Improved FRU-Net Using spatial Attention Mechanism

Pijush Bhuyan
IIIT Delhi
pijush22049@iiitd.ac.in

Abu Osama Siddiqui
IIIT Delhi
osama22006@iiitd.ac.in

Kaushik Dey
IIIT Delhi
kaushik22034@iiitd.ac.in

1 Introduction

Vessel segmentation refers to the task of precisely identifying and separating blood vessels from medical images, which may include retinal images, angiograms, and CT/MRI scans. The goal is to create a binary image that shows the blood vessels as foreground objects while the remaining areas are represented as the background.

Vessel segmentation is a challenging task as it requires accurate differentiation of blood vessels from other structures in the image, which can have varying sizes, shapes, and intensities. In addition, the images may have noise, non-uniform illumination, and other artifacts that further complicate the segmentation process. However, accurate vessel segmentation is crucial for medical applications such as diagnosing and treating diseases like diabetic retinopathy and hypertension, and can provide valuable information about the circulatory system.
.

2 Related Work

2.1 Vessel Segmentation based on Deep Learning

In early image segmentation methods, an image was divided into patches and the center pixel of the patches was predicted using a network of convolutional and fully connected layers. FCN, a fully convolutional network, was later proposed to solve the problem. U-Net, which is an encoder-decoder network with skip-connections, became popular in medical image segmentation. Several network models, such as Bi-directional ConvLSTM U-Net and SCS-Net, were proposed for vessel segmentation. Mou et al. added a dual self-attention mechanism for adaptive integration of local and global features of the vessel image.

For retinal vessel and coronary angiogram segmentation, Pearl et al. proposed VSSC Net which utilizes two-vessel extraction layers added to the VGG-16 network. Kamran et al. proposed RV-GAN, a new multiscale generative architecture for accurate retinal vessel segmentation. Zhou et al. designed the pipeline of synthesizing noisy labels and proposed a Study Group Learning (SGL) scheme to improve the performance of the model trained with imperfect labels.

For retinal vessel and coronary angiogram segmentation, Pearl et al. proposed VSSC Net which utilizes two-vessel extraction layers added to the VGG-16 network. Kamran et al. proposed RV-GAN(Kamran et al., 2021), a new multiscale generative architecture for accurate retinal vessel segmentation. Zhou et al. designed the pipeline of synthesizing noisy labels and proposed a Study Group Learning (SGL) scheme to improve the performance of the model trained with imperfect labels.

2.2 High/Full Resolution Network

Extracting detailed information at a higher resolution with better accuracy is crucial for semantic segmentation. Sun proposed high-resolution networks (HRNet(Wang et al., 2019)) that maintain high-resolution representations by connecting high-to-low-resolution convolutions in parallel, which has become a popular backbone of network architecture design due to its powerful high-resolution feature learning ability. UNet++(Zhou et al., 2018) redesigns with rich skip connections to aggregate features of varying semantic scales at the decoder subnetworks, leading to a highly flexible feature fusion scheme. The first stage of the network efficiently integrates semantic representations of different depths using deep supervision.

3 Methods

We have taken the following models as the baselines:

1. **Bi-Directional ConvLSTM U-Net with Densely Connected Convolutions (Azad et al., 2019)**
   The U-Net architecture (Ronneberger et al., 2015), initially proposed by Ronneberger et al. in 2015, has been expanded in this approach to fully utilize Bi-directional ConvLSTM and Dense convolutions. Bi-directional Convolutional LSTM nodes have been incorporated to merge the feature maps obtained from the corresponding encoding path and the previous decoding up-convolutional layer of the decoding path, using a non-linear method instead of the traditional simple concatenation of feature maps. In addition, densely connected convolutions have been employed in the last convolutional layer of the encoding path to enhance feature propagation and reuse. To speed up the convergence rate, Batch Normalization (BN) has also been implemented. The figure below illustrates the model's architecture.

2. **SA-UNet: Spatial Attention U-Net for Retinal Vessel Segmentation (Guo et al., 2020)**
   This is a compact network that utilizes a modified version of structured dropout convolutional block incorporating DropBlock to tackle overfitting and batch normalization (BN) as a substitute for the original U-Net convolutional block. Furthermore, a spatial attention module has been inserted between the encoder and decoder blocks, which enables the network to concentrate on significant features and suppress redundant ones, resulting in an enhanced representation capability.

The proposed models comprise of the following components:

3.1 Full Resolution UNet

The architecture of FR-UNet is depicted in Fig. 1. The network grows both horizontally and vertically through convolution, down-sampling, and up-sampling, similar to the structure of UNet++(Zhou et al., 2018). The shallow stage provides detailed semantic information, while the deep stages enhance high-level contextual information and expand the local receptive field of the feature maps. Each stage integrates the feature maps of neighboring locations in parallel expansions and learns hierarchical representations. FR-UNet comprises feature aggregation modules, residual blocks, up-sampling, and down-sampling. The network's up-sampling and down-sampling involve a convolution block with a Conv layer, batch normalization layer, and LeakyReLU activation function with a negative slope of 0.1. The down-sampling involves a 2×2Conv with a stride of 2, reducing the number of channels by half and doubling the spatial size, while the Conv layer in the up-sampling is a 2×2 deconvolution with a stride of 2, doubling the number of feature map channels and halving the spatial size.

3.2 Feature Aggregation Module

Feature Aggregation Module: The feature aggregation module of the network merges feature maps from the previous residual block with those from the up-sampling and down-sampling operations of adjacent stages. This step is followed by three different types of convolutions: 1x1, 3x3, and 3x3 atrous with a dilated rate of 2. The resulting feature maps from the three convolution modes are then added, and passed through a BN layer for normalization. The modified residual block that integrates a dropout layer (with a dropout rate of 20%) after each BN layer is employed to prevent overfitting.

3.3 Deep Supervision

In the first stage, probability maps are generated using a 1×1 Conv based on the feature maps of the last few convolution blocks, producing outputs I to V. The final probability prediction is obtained by computing a weighted sum of these probability maps. The loss function used to train the model is binary cross-entropy, which is calculated on the probability map as:

$$L(y, \hat{y}) = -\frac{1}{N}\sum_{i=1}^{n} [y_i \log(\hat{y}_i) + (1 - y_i) \log(1 - \hat{y}_i)]$$

where y and $\hat{y}$ indicate predicted probability and ground truth of ith image, respectively; N denotes the batch size.

3.4 Full Resolution UNet (FRU-Net) with Spatial Attention Mechanism-

We have incorporated the spatial attention module into the base FRU-Net architecture to help the encoder-decoder model emphasize more on the important features and thus suppress the redundant ones.

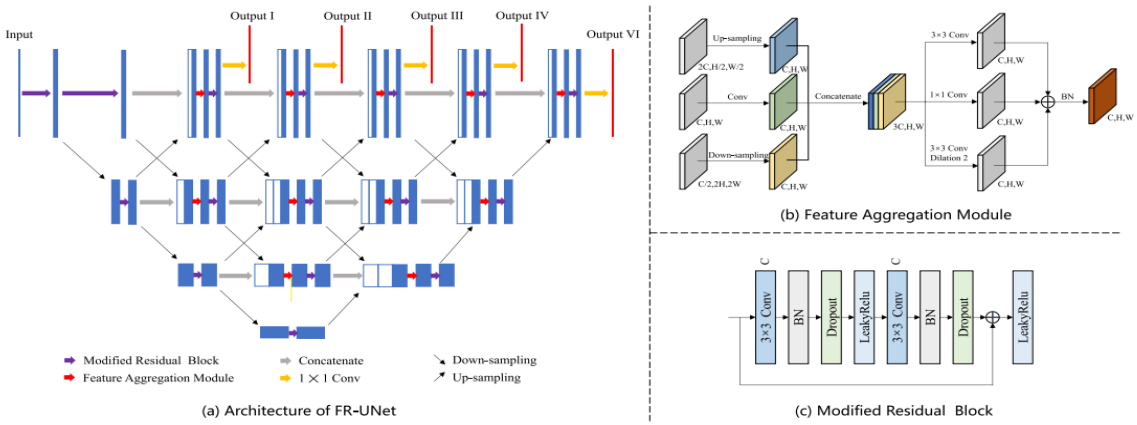

Figure 1: Architecture of FR-UNet

4 Experimental Setup

4.1 Dataset

The DRIVE (Digital Retinal Images for Vessel Extraction) dataset is intended for retinal vessel segmentation and comprises 40 color fundus images in JPEG format, including 7 abnormal pathology cases, obtained from a diabetic retinopathy screening program in the Netherlands. Each image has a resolution of 584*565 pixels with 8 bits per color channel (3 channels). The dataset was divided equally into a training set and a testing set, with each set containing 20 images. Each image in both sets has a circular field of view (FOV) mask of approximately 540 pixels in diameter. For each image in the training set, a manual segmentation by an ophthalmological expert was performed, while for each image in the testing set, two manual segmentations were conducted by two different observers, with the first observer's segmentation considered as the ground truth for performance evaluation.

4.1.1 Training

Training of BCDUNet: The code runs through the entire train test datasets and then saves them as hdf5 file and saves it in the appropriate folder. Random patches of size 64 x64 were extracted and saved as a numpy file. 20% of the training samples were used as validation dataset and trained for 50 epochs.

Training of SA-UNet: To augment the training data, four different augmentation techniques were applied, namely Random Rotation, Gaussian Noise, Color Jittering, and Horizontal, Vertical, and Diagonal Flips. The model was then trained from scratch using this augmented training set, with the binary cross-entropy loss function and Adam optimizer. The training was conducted over 150 epochs, with a learning rate of 0.001 for the first 100 epochs and 0.0001 for the remaining 50 epochs. The DropBlock method was used with a block size of 7 and a dropout rate of 0.18, while the batch size was set to 8.

Training of FR-UNet: To augment the training data, four different augmentation techniques were applied, namely Random Rotation, Center Cropping and Horizontal, Vertical Flips. Moreover patches were extracted from the training dataset to increase the amount of training data.The model was then trained from scratch using this augmented training set, with the binary focal loss function and Adam optimizer with weight decay and a learning rate scheduler.

Training of FR-UNet with spatial attention : The same augmentations and patch extraction techniques were applied and the model was trained from scratch using binary focal loss function and Adam optimizer with weight decay and a learning rate scheduler.

4.2 Results

We conducted vessel segmentation experiments on the most popular networks, including U-Net(Ronneberger et al., 2015), UNet++(Zhou et al., 2018), Attention U-Net (Oktay et al., 2018) and HRNet(Wang et al., 2019). Table I provides the qualitative results of vessel segmentation for retinal vessel dataset DRIVE.

* Complete results could not be generated due to resource limitations .
Although the training process could not be com-

Output	F1 Score	Sensitivity	Specificity	Accuracy	AUC
SA-UNet(Guo et al., 2020)	0.8221	0.8234	0.9840	0.9708	0.9872
BCD-UNet(Azad et al., 2019)	0.8222	0.8012	0.9784	0.9559	0.9788
FR-UNet (Liu et al., 2022)	0.8316	0.8356	0.9837	0.9705	0.9889
FR-UNet replicated*	0.0430	0.0222	0.9832	0.3211	0.5027
FR-UNet with spatial attention* (ours)	0.0570	0.0296	0.9821	0.3259	0.5059

Table 1: Performance of the existing baseline models and ours for retinal vessel segmentation.

pleted for both the FR-UNet as well as the variant with spatial attention mechanism, it was found that the later converge relatively faster in terms of loss as well as evaluation metrics during the training phase. The same can be interpreted from the testing results as shown in the Table 1.

5 Error Analysis

The model is overfitting on the training data clearly as we can see that model performs better during train time but gives poor results during validation and also during test. May be training the model for more iterations and also reducing the strides during patch extraction will have solved this problem but increased computational load.

Contributions

The baselines were trained and results were replicated by Abu Osama Siddiqui and Pijush Bhuyan. The architecture for the FRU-Net was written from scratch and trained by Kaushik Dey and Pijush Bhuyan. The spatial attention module was developed and the modified architecture was trained by Abu Osama Siddiqui and Kaushik Dey. The report was created by the joint effort of all the members.

References

Reza Azad, Maryam Asadi-Aghbolaghi, Mahmood Fathy, and Sergio Escalera. 2019. Bi-directional convlstm u-net with densley connected convolutions.

Changlu Guo, Márton Szemenyei, Yugen Yi, Wenle Wang, Buer Chen, and Changqi Fan. 2020. Sa-unet: Spatial attention u-net for retinal vessel segmentation.

Sharif Amit Kamran, Khondker Fariha Hossain, Alireza Tavakkoli, Stewart Lee Zuckerbrod, Kenton M. Sanders, and Salah A. Baker. 2021. RV-GAN: Segmenting retinal vascular structure in fundus photographs using a novel multi-scale generative adversarial network. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2021*, pages 34–44. Springer International Publishing.

Wentao Liu, Huihua Yang, Tong Tian, Zhiwei Cao, Xipeng Pan, Weijin Xu, Yang Jin, and Feng Gao. 2022. Full-resolution network and dual-threshold iteration for retinal vessel and coronary angiograph segmentation. *IEEE journal of biomedical and health informatics*, PP.

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. 2018. Attention u-net: Learning where to look for the pancreas. *CoRR*, abs/1804.03999.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 2015. U-net: Convolutional networks for biomedical image segmentation. *CoRR*, abs/1505.04597.

Jingdong Wang, Ke Sun, Tianheng Cheng, Borui Jiang, Chaorui Deng, Yang Zhao, Dong Liu, Yadong Mu, Mingkui Tan, Xinggang Wang, Wenyu Liu, and Bin Xiao. 2019. Deep high-resolution representation learning for visual recognition. *CoRR*, abs/1908.07919.

Zongwei Zhou, Md Mahfuzur Rahman Siddiquee, Nima Tajbakhsh, and Jianming Liang. 2018. Unet++: A nested u-net architecture for medical image segmentation. *CoRR*, abs/1807.10165.