Enhancing Retinal Vessel Segmentation with a Simplified nnU-Net Model and CBAM

Haotian Wu

Department of Computer Science and Engineering, New York University, New York, USA

1809040304@stu.hrbust.edu.cn

Abstract. To better diagnose and monitor ocular diseases, segmentation of blood vessels on the retina is crucial, but manual annotation is labor-intensive and error-prone. To deal with these problems, many automatic segmentation algorithms have already been developed, with deep learning techniques, leading the advancements. This study aims to enhance retinal vessel segmentation by refining the no-new-U-Net (nnU-Net) model through two key modifications: the integration of the Convolutional Block Attention Module (CBAM) attention blocks and the replacement of the original Stage 1 nnU-Net with a simplified Residual Neural Network (ResNet)-like network. The incorporation of CBAM improves feature extraction by focusing on the most relevant features, while the ResNet-like architecture simplifies the network structure, making it more efficient. Experimental results on the Retina Blood Vessel Dataset reveal that the modified nnU-Net model significantly outperforms the original in terms of both Dice score and prediction time. Specifically, it achieves superior accuracy in identifying fine blood vessels and demonstrates reduced memory usage, showcasing its potential for delivering more precise and efficient retinal vessel segmentation.

Keywords: Retinal Vessel Segmentation, nnU-Net, Convolutional Block Attention Module, Deep Learning.

1. Introduction

Vessel segmentation in medical image analysis is crucial since it aids the diagnosis and identification of blood vessels and related diseases. Especially, retinal vessel segmentation enables fast detection, and monitoring of related diseases [1]. With accurate segmentation of the retina vessels, doctors can assess the extent of the disease, track its progression, and develop treatment plans, ultimately speeding up the process and lowering the risk of vision loss.

However, generating segmentation results by manual annotation usually involves many problems due to its time-consuming nature, inconsistency, and proneness to errors in most medical scenes. Then in the past decades, lots of automatic retinal vessel segmentation algorithms have been established and these algorithms applied a variety of techniques aimed at accuracy and efficiency. On the one hand, traditional vessel segmentation algorithms like Matched filtering enhance vessel areas using certain filters that can highlight tube structures [2]. Multi-scale approaches analyze images in different scales to detect vessels of various sizes [3]. Active contour models, or snakes, try to fit the boundaries of vessels by minimizing an energy function [4]. On the other hand, convolutional neural networks (CNNs) and other deep learning methods, are prominent due to their high accuracy and ability to extract complicated features from vision data. As one of the most famous neutral networks, U-Net architecture [5] is used in many segmentation tasks on biomedical images. It has demonstrated significant success in vessel segmentation. Lots of variants of U-Net, such as Residual Neural Network (ResUNet) [6] and Attention Unit [7], have been developed to achieve better performance.

In practice, traditional algorithms and deep learning algorithms have their advantages and disadvantages: The number of situations that traditional algorithms can solve is limited while deep learning algorithms rely on massive training data to work properly. Some algorithms combined both techniques and have shown some progressive results. For example, deep learning models can be augmented with conditional random fields to refine segmentation [8].

Derived from Unet architecture, The no-new-U-Net (nnU-Net) is a deep learning model designed for segmentation tasks on medical images. It has demonstrated outstanding performance across a

series of segmentation tasks on medical images by adapting its automatic configuration including preprocessing, data augmentation, and postprocessing to different situations [9].

This study's focus is to enhance the performance of nnU-Net for retinal vessel segmentation tasks while simplifying its network structure. Specifically, the Convolutional Block Attention Module (CBAM) block is incorporated into the network to improve its feature extraction capabilities [10]. Additionally, the model in stage 1 of nnU-Net is replaced with a simpler CNNs, which extracts a feature map from the entire image. This feature map then replaces the coarse segmentation result typically produced in stage 1 and is passed to stage 2, reducing the overall number of network parameters. Experimental results show that the modified model, with its attention mechanism and streamlined architecture, can more accurately identify tiny blood vessels and requires less memory compared to the original network.

2. Methodology

2.1. Dataset Description and Preprocessing

The dataset applied in the experiment is chosen from Kaggle and is called the Retina Blood Vessel Dataset [11]. It comprises a collection of high-resolution retinal fundus images with blood vessel annotations. All images in the dataset have a fixed size of 512 x 512. The annotations are in a format of binary mask in which vessel pixels are marked as 1 while other things are labeled as 0. The main task in this study is to identify all retina vessels from these images. One image and corresponding mask are shown in Fig. 1.



Figure 1. Sample image and mask

2.2. Proposed Approach

Two major improvements made by this proposed approach are feature fusion and attention mechanism. In stage 1 of the nnU-Net, a simplified ResNet-like network replaces the original network. The reason why resnet is chosen is due to its powerful feature extraction capabilities [12]. Then since more feature maps are added in stage 2's model alone the channel dimension, the addition of CBAM to the network is particularly useful for selecting valuable feature layers given the increased number of feature maps, ensuring effective feature utilization [10]. The pipeline is shown in the Fig. 2.

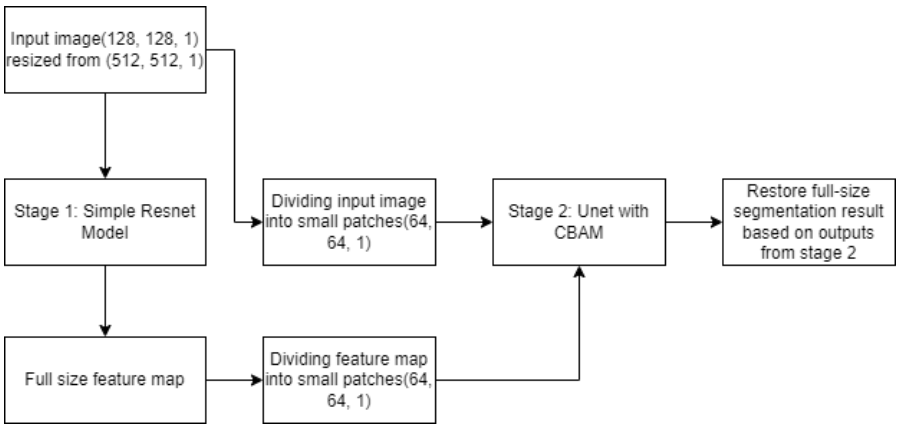


Figure 2. Pipeline of the model

2.2.1 Local and global feature fusion

The original nnU-Net proposed a dual-network structure, where one Unet extracts global information from full images and the second uNet processes local information from a small patch of the image with corresponding segmentation results in the same patch generated from the first network. One major change proposed in this study is to replace the complete Unet structure in stage 1 with a simplified resnet as the global feature extractor. One advantage is that this method is able to significantly reduce the amount of network parameters. Additionally, the original nnU-Net performs encoding and decoding operations twice on the same image region through two U-Nets in stage 1 and 2. This method considers that the decoding operation in the first stage can be effectively completed in the decoding layer of the network in stage 2 along with the local features. Therefore, this improvement can eliminate redundant decoding operations.

2.2.2 Attention mechanism

CBAM is a lightweight yet highly effective attention mechanism designed to bolster the performance of CNNs. This attention mechanism comprises two primary components: channel attention and spatial attention. Each module serves a distinct purpose to boost the network's performance by making the model focus on the most valuable features. The channel attention module operates by evaluating the importance of different feature channels. It generates a channel-wise attention map that emphasizes informative channels while suppressing less useful ones. This process is achieved by both global average and max pooling and a shared multi-layer perceptron (MLP) to learn the significance of each channel. By doing so, the channel attention module ensures that the model focuses on the channels that contribute mostly to the representation of the input data.

Secondly, the other attention module works on the spatial dimensions of the feature maps. It generates an attention map and it highlights the interesting areas while diminishing the less important areas. This is accomplished by aggregating feature information across channels through average pooling and max pooling and then applying a convolutional layer and generate an attention map in spatial dimension. This improvement helps the model to focus on important regions and boosts the model’s ability to capture spatial dependencies effectively. When integrated into a CNNs, CBAM enhances the network's capability to focus on meaningful features across both channel and spatial dimensions. In networks with multiple layers of feature maps, CBAM proves particularly beneficial. As the number of feature map layers increases, CBAM selectively emphasizes the most relevant features, thereby optimizing the ability of the model. This emphasis makes sure that additional layers have a positive contribution to the overall network performance, rather than overwhelming the system with redundant or less critical information. Consequently, CBAM is a valuable tool in improving the efficacy of CNNs by refining feature extraction and representation processes.

2.2.3 Loss function

In this experiment, loss based on dice score and Binary Cross Entropy (BCE) loss are applied as the main losses during the training of the model since they are very effective in handling segmentation tasks.

Dice loss is widely used in many segmentation tasks. This loss function is effective in dealing with class imbalance issues. It is based on the dice coefficient and is defined as:

$$DiceLoss = 1 - \frac{2*|X \cap Y|}{|X|*|Y|} \tag{1}$$

In this function, X is the result of the prediction. And Y represents the target value. This loss is effective in dealing with class imbalance issues. BCE Loss is also commonly applied on binary classification tasks and is defined as:

$$BCELoss = -\frac{1}{N}\sum_{i=1}^{N}[y_i\, log(\, p_i) + (1 - y_i)\, log(\, 1 - p_i)] \tag{2}$$

y_i is target value, p_i is the result from the prediction, and N represents the total amount of samples.

2.3. Implementation of Details

For data preprocessing step, images are resized to 128 x 128 for the model in stage 1 and are divided into 64 x 64 patches for the model in stage 2. Images are also converted to grayscale for parameters and calculation reduction. Random lightness adjustment and random rotation are added as the data augmentation methods. 0.001 is set as the initial value of the learning rate at the beginning of the training process. and Adam is chosen as the optimizer. In stage 1, the simplified Resnet model will generate a feature map with a size of 512 x 512 x 32. Then the model in stage 2 takes the patches of the original image and the feature from stage 1 as the input. One key point of this model is that the patch of the feature map is combined with other information generated by the first layer and sent to the next layer of the model. This paper believes that instead of concatenating binary mask with image data together as the input of the model, combining similar feature maps together during the middle of inference brings a few advantages. For example, it is easier for the model to learn context information through a bunch of feature maps that share similar properties than a combination of image data and binary mask data.

3. Result and Discussion

In this research, the modified nnUnet model is applied to achieve retina vessel segmentation tasks through 100 retina images with retina vessel annotations. The experiment includes a comparison among the original Unet network, nnUnet network, and modified nnUnet network. The loss curves during the training are demonstrated in Fig. 3 and Fig. 4 below.

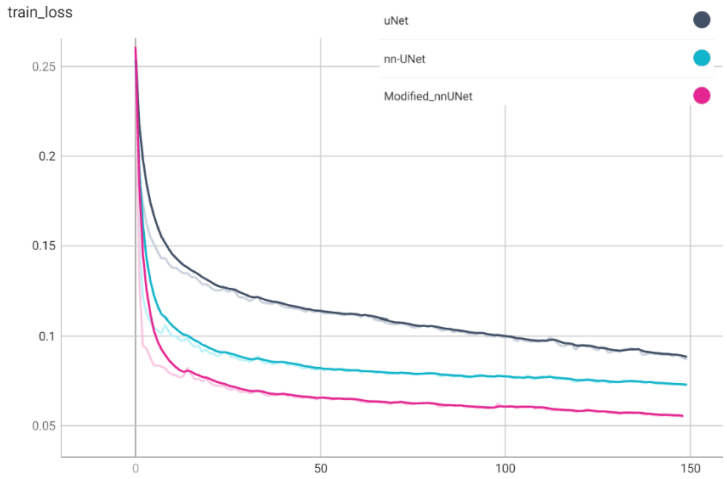


Figure 3. Loss curve for Unet, nnUNet, modified nnUNet

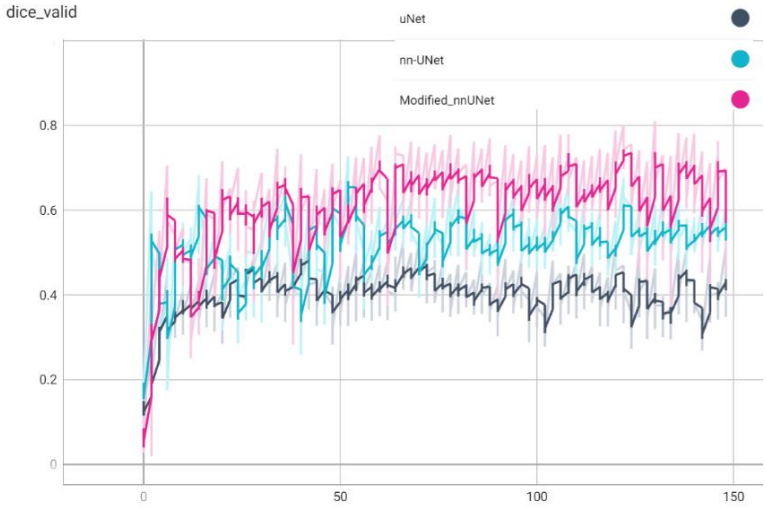


Figure 4. Validation dice score for Unet, nnUNet, modified nnUNet

According to the result from Fig. 3, the modified uuUnet shows a better performance through the loss curve on the training dataset with a faster loss descending speed. And according to Fig. 4, in general, the modified nnUnet achieves the best dice score of 80% compared with the original Unet and nnUnet models. This result can be attributed to the merging of feature maps on local and global scales. By doing so, the model can receive extra information from different views and conduct more precise results based on two kinds of features. In particular, the local feature extracted from the model in stage 2 allows for precise segmentation on tiny vessels. On the other hand, the feature from the global view helps the model to keep the continuances of vessels when they cross among patches. Also, simplification of the model in stage 1 allows faster calculation time as shown in Table 1.

Table 1. Dice score for uNet, nnUnet, modified nnUnet

Model	Best Dice Score	Average prediction time per image
Unet (128 * 128)	0.5322	0.54s
nnUNet (stage 1: 128 * 128, stage 2: 64 * 64)	0.7264	1.23s
Modified nnUnet (stage 1: 128 * 128, patch: 64 * 64)	0.8014	0.81s

4. Conclusion

In this study, two significant modifications were applied to the original nnU-Net architecture: the incorporation of the CBAM into the Stage 2 model and the integration of global and local feature maps. The enhanced nnU-Net model demonstrated notable improvements when evaluated on the Retina Blood Vessel Dataset. Specifically, it outperformed both the original nnU-Net and the traditional U-Net models by 27% and 8% in the Dice score, respectively, and achieved reductions of 0.3 seconds and 0.4 seconds in prediction time. These improvements are attributed to the strategic combination of feature maps at different scales. By integrating global and local feature maps, the modified model effectively captures both low-level features from detailed image patches and high-level features from broader image contexts. This dual approach offers a more comprehensive interpretation of features from input images to boost the model's performance. Moreover, the CBAM blocks serve a crucial role as selectors, prioritizing the most valuable feature maps from the hybrid feature sets during inference. This selective attention mechanism ensures that the model focuses on the most pertinent features, further boosting accuracy and efficiency. For future research, exploring how and where to best combine local and global feature maps presents a valuable avenue for further improvement. For instance, investigating the effects of multiple feature fusion strategies at various network layers could provide additional performance gains. Additionally, developing more efficient methods for feature extraction could further refine the model's capabilities, offering potential enhancements in both accuracy and processing speed.

References

[1] Conor H. John F. Michael O. Mark C. Characterization of changes in blood vessel width and tortuosity in retinopathy of prematurity using image analysis. Medical Image Analysis, 2002, 6 (4): 407-429.

[2] Chaudhuri S. Chatterjee S. Katz N. et al. Detection of blood vessels in retinal images using two-dimensional matched filters. IEEE Transactions on medical imaging, 1989, 8 (3): 263-269.

[3] Mendonca A.M. Campilho A. Segmentation of retinal blood vessels by combining the detection of centerlines and morphological reconstruction. IEEE transactions on medical imaging, 2006, 25 (9): 1200-1213.

[4] Kass M. Witkin A. Terzopoulos D. Snakes: Active contour models. International journal of computer vision, 1988, 1 (4): 321-331.

[5] Ronneberger O. Fischer P. & Brox T. U-Net: Convolutional networks for biomedical image segmentation. International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI), 2015, 234-241.

[6] Diakogiannis F.I. Waldner F. Caccetta P. et al. ResUNet-a: A deep learning framework for semantic segmentation of remotely sensed data. ISPRS Journal of Photogrammetry and Remote Sensing, 2020, 162: 94-114.

[7] Oktay O. Schlemper J. Folgoc L.L. et al. Attention u-net: Learning where to look for the pancreas. 2018, arXiv preprint arXiv: 1804.03999.

[8] Liskowski P. Krawiec K. Segmenting retinal blood vessels with deep neural networks. IEEE transactions on medical imaging, 2016, 35 (11): 2369-2380.

[9] Isensee F. Petersen J. Klein A. et al. nnu-net: Self-adapting framework for u-net-based medical image segmentation. 2018, arXiv preprint: 1809.10486.

[10] Woo S. Park J. Lee J.Y. et al. Cbam: Convolutional block attention module. Proceedings of the European conference on computer vision (ECCV). 2018: 3-19.

[11] Kaggle dataset, “Retina Blood Vessel”, 2023, https://www.kaggle.com/datasets/abdallahwagih/retina-blood-vessel\.

[12] He K. Zhang X. Ren S. et al. Deep residual learning for image recognition. Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 770-778.