak the blood in the retina

[2]causes blindness, and is the next stage to NPDR.

[3]https://idrid.grand-challenge.org/Segmentation/

multiscaled input image pyramid is incorporated for a better precision-recall balance. A deep-supervision based layer-wise training scheme, to force the intermediate layers to be semantically discriminative at every scale, is also introduced. Our main contributions are highlighted as follows.

- We incorporated multi resolution inputs in the U-Net architecture. It provides a better representation of fine-grain contextual information and eventually helped in segmenting pathologies with variable sizes and shapes.
- We proposed a deep supervision based training to provide direct supervision to the hidden layers and propagate it to lower layers, instead of just doing it at the output layer. This allows each layer to use a less "diluted" gradient to learn.
- Gradients originating from background regions are down-weighted during the backward pass of the deep supervision training. This allows model parameters in prior layers to be updated based on spatial regions that are relevant to a given task.

The rest of the paper is organized as follows. Sec. II provides details about the dataset and the proposed deep multiscale attention-based architecture for the segmentation of DR pathologies. Sec. III presents the experimental setup, results, and discussion. The article is concluded in Sec. IV.

II. Materials and Methods

This section presents the details of the dataset used, followed by a description of the proposed deeply-supervised Multiscale Attention U-Net (Mult-Attn-U-Net).

A. Dataset

Training and testing of the Mult-Attn-U-Net was performed on the IDRiD database, encompassing 54 training and 27 testing color Fundus images annotated with the ground truth mask for each of the four pathologies MA, HE, EX and SE. The images were of 4288×2848 resolution. However, not all pathologies were present in all images. In particular, 54 images had MA and EX, 53 contained HE, and only 26 of them exhibited SE. While MA was observed to be distributed over the entire retinal region in small portions, the number of images containing SE was comparatively fewer. This made the task of segmentation even more difficult. Fig. 2 illustrates samples of the four kinds of lesions, appearing in the database. The 54 training images were randomly further split into 80% for training and 20% for validation, while ensuring that the validation set contained images representative of all pathologies (for their proper evaluation). As each image was of very high resolution, with the number of images being rather small to train the network, we extracted overlapping patches (of resolution 400×400) from each image with a stride of 200 to build the training and validation datasets. Non-overlapping patches of 1600×1600 were used for testing.

B. Mult-Attn-U-Net for segmentation

Typically the standard U-Net architecture gradually down-samples feature maps to capture a sufficiently large receptive field of the input image, such that features at the coarser spatial grid level serve to represent the location. Often this leads to difficulty in reducing false-positive predictions for smaller regions like MA and HE, which exhibit larger shape variability. In order to alleviate this problem, we integrate Attention Gates (AGs) into the U-Net framework at different scales. The AG acts as a filter to the response from the encoding stage of the U-Net, and is concatenated to the decoding stage via skip connection. The gates are scalar matrices, with values lying in the range $[0, 1]$, and represent probabilities of salient object(s) required for a task. The saliency maps, when multiplied by the input response before concatenation with the decoding path, help in attenuating feature responses irrelevant to the task. As the maps get generated during testing, these can be learned end-to-end via deep learning framework.

Fig. 1 depicts the schematic of the proposed Multiscale Attention U-Net (Mult-Attn-U-Net), with Attention Gates at multiple levels incorporated into the U-Net architecture. Multi resolution inputs are incorporated in the U-Net architecture for capturing better fine-grain contextual information, which eventually helped in segmenting pathologies with variable size and shapes. The novel "deeply-supervised" learning enforces the intermediate layers to be semantically discriminative at every scale. Multiple losses are computed, based on the outputs from each of the decoding blocks, by comparing them with the corresponding segmentation maps of the same resolution. Since max-pooling is applied after each encoding stage, the input image gets downsampled and concatenated with the feature maps. This improves the segmentation accuracy, as small features do not get lost in cascading max-pooling. It was found that deep supervision is advantageous in enhancing the segmentation accuracy of smaller pathologies like the MA and HE.

The Attention Gates, which help preserve feature responses relevant only to the task, are highlighted in Fig. 1. An input features x (represented by the dotted input to AG) is scaled with attention coefficient α in the Attention gates (AG). Relevant spatial regions are passed after analysing the activation of the gating signal g (denoted by the blue input to AG), which contains rich spatial and contextual information provided by the input signal x. The decoder path is divided into three levels, viz. "lower" (l), "middle" (m), and "upper" (u) layers, in order to incorporate deep supervision. Along with the predicted segmentation maps generated from the final upper layer, there arrive two smaller resolution prediction maps from the middle and lower layers, respectively. Independent loss functions are computed at each layer, *viz.* $\mathcal{L}_l$, $\mathcal{L}_m$ and $\mathcal{L}_u$, by comparing the predictions with the corresponding resized ground truths. This results in a more efficient gradient back propagation, along with accurate detection of smaller irregular pathologies. Let W be the weight of the main network, and w^l, w^m and w^u be the weights of the three classifiers used in the lower, middle and upper level outputs respectively. The final loss used for the deep supervision is represented as a weighted sum of

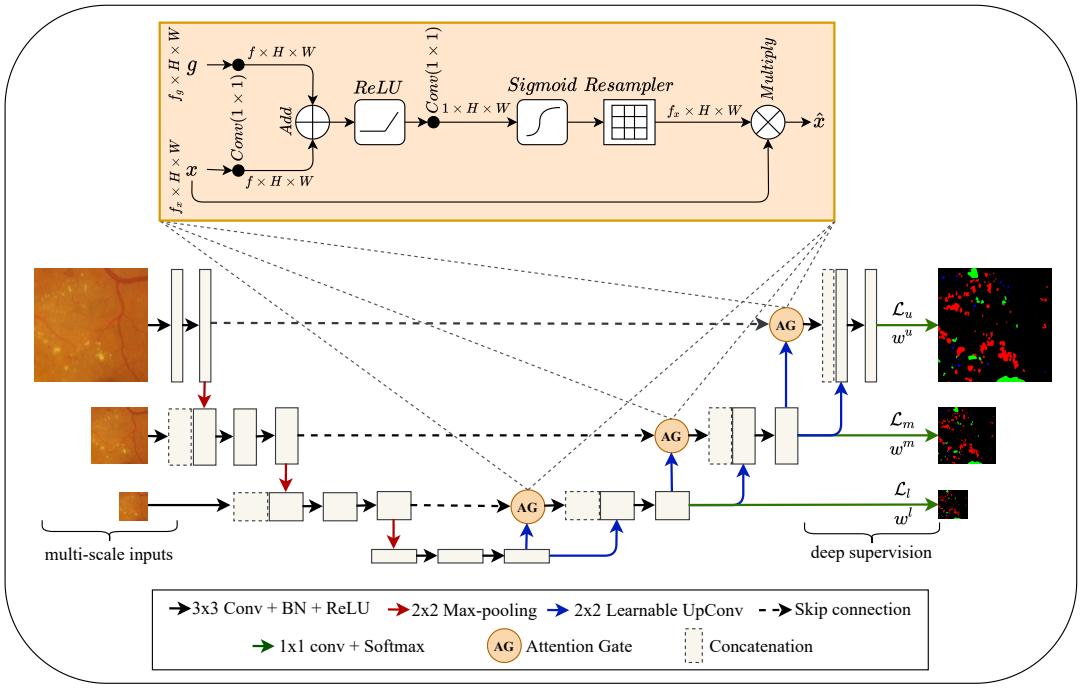

Fig. 1. Multiscale Attention U-Net (Mult-Attn-U-Net) with deep supervision.

individual losses, and expressed as

$$\mathcal{L}(\chi; W, w^l, w^m, w^u) = \sum_{c\in\{l,m,u\}} \alpha_c \mathcal{L}_c(\chi; W, w^c) + \lambda(\psi(W) + \sum_{c\in\{l,m,u\}} \psi(w^c). \quad (1)$$

Here α_l, α_m, α_u are the weights for the associated loss, χ represents the training samples, $\psi()$ is the L_2 regularization term, and hyperparameter λ acts as a trade-off coefficient.

Evaluation of segmentation is made in terms of the multi-class extension of Generalized Dice loss (GDL) [10], where the weight of each class (ω) is inversely proportional to the square of label frequencies, that efficiently handles the associated class imbalance problem. The GDL is defined as

$$\mathcal{L}_c = \frac{\sum_{s=1}^{S} \omega_s \sum_{i=1}^{N} y_{s,i} p_{s,i}}{\sum_{s=1}^{S} \omega_s \sum_{i=1}^{N} y_{s,i} + p_{s,i} + \epsilon}, \quad (2)$$

where $\omega_s = \frac{1}{(\sum_{i=1}^{N} y_{s,i})^2}$, with p and y denoting the predicted probability map and ground truth values, respectively. Here ϵ ensures the loss function stability, S represents the classes viz. $\{Background, MA, HE, EX, SE\}$, and N denotes the total number of pixels in the image.

III. Experimental Setup and Results

The proposed deeply-supervised Mult-Attn-U-Net was developed using TensorFlow, with a wrapper library Keras in Python. The experiments were performed on a Dell Precision 7810 Tower with 2x Intel Xeon E5-2600 v3, totalling 12 cores, 128GB RAM, and NVIDIA Quadro K6000 GPU with 12GB VRAM. Adam optimization algorithm was employed for hyperparameter optimization for training, with an initial learning rate 10^{-3}, and decayed according to cosine annealing. Real time data augmentation was used in terms of random rotation, scaling, and mirroring. Area Under the Curve of Precision-Recall (AUCPR) [6] is used to evaluate the performance of the model. We have compared our results with the top scoring methods from IDRiD challenge on the test data set as reported on the leaderboard[4] as summarized in Table III.

As observed from the Table III, the proposed method achieved the best scores for segmentation of MA, HE and SE. A huge gain (around 15%) is achieved in case of MA segmentation compared with the second best performing model (IFLYTEK). Detection of DR in the preliminary stage actually depends on correct detection of MA, which is the earliest visible sign of retinal damage. In case of HE and SE significantly improved segmentation results were observed. It should be noted that we are using smaller size patches (400×400) as compared with the VRT (1200×1200) (the second best performing method). Fig. 2 displays qualitative segmentation results for two sample images from the dataset.

We also performed three ablation studies to report the effect of training patch size – two patch sizes *viz.* 256×256 and 512×512 were considered, training without deep supervision and with increased depth reported in Table II. As observed from Table II the proposed model with increased depth attained the best score for the EX but performed poorly for MA and HE. For bigger patch sizes we did not observe any significant improvements in the results and deep supervision actually enhances the model performance.

[4]https://idrid.grand-challenge.org/Leaderboard/

TABLE I

COMPARISONS OF THE PROPOSED METHOD AND THE TOP 4 TEAMS FOR SEGMENTATION OF DIFFERENT LESION (MA, HE, SE AND EX) ON THE TESTING DATASET.

Method	Lesion				Approach	Training patch size
	MA	HE	SE	EX		
VRT	0.4951	0.6804	0.6995	0.7127	U-Net	1200×1200
iFLYTEK	0.5017	0.5588	0.6588	0.8741	Cascaded CNN with Ensemble	320×320
PATech	0.4740	0.6490	-	**0.8850**	DenseNet+U-Net	256×256
SDNU	0.4111	0.4572	0.5374	0.5018	Mask R-CNN	3584×2380
Proposed	**0.6500**	**0.6984**	**0.7201**	0.8545	Attention U-Net with deep supervision	400×400

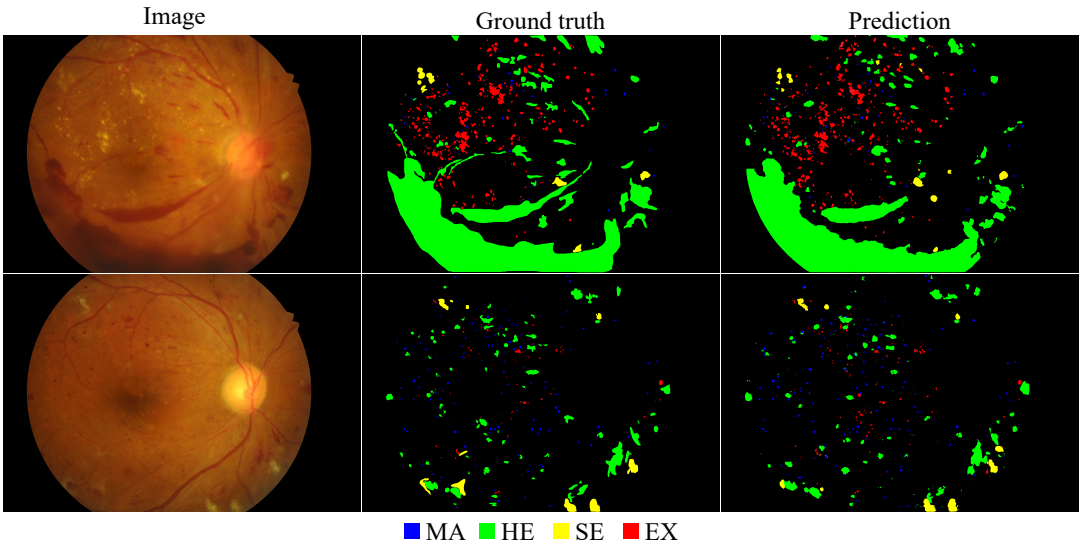

Fig. 2. Qualitative segmentation results of two sample images.

TABLE II

ABLATION STUDIES TO STUDY THE EFFECT OF PATCH SIZE, DEEP SUPERVISED TRAINING AND MODEL DEPTH.

Networks \ Pathologies	MA	HE	SE	EX
Trained on patches of size 256×256	0.46	0.65	0.38	0.82
Trained on patches of size 512×512	0.51	0.67	0.45	0.68
Mult-Attn-U-Net without deep supervision	0.44	0.62	0.59	0.76
Mult-Attn-U-Net with increasd depth	0.43	0.52	0.71	**0.89**

IV. CONCLUSION

This paper presented a novel CNN model called deeply-supervised multiscale attention U-Net (Mult-Attn-U-Net) for segmentation of four DR pathologies *viz.* MA, HE, SE, and EX from fundus images. Novel concepts such as deeply supervised training and multiscaller attention based networks were used for this purpose. An aggregated loss function was also proposed for the deeply supervised training which provides direct supervision to the hidden layers. We compared the proposed model with the four state-of-the-art models based on a publicly available dataset (IDRiD). The proposed model achieved the best segmentation accuracy for small pathologies such as MA, HE, and SE. A huge gain (around 15%) is achieved in case of MA segmentation compared with the second best performing model (IFLYTEK).

REFERENCES

[1] G. G. Gardner, D. Keating, T. H. Williamson, *et al.*, "Automatic detection of diabetic retinopathy using an artificial neural network: a screening tool.," *British Journal of Ophthalmology*, vol. 80, no. 11, pp. 940–944, 1996.

[2] Y. Yang, T. Li, W. Li, H. Wu, *et al.*, "Lesion detection and grading of diabetic retinopathy via two-stages deep convolutional neural networks," in *Medical Image Computing and Computer Assisted Intervention - MICCAI 2017*, pp. 533–540, Springer International Publishing, 2017.

[3] B. Harangi, J. Toth, A. Baran, and A. Hajdu, "Automatic screening of fundus images using a combination of convolutional neural network and hand-crafted features," in *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pp. 2699–2702, 2019.

[4] J. de La Torre, A. Valls, and D. Puig, "A deep learning interpretable classifier for diabetic retinopathy disease grading," *Neurocomputing*, vol. 396, pp. 465–476, 2020.

[5] J. H. Tan, H. Fujita, S. Sivaprasad, *et al.*, "Automated segmentation of exudates, haemorrhages, microaneurysms using single convolutional neural network," *Information Sciences*, vol. 420, pp. 66–76, 2017.

[6] P. Porwal, S. Pachade, M. Kokare, *et al.*, "IDRiD: Diabetic retinopathy–segmentation and grading challenge," *Medical Image Analysis*, vol. 59, p. 101561, 2020.

[7] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 234–241, Springer, 2015.

[8] E. Shelhamer, J. Long, and T. Darrell, "Fully convolutional networks for semantic segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 4, pp. 640–651, 2017.

[9] O. Oktay, J. Schlemper, L. L. Folgoc, *et al.*, "Attention U-Net: Learning where to look for the pancreas," *arXiv preprint arXiv:1804.03999*, 2018.

[10] S. Banerjee and S. Mitra, "Novel volumetric sub-region segmentation in brain tumors," *Frontiers in Computational Neuroscience*, vol. 14, p. 3, 2020.