

Bi-Directional ConvLSTM U-Net with Densely Connected Convolutions

Reza Azad*
 Sharif University
 of Tech., Iran
 rezazad68@gmail.com

Maryam Asadi-Aghbolaghi*
 Inst. for Research in
 Fundamental Sciences (IPM), Iran
 masadi@ipm.ir

Mahmood Fathy
 Iran University of
 Science and Tech., Iran
 mahfathy@iust.ac.ir

Sergio Escalera
 Universitat de Barcelona and
 Computer Vision Center, Spain
 sergio@maia.ub.es

Abstract

In recent years, deep learning-based networks have achieved state-of-the-art performance in medical image segmentation. Among the existing networks, U-Net has been successfully applied on medical image segmentation. In this paper, we propose an extension of U-Net, Bi-directional ConvLSTM U-Net with Densely connected convolutions (BCDU-Net), for medical image segmentation, in which we take full advantages of U-Net, bi-directional ConvLSTM (BConvLSTM) and the mechanism of dense convolutions. Instead of a simple concatenation in the skip connection of U-Net, we employ BConvLSTM to combine the feature maps extracted from the corresponding encoding path and the previous decoding up-convolutional layer in a non-linear way. To strengthen feature propagation and encourage feature reuse, we use densely connected convolutions in the last convolutional layer of the encoding path. Finally, we can accelerate the convergence speed of the proposed network by employing batch normalization (BN). The proposed model is evaluated on three datasets of: retinal blood vessel segmentation, skin lesion segmentation, and lung nodule segmentation, achieving state-of-the-art performance.

1. Introduction

Medical images play a key role in medical treatment and diagnosis. The goal of Computer-Aided Diagnosis (CAD) systems is providing doctors with precise interpretation of medical images to have better treatment of a large number of people. Moreover, automatic processing of medical images results in reducing the time, cost, and error of human-based processing. One of the main research areas in this

field is medical image segmentation, being a critical step in numerous medical imaging studies. Like other fields of research in computer vision, deep learning networks achieve outstanding results and use to outperform non-deep state-of-the-art methods in medical imaging. Deep neural networks are mostly utilized in classification tasks, where the output of the network is a single label or probability values associated labels to a given input image. These networks work fine thanks to some structural features [2] such as: activation function, different efficient optimization algorithms, and dropout as a regularizer for the network. These networks require a large amount of data to train and provide a good generalization behavior given the huge number of network parameters. A critical issue in medical image segmentation is the unavailability of large (and annotated) datasets. In medical image segmentation, per pixel labeling is required instead of image level label.

Fully convolutional neural network (FCN) [17] was one of the first deep networks applied to image segmentation. Ronneberger et al. [21] extended this architecture to U-Net, achieving good segmentation results leveraging the need of a large amount of training data. Their network consists of encoding and decoding paths. In the encoding path a large number of feature maps with reduced dimensionality are extracted from the input data. The decoding path is used to produce segmentation maps (with the same size as the input) by performing up-convolutions. Many extensions of U-Net have been proposed so far [2, 19]. The most important modification is mainly about the skipping connections. In some extended versions of U-Net, the extracted feature maps in the skip connection are first fed to a processing step (e.g. attention gates [19]) and then concatenated. The main drawback of these networks is that the processing step is performed individually for the two sets of feature maps, and these features are then simply concatenated.

*These two authors contributed equally.

In this paper, we propose BCDU-Net, an extended version of the U-Net, by including BConvLSTM [23] in the skip connection and reusing feature maps with densely convolutions. The feature maps from the corresponding encoding layer have higher resolution while the feature maps extracted from the previous up-convolutional layer contain more semantic information. Instead of a simple concatenation, combining these two kinds of feature maps with non-linear functions may result in more precise segmentation output. Therefore, in this paper we extend the U-Net architecture by adding BConvLSTM in the skip connection to combine these two kinds of feature maps.

Having a sequence of convolutional layers may help the network to learn more kinds of features; however, in many cases, the network learns redundant features. To mitigate this problem and enhance information flow through the network, we utilize the idea of densely connected convolutions [12]. In the last layer of the contracting path, convolutional blocks are connected to all subsequent blocks in that layer via channel-wise concatenation. Features which are learned in each block are passed forward to the next block. This strategy helps the method to learn a diverse set of features based on the collective knowledge gained by previous layers, and therefore, avoiding learning redundant features..

Furthermore, we accelerate the convergence speed of the network by employing BN after the up-convolution filters. We evaluate the proposed BCDU-Net on three different applications of: retinal blood vessel segmentation (DRIVE dataset), Skin lesion segmentation (ISIC 2018 dataset) and lung nodule segmentation (Lung dataset). The experimental results demonstrate that the proposed network achieves superior performance than state-of-the-art alternatives.¹

2. Related Work

One of the most crucial tasks in medical imaging is semantic segmentation. Before the revolution of deep learning in computer vision, traditional handcrafted features were exploited for semantic segmentation. During the last few years, deep learning-based approaches have outstandingly improved the performance of classical image segmentation strategies. Based on the exploited deep architecture, these approaches can be divided into three groups of: convolutional neural network (CNN), fully convolutional network (FCN), and recurrent neural network (RNN).

2.1. Convolutional Neural Network (CNN)

Cui et al. [10] exploited CNN for automatic segmentation of brain MRI images. The authors first divided the input images into some patches and then utilized these patches for training CNN. To handle an arbitrary number of modalities as the input data, Kleesiek et al. [14] proposed a 3D

CNN for brain lesion segmentation. To process MRI data, the network consists of four channels: non-enhanced and contrast-enhanced T1w, T2w and FLAIR contrasts. Roth et al. [22] proposed a multi-level deep convolutional networks for pancreas segmentation in abdominal CT scans as a probabilistic bottom-up approach.

2.2. Fully Convolutional Network (FCN)

One of the main problems of the CNN models for segmentation tasks is that the spatial information of the image is lost when the convolutional features are fed into the fc layers. To overcome this problem the *fully convolutional network* (FCN) was proposed by Long et al. [17]. This network is trained end-to-end and pixels-to-pixels for semantic segmentation. All fc layers of the CNN architecture are replaced with convolutional and deconvolutional ones to keep the original spatial resolutions. Therefore, the original spatial dimension of the features maps are recovered while the network is performing the segmentation task. FCN has been frequently utilized for segmentation of medical and biomedical images [27, 28]. Zhou et al. [27] exploited FCN for segmentation of anatomical structures on 3D CT images. An FCN with convolution and de-convolution parts is trained end-to-end, performing voxel-wise multiple-class classification to map each voxel in a CT image to an anatomical label. Drozdzal et al. [11] proposed very deep FCN by using short skip connections. The authors showed that a very deep FCN with both long and short skip connections achieved better result than the original one.

U-Net, proposed by Ronneberger et al. [21], is one of the most popular FCNs for medical image segmentation. This network consists of contracting and expanding paths. U-Net has some advantages than the other segmentation-based network [2]. It works well with few training samples and the network is able to utilize the global location and context information at the same time. Milletari et al. [18] proposed V-Net, a 3D extension version of U-Net to predict segmentation of a given volume at once. In that network, the authors proposed an end-to-end 3D image segmentation network based on a volumetric (MRI volumes), fully convolutional, neural network. 3D U-Net [7] is proposed for processing 3D volumes instead of 2D images as input. In which, all 2D operations of U-Net are replaced with their 3D counterparts. VoxResNet [5], a deep voxel-wise residual network, was proposed for brain segmentation from MR. This 3D network is inspired by deep residual learning, performing summation of feature maps from different layers.

2.3. Recurrent Neural Network (RNN)

Pinheiro et al. [20] proposed an end-to-end feed forward deep network consisting of an RNN that can take into account long range label dependencies in the scenes while limiting the capacity of the model. Visin et al. [25] pro-

¹Source code is available on <https://github.com/rezazad68/BCDU-Net>.

posed ReSeg for semantic segmentation. In that network, the input images are processed with a pre-trained VGG-16 model and its resulting feature maps are then fed into one or more ReNet layers. DeepLab architecture [6] contains a deep convolutional neural network in which all fully connected layers are replaced by convolutional layers and then the feature resolution is increased through atrous convolutional layers. Alom et al. [2] proposed Recurrent Convolutional Neural Network (RCNN) and Recurrent Residual Convolutional Neural Network (R2CNN) based on U-Net models for medical image segmentation. Bai et al. [4] combined an FCN with an RNN for medical image sequence segmentation, which is able to incorporate both spatial and temporal information for MR images.

In this paper, BCDU-Net is proposed as an extension of U-Net, showing better performance than state-of-the-art alternatives for the segmentation task. Moreover, BN has a significant effect on the convergence speed of the network.

3. Proposed Method

Inspired by U-Net [21], BConvLSTM [23], and dense convolutions [12], we propose the BCDU-Net as shown in Figure 1. The network utilizes the strengths of both BConvLSTM states and densely connected convolutions. We detail different parts of the network in the next sub sections.

3.1. Encoding Path

The contracting path of BCDU-Net includes four steps. Each step consists of two convolutional 3×3 filters followed by a 2×2 max pooling function and ReLU. The number of feature maps are doubled at each step. The contracting path extracts progressively image representations and increases the dimension of these representations layer by layer. Ultimately, the final layer in the encoding path produces a high dimensional image representation with high semantic information. The original U-Net contains a sequence of convolutional layers in the last step of encoding path. Having a sequence of convolutional layers in a network yields the method learn different kinds of features. Nevertheless, the network might learn redundant features in the successive convolutions. To mitigate this problem, densely connected convolutions are proposed [12]. This helps the network to improve its performance by the idea of “collective knowledge” in which the feature maps are reused through the network. It means feature maps learned from all previous convolutional layers are concatenated with the feature map learned from the current layer and then are forwarded to use as the input to the next convolution.

The idea of densely connected convolutions has some advantages over the regular convolutions [12]. First of all, it helps the network to learn a diverse set of feature maps instead of redundant features. Moreover, this idea improves the network’s representational power by allowing informa-

tion flow through the network and reusing features. Furthermore, dense connected convolutions can benefit from all the produced features before it, which prompt the network to avoid the risk of exploding or vanishing gradients. In addition, the gradients are sent to their respective places in the network more quickly in the backward path. We employ the idea of densely connected convolutions in the proposed network. To do that, we introduce one block as two consecutive convolutions. There are a sequence of N blocks in the last convolutional layer of the encoding path, shown in Figure 2. These blocks are densely connected. We consider \mathcal{X}_e^i as the output of the i^{th} convolutional block. The input of the i^{th} ($i \in \{1, \dots, N\}$) convolutional block receives the concatenation of the feature maps of all preceding convolutional blocks as its input, i.e., $[\mathcal{X}_e^1, \mathcal{X}_e^2, \dots, \mathcal{X}_e^{i-1}] \in \mathbb{R}^{(i-1)F_l \times W_l \times H_l}$, and the output of the i^{th} block is $\mathcal{X}_e^i \in \mathbb{R}^{F_l \times W_l \times H_l}$. In the remaining part of the paper we use simply \mathcal{X}_e instead of \mathcal{X}_e^N .

3.2. Decoding Path

Each step in the decoding path starts with performing an up-sampling function over the output of the previous layer. In the standard U-Net, the corresponding feature maps in the contracting path are cropped and copied to the decoding path. These feature maps are then concatenated with the output of the up-sampling function. In BCDU-Net, we employ BConvLSTM to process these two kinds of feature maps in a more complex way. Let $\mathcal{X}_e \in \mathbb{R}^{F_l \times W_l \times H_l}$ be the set of feature maps copied from the encoding part, and $\mathcal{X}_d \in \mathbb{R}^{F_{l+1} \times W_{l+1} \times H_{l+1}}$ be the the set of feature maps from the previous convolutional layer, where F_l is number of feature maps at layer l , and $W_l \times H_l$ is the size of each feature map at layer l . It is worth mentioning that $F_{l+1} = 2 * F_l$, $W_{l+1} = \frac{1}{2} * W_l$, and $H_{l+1} = \frac{1}{2} * H_l$. Based on Figure 3, \mathcal{X}_d is first passed to an up-convolutional layer in which an up-sampling function followed by a 2×2 convolution are applied, doubling the size of each feature map and halving the number of feature channels, i.e., producing $\mathcal{X}_d^{up} \in \mathbb{R}^{F_l \times W_l \times H_l}$. In other words, the expanding path increases the size of the feature maps layer by layer to reach the original size of the input image after the final layer.

3.2.1 Batch Normalization:

After up-sampling, \mathcal{X}_d^{up} goes through a BN function and produces $\hat{\mathcal{X}}_d^{up}$. A problem in the intermediate layers in training step is that the distribution of the activations varies. This problem makes the training process very slow since each layer in every training step has to learn to adapt themselves to a new distribution. BN [13] is utilized to increase the stability of a neural network, which standardizes the inputs to a layer in the network by subtracting the batch mean and dividing by the batch standard deviation. BN affect-

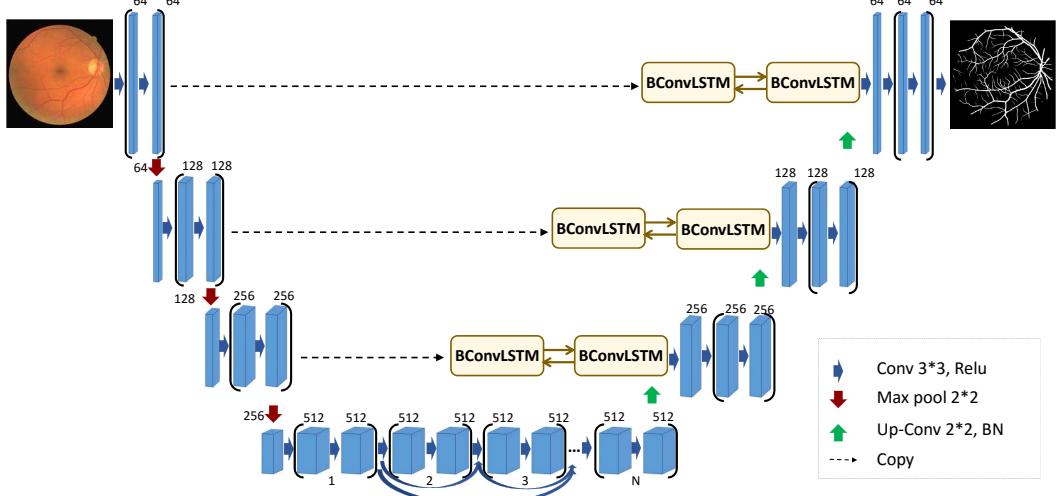


Figure 1. BCDU-Net with bi-directional ConvLSTM in the skip connections and densely connected convolution.

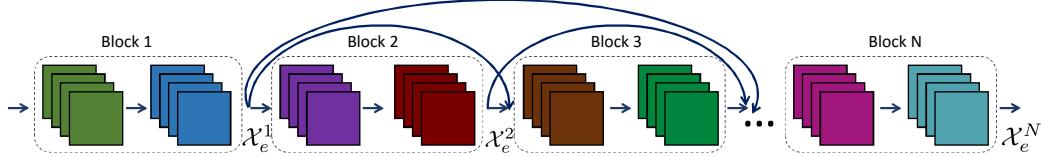


Figure 2. Dense layer of the BCDU-Net.

edly accelerates the speed of training process of a neural network. Moreover, in some cases the performance of the model is improved thanks to the modest regularization effect. More details can be found in [13].

3.2.2 Bi-Directional ConvLSTM:

The output of the BN step ($\hat{\mathcal{X}}_d^{up} \in \mathbb{R}^{F_l \times W_l \times H_l}$) is now fed to a BConvLSTM layer. The main disadvantage of the standard LSTM is that these networks does not take into account the spatial correlation since these models use full connections in input-to-state and state-to-state transitions. To solve this problem, ConvLSTM [26] was proposed which exploited convolution operations into input-to-state and state-to-state transitions. It consists of an input gate i_t , an output gate o_t , a forget gate f_t , and a memory cell C_t . Input, output and forget gates act as controlling gates to access, update, and clear memory cell. ConvLSTM can be formulated as follows (for convenience we remove the subscript and superscript from the parameters):

$$\begin{aligned}
 i_t &= \sigma(\mathbf{W}_{xi} * \mathcal{X}_t + \mathbf{W}_{hi} * \mathcal{H}_{t-1} + \mathbf{W}_{ci} * \mathcal{C}_{t-1} + b_i) \\
 f_t &= \sigma(\mathbf{W}_{xf} * \mathcal{X}_t + \mathbf{W}_{hf} * \mathcal{H}_{t-1} + \mathbf{W}_{cf} * \mathcal{C}_{t-1} + b_f) \\
 \mathcal{C}_t &= f_t \circ \mathcal{C}_{t-1} + i_t \tanh(\mathbf{W}_{xc} * \mathcal{X}_t + \mathbf{W}_{hc} * \mathcal{H}_{t-1} + b_c) \\
 o_t &= \sigma(\mathbf{W}_{xo} * \mathcal{X}_t + \mathbf{W}_{ho} * \mathcal{H}_{t-1} + \mathbf{W}_{co} * \mathcal{C}_t + b_o) \\
 \mathcal{H}_t &= o_t \circ \tanh(\mathcal{C}_t),
 \end{aligned} \tag{1}$$

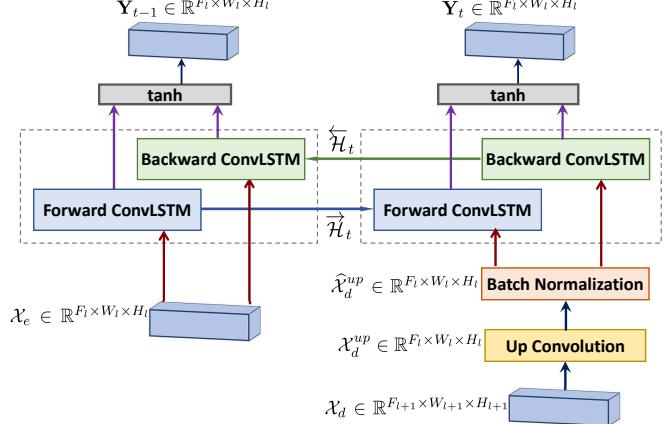


Figure 3. Bi-directional ConvLSTM in CUA-Net.

where $*$ and \circ denote the convolution and Hadamard functions, respectively. \mathcal{X}_t is the input tensor (in our case \mathcal{X}_e and $\hat{\mathcal{X}}_d^{up}$), \mathcal{H}_t is the hidden state tensor, \mathcal{C}_t is the memory cell tensor, and, \mathbf{W}_{x*} and \mathbf{W}_{h*} are 2D Convolution kernels corresponding to the input and hidden state, respectively, and b_i , b_f , b_o , and b_c are the bias terms.

In this network, we employ BConvLSTM [23] to encode \mathcal{X}_e and $\hat{\mathcal{X}}_d^{up}$. BConvLSTM uses two ConvLSTMs to process the input data into two directions of forward and backward paths, and then makes a decision for the current input by dealing with the data dependencies in both directions. In a standard ConvLSTM, only the dependencies of the for-

ward direction are processed. However, all the information in a sequence should be fully considered, therefore, it might be effective to take into account backward dependencies. It has been proved that analyzing both forward and backward temporal perspectives enhanced the predictive performance [9]. Each of the forward and backward ConvLSTM can be considered as a standard one. Therefore, we have two sets of parameters for backward and forward states. The output of the BConvLSTM is calculated as

$$\mathbf{Y}_t = \tanh \left(\mathbf{W}_y^{\vec{\mathcal{H}}} * \vec{\mathcal{H}}_t + \mathbf{W}_y^{\overleftarrow{\mathcal{H}}} \overleftarrow{\mathcal{H}}_t + b \right) \quad (2)$$

where $\vec{\mathcal{H}}_t$ and $\overleftarrow{\mathcal{H}}_t$ denote the hidden state tensors for forward and backward states, respectively, b is the bias term, and $\mathbf{Y}_t \in \mathbb{R}^{F_l \times W_l \times H_l}$ indicates the final output considering bidirectional spatio-temporal information. Moreover, tanh is the hyperbolic tangent which is utilized here to combine the output of both forward and backward states through a non-linear way. We utilize the energy function like the original U-Net to train the network.

4. Experimental Results

We evaluate BCDU-Net on DRIVE, ISIC 2018, and a lung segmentation public benchmark datasets. DRIVE is a dataset for blood vessel segmentation from retina images, ISIC is for skin cancer lesion segmentation, and the last dataset consists of diagnostic and lung cancer screening thoracic computed tomography (CT) scans with marked-up annotated lesions. The empirical results show that the proposed method outperforms state-of-the-art alternatives. Keras with TensorFlow backend is utilized for this implementation. The network is trained from scratch for all datasets. We consider several performance metrics to perform the experimental comparative, including accuracy (AC), sensitivity (SE), specificity (SP), F1-Score, Jaccard similarity (JS), and area under the curve (AUC). We stop the training of the network when the validation loss remains the same in 10 consecutive epochs which is 50, 100, and 25 for DRIVE, ISIC, and Lung datasets, respectively.

4.1. DRIVE Dataset

DRIVE [24] is a dataset for blood vessel segmentation from retina images. It includes 40 color retina images, from which 20 samples are used for training and the remaining 20 samples for testing. The original size of images is 565×584 pixels. It is clear that a dataset with this number of samples is not sufficient for training a deep neural network. Therefore, we use the same strategy as [2] for training our network. The input images are first randomly divided into a number of patches. In total, around 190,000 patches are produced from 20 training images, from which 171,000 patches are used for training, and the remaining

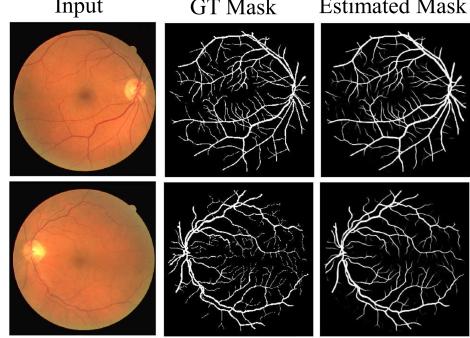


Figure 4. Segmentation result of BCDU-Net on DRIVE.

19,000 patches are used for validation. The size of batches utilized as the input data to the network is 64×64 .

Some precise and promising segmentation results of the experimental output of the proposed network are shown in Figure 9. The first column is the original color image, the second one is the ground truth mask and the third column is the output of the proposed BCDU-Net. Table 1 lists the quantitative results obtained by different methods and the proposed network on DRIVE dataset. We evaluate the network with $d = 1$ and $d = 3$ as the number of dense blocks in the network. With $d = 1$ we have one convolutional block without any dense connection in that layer, i.e., like the last encoding layer of the standard U-Net. With $d = 3$ we have three convolutional blocks and two dense connections in that layer. It is shown that the BCDU-Net (with both $d = 1$ and $d = 3$) outperforms w.r.t. the state-of-the-art alternatives for most of the evaluation metrics. Moreover, it can be seen that the network with $d = 3$ works better than the network without dense block.

To ensure the proper convergence of the proposed network, the training and validation accuracy for DRIVE dataset is shown in Figure 5 (a). It is shown that the network converges very fast, i.e., after the 30th epoch, the network is almost converged. We also can see that in the first 15 epochs the validation accuracy is larger than the training one. This fact is mostly because of the small size of dataset since we use a small set of images as the validation set. Moreover, it might be related to the fact that we evaluate the validation set at the end of epoch. To show the overall performance of the BCDU-Net on DRIVE dataset, ROC curves is shown in Figure 6 (a). ROC is the plot of the true positive rate (TPR) against the false positive rate (FPR). AUC (reported in Table 1) is the area under the ROC curve and is a measure of how well the network can segment the input data.

4.2. ISIC 2018 Dataset

The ISIC dataset [8] was published by the International Skin Imaging Collaboration (ISIC) as a large-scale dataset of dermoscopy images. This dataset is taken from a challenge on lesion segmentation, dermoscopic feature detection, and disease classification. It includes 2594 images

Table 1. Performance comparison of the proposed network and the state-of-the-art methods on DRIVE dataset.

Methods	F1-Score	Sensitivity	Specificity	Accuracy	AUC
COSFIRE filters [3]	-	0.7655	0.97048	0.9442	0.9614
Cross-Modality [15]	-	0.7569	0.9816	0.9527	0.9738
U-net [21]	0.8142	0.7537	0.9820	0.9531	0.9755
Deep Model [16]	-	0.7763	0.9768	0.9495	0.9720
RU-net [2]	0.8149	0.7726	0.9820	0.9553	0.9779
R2U-Net [2]	0.8171	0.7792	0.9813	0.9556	0.9782
BCDU-Net (d=1)	0.8222	0.8012	0.9784	0.9559	0.9788
BCDU-Net (d=3)	0.8224	0.8007	0.9786	0.9560	0.9789

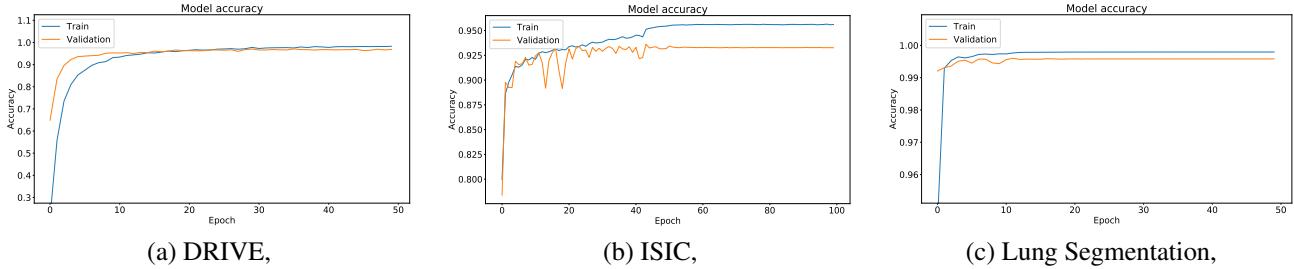


Figure 5. Training and validation accuracy of BCDU-Net for three datasets.

where like previous approaches[2], we used 1815 images for training, 259 for validation and 520 for testing. The original size of each sample is 700×900 . We use the same pre-processing as [2] on the input image, and resize images to 256×256 . The training data consists of the original images and corresponding ground truth annotations (i.e., binary images containing cancer or non-cancer lesions).

For qualitative analysis, Figure 7 shows some promising example outputs of the proposed BCDU-Net on ISIC dataset. Table 2 lists the quantitative results obtained by different methods and the proposed network on ISIC dataset. A large improvement is achieved by the BCDU-Net (with both $d = 1$ and $d = 3$) w.r.t. state-of-the-art alternatives for all of the evaluation metrics. It is clear that the network with $d = 3$ works better than the one with $d = 1$. It is worth mentioning that there was a challenge on ISIC dataset and the best result achieved by the participants was $JS = 0.802$. Compare to this result, there is a good gap between the JS achieved by the BCDU-Net (0.936) and the best result of the ISIC challenge.

The training and validation accuracy of the proposed network for ISIC dataset is shown in Figure 5 (b). Like DRIVE dataset, the convergence speed of the network for ISIC dataset is fast (after 40 epochs). The validation accuracy over the training process is variable. The reason behind this fact is that the validation set contains some images totally different from the ones in training set, therefore, during the first learning iterations the model has some problems about segmenting those images. To show the overall performance of the BCDU-Net on ISIC dataset, the ROC curves are shown in Figure 6 (b).

4.3. Lung Segmentation Dataset

A lung segmentation dataset is introduced in the Lung Nodule Analysis (LUNA) competition at the Kaggle Data Science Bowl in 2017. This dataset consists of 2D and 3D CT images with respective label images for lung segmentation [1]. We use 70% of the data as the train set and the remaining 30% as the test set. The size of each image is 512×512 . Since the lung region in CT images have almost the same Hausdorff value with non-object of interests such as bone and air, it is worth to learn lung region by learning its surrounding tissues. To do that first we extract the surrounding region by applying algorithm 1 and then make a new mask for the training sets. We train the model on these new masks and on the testing phase, and estimate the lung region as a region inside the estimated surrounding tissues. A sample is shown in Figure 8 .

Algorithm 1 Pre-processing over lung dataset.

- 1: Input = X and GT_Mask
 - 2: Output = $Surrounding_Mask$
 - 3: $X(X > 512) = 512$
 - 4: $X = Norm(X)$ ▷ Normalize X
 - 5: $X = image2binary(X)$ ▷ Convert to binary
 - 6: $X = X \cup GT_Mask$
 - 7: $X = Morphology(X)$ ▷ Remove noise
 - 8: $Surrounding_Mask = X - GT_Mask$
-

Figure 9 shows some segmentation outputs of the proposed network for lung dataset. The quantitative results of the proposed BCDU-Net is compared with other methods in Table 3. It is clear that the BCDU-Net (with both $d = 1$ and $d = 3$) outperforms the other methods. Moreover, the

Table 2. Performance comparison of the proposed network and the state-of-the-art methods on ISIC dataset.

Methods	F1-Score	Sensitivity	Specificity	Accuracy	PC	JS
U-net [21]	0.647	0.708	0.964	0.890	0.779	0.549
Attention U-net [19]	0.665	0.717	0.967	0.897	0.787	0.566
R2U-net [2]	0.679	0.792	0.928	0.880	0.741	0.581
Attention R2U-Net [2]	0.691	0.726	0.971	0.904	0.822	0.592
BCDU-Net (d=1)	0.847	0.783	0.980	0.936	0.922	0.936
BCDU-Net (d=3)	0.851	0.785	0.982	0.937	0.928	0.937

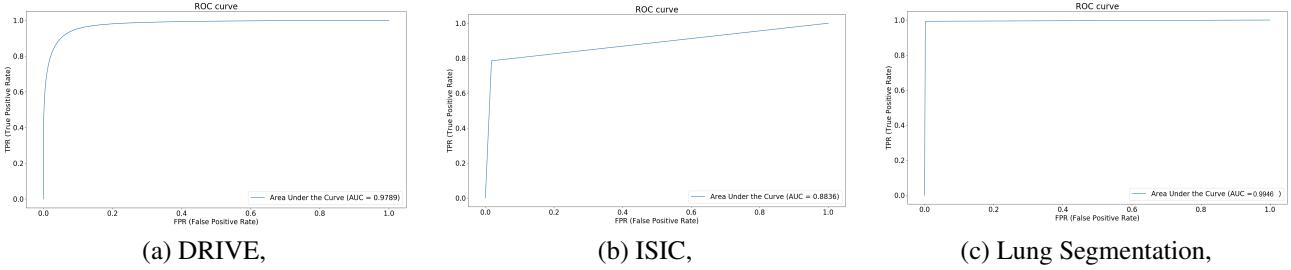


Figure 6. ROC diagrams of the proposed BCDU-Net for three dataset.

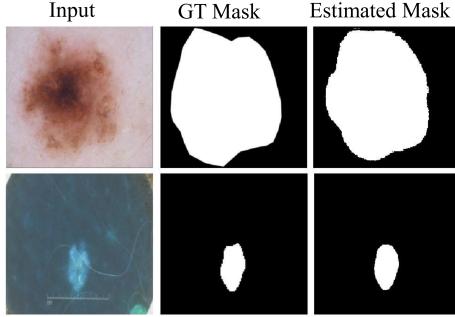


Figure 7. Segmentation result of BCDU-Net on ISIC.

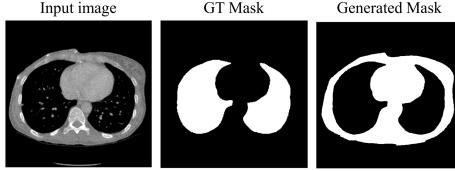


Figure 8. A sample of generated mask for Lung dataset.

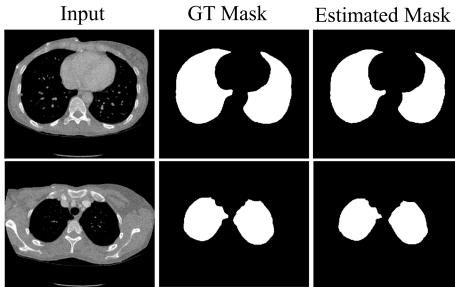


Figure 9. Segmentation result of BCDU-Net on Lung dataset.

network with dense connections works better. The training and validation accuracy for this dataset is shown in Figure 5 (c). To show the overall performance of the network on this dataset, ROC curves is shown in Figure 6 (c).

4.4. Discussion

The proposed network has some modifications from the original U-Net. We summarized the "Accuracy" and "F1-Score" of the original U-Net and its modifications for three utilized datasets in Table 4. We evaluate each modified part of the network to analyze its influence on the result. In Table 4, it can be seen that the result of the standard U-Net is improved by inserting BConvLSTM in the skip connections. Figure 10 shows the output segmentation mask of the original U-Net and BCDU-Net for two samples of the ISIC dataset. It shows a more precise and fine segmentation output of the proposed network than the original U-Net. After the skip connections, there are two kinds of features to combine, i.e., the features from the previous decoding layer and the features from the corresponding encoding layer. For convenience, we call them the encoded and decoded features. In the original U-Net, a simple concatenation function is used to combine these two kinds of features.

In the proposed network, we used a set of BConvLSTMs to combine encoded and decoded features. The encoded features have higher resolution and therefore contain more local information of the input image, while the decoded features have more semantic information about the input images. The affection of these two features over each other might result in a set of feature maps rich in both local and semantic information. Therefore, instead of a simple concatenation, we utilize BConvLSTM to combine the encoded and decoded features. In BConvLSTM, a set of convolution filters are applied on each kind of features. Therefore each ConvLSTM state, corresponds to one kind of features (e.g. encoded features), is able to encode relevant information about the other kind of features (e.g. decoded features). The convolutional filters along with the hyperbolic tangent functions help the network to learn complex data structures.

Table 3. Performance comparison of the proposed network and the state-of-the-art methods on Lung dataset.

Methods	F1-Score	Sensitivity	Specificity	Accuracy	AUC	JS
U-net [21]	0.9658	0.9696	0.9872	0.9872	0.9784	0.9858
RU-net [2]	0.9638	0.9734	0.9866	0.9836	0.9800	0.9836
R2U-Net [2]	0.9832	0.9944	0.9832	0.9918	0.9889	0.9918
BCDU-Net (d=1)	0.9889	0.9901	0.9979	0.9967	0.9940	0.9967
BCDU-Net (d=3)	0.9904	0.9910	0.9982	0.9972	0.9946	0.9972

Table 4. Performance comparison of U-Net and its modifications in our work.

Methods	DRIVE		ISIC		Lung	
	F1-Score	AC	F1-Score	AC	F1-Score	AC
U-net	0.8142	0.9531	0.6470	0.8900	0.9658	0.9828
U-Net + BConvLSTM (d=1)	0.8222	0.9559	0.8470	0.9360	0.9889	0.9967
U-Net + BConvLSTM + Dense Conv (d=3)	0.8243	0.9560	0.8506	0.9374	0.9904	0.9972

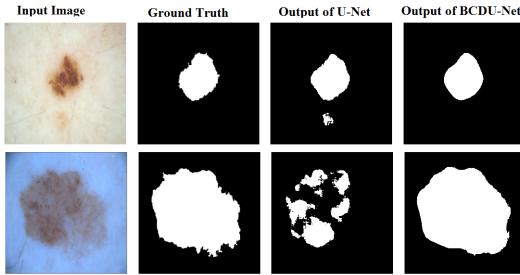


Figure 10. Visual effect of BConvLSTM in BCDU-Net.

We included BN after each up-convolutional layer to speed up the network learning process. To evaluate the effect of this function, we train the network with and without BN. Figure 11 (a) shows the training and validation accuracy of BCDU-Net for ISIC dataset without BN and Figure 11 (b) shows the same contend for the network with BN. BCDU-Net converged after 200 epochs without BN while this number is about 30 with BN, i.e., BN yields the network to converge 6.6 times faster. Moreover, it can be seen that BN has improved the accuracy of the BCDU-Net. The variations among data in the ISIC dataset is larger than the other datasets. BN manages these variations by standardizing data through controlling the mean and variance of distributions of inputs which results in a small regularization and reducing generalization error. Therefore, BN helps the network to improve the performance.

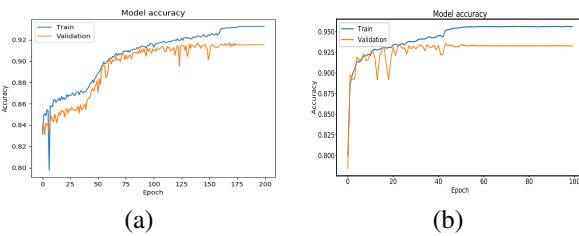


Figure 11. Training and validation accuracy of BCDU-Net (a) without and (b) with BN.

Table 4 shows that the network with dense connections improve the accuracy and F1-Score for the three datasets.

The key idea of dense convolutions is sharing feature maps between blocks through direct connection between convolutional block. Consequently, each dense block receives all preceding layers as input, and therefore, produces more diversified and richer features. Thus, it helps the network to increase the representational power of deeper models.

We have more feature propagation both in backward and forward paths through dense blocks. The network performs a kind of deep supervision in backward path since dense block receives additional supervision from loss function through shorter connections [12]. The error signal is propagated to earlier layers more directly, hence, earlier layers can get direct supervision from the final softmax layer, and moreover, it results in decreasing the vanishing-gradient problem. In addition, compared to other deep architectures like residual connections, dense convolutions require fewer parameters while improving the accuracy of the network.

5. Conclusion

We proposed BCDU-Net for medical image segmentation. We showed that by including BConvLSTM in the skip connection and also inserting a densely connected convolutional blocks, the network is able to capture more discriminative information which resulted in more precise segmentation results. Moreover, we were able to speed up the network about six times by utilizing BN after the up-convolutional layer. The experimental results on three public benchmark datasets showed high gain in semantic segmentation in relation to state-of-the-art alternatives.

6. Acknowledgment

This work has been partially supported by the Spanish project TIN2016-74946-P (MINECO/FEDER, UE) and CERCA Programme/Generalitat de Catalunya. We gratefully acknowledge the support of NVIDIA Corporation with the donation of the GPU used for this research. This work is partially supported by ICREA under the ICREA Academia programme.

References

- [1] <https://www.kaggle.com/kmader/finding-lungs-in-ct-data>.
- [2] M. Z. Alom, M. Hasan, C. Yakopcic, T. M. Taha, and V. K. Asari. Recurrent residual convolutional neural network based on u-net (r2u-net) for medical image segmentation. *arXiv preprint arXiv:1802.06955*, 2018.
- [3] G. Azzopardi, N. Strisciuglio, M. Vento, and N. Petkov. Trainable cosfire filters for vessel delineation with application to retinal images. *Medical image analysis*, 19(1):46–57, 2015.
- [4] W. Bai, H. Suzuki, C. Qin, G. Tarroni, O. Oktay, P. M. Matthews, and D. Rueckert. Recurrent neural networks for aortic image sequence segmentation with sparse annotations. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 586–594. Springer, 2018.
- [5] H. Chen, Q. Dou, L. Yu, J. Qin, and P.-A. Heng. Voxresnet: Deep voxelwise residual networks for brain segmentation from 3d mr images. *NeuroImage*, 170:446–455, 2018.
- [6] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, 40(4):834–848, 2017.
- [7] Ö. Çiçek, A. Abdulkadir, S. S. Lienkamp, T. Brox, and O. Ronneberger. 3d u-net: learning dense volumetric segmentation from sparse annotation. In *International conference on medical image computing and computer-assisted intervention*, pages 424–432. Springer, 2016.
- [8] N. C. Codella, D. Gutman, M. E. Celebi, B. Helba, M. A. Marchetti, S. W. Dusza, A. Kalloo, K. Liopyris, N. Mishra, H. Kittler, et al. Skin lesion analysis toward melanoma detection: A challenge at the 2017 international symposium on biomedical imaging (isbi), hosted by the international skin imaging collaboration (isic). In *ISBI 2018*, pages 168–172. IEEE, 2018.
- [9] Z. Cui, R. Ke, and Y. Wang. Deep bidirectional and unidirectional lstm recurrent neural network for network-wide traffic speed prediction. *arXiv preprint arXiv:1801.02143*, 2018.
- [10] Z. Cui, J. Yang, and Y. Qiao. Brain mri segmentation with patch-based cnn approach. In *2016 35th Chinese Control Conference (CCC)*, pages 7026–7031. IEEE, 2016.
- [11] M. Drozdzal, E. Vorontsov, G. Chartrand, S. Kadoury, and C. Pal. The importance of skip connections in biomedical image segmentation. In *Deep Learning and Data Labeling for Medical Applications*, pages 179–187. Springer, 2016.
- [12] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017.
- [13] S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. pages 448–456, 2015.
- [14] J. Kleesiek, G. Urban, A. Hubert, D. Schwarz, K. Maier-Hein, M. Bendszus, and A. Biller. Deep mri brain extraction: a 3d convolutional neural network for skull stripping. *NeuroImage*, 129:460–469, 2016.
- [15] Q. Li, B. Feng, L. Xie, P. Liang, H. Zhang, and T. Wang. A cross-modality learning approach for vessel segmentation in retinal images. *IEEE transactions on medical imaging*, 35(1):109–118, 2015.
- [16] P. Liskowski and K. Krawiec. Segmenting retinal blood vessels with deep neural networks. *IEEE transactions on medical imaging*, 35(11):2369–2380, 2016.
- [17] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015.
- [18] F. Milletari, N. Navab, and S.-A. Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *2016 Fourth International Conference on 3D Vision (3DV)*, pages 565–571. IEEE, 2016.
- [19] O. Oktay, J. Schlemper, L. L. Folgoc, M. Lee, M. Heinrich, K. Misawa, K. Mori, S. McDonagh, N. Y. Hammerla, B. Kainz, et al. Attention u-net: Learning where to look for the pancreas. *arXiv preprint arXiv:1804.03999*, 2018.
- [20] P. O. Pinheiro and R. Collobert. Recurrent convolutional neural networks for scene labeling. Technical report, 2014.
- [21] 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*, pages 234–241. Springer, 2015.
- [22] H. R. Roth, L. Lu, A. Farag, H.-C. Shin, J. Liu, E. B. Turkbey, and R. M. Summers. Deeporgan: Multi-level deep convolutional networks for automated pancreas segmentation. In *International conference on medical image computing and computer-assisted intervention*, pages 556–564. Springer, 2015.
- [23] H. Song, W. Wang, S. Zhao, J. Shen, and K.-M. Lam. Pyramid dilated deeper convlstm for video salient object detection. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 715–731, 2018.
- [24] J. Staal, M. D. Abràmoff, M. Niemeijer, M. A. Viergever, and B. Van Ginneken. Ridge-based vessel segmentation in color images of the retina. *IEEE transactions on medical imaging*, 23(4):501–509, 2004.
- [25] F. Visin, M. Ciccone, A. Romero, K. Kastner, K. Cho, Y. Bengio, M. Matteucci, and A. Courville. Reseg: A recurrent neural network-based model for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 41–48, 2016.
- [26] S. Xingjian, Z. Chen, H. Wang, D.-Y. Yeung, W.-K. Wong, and W.-c. Woo. Convolutional lstm network: A machine learning approach for precipitation nowcasting. In *Advances in neural information processing systems*, pages 802–810, 2015.
- [27] X. Zhou, T. Ito, R. Takayama, S. Wang, T. Hara, and H. Fujita. Three-dimensional ct image segmentation by combining 2d fully convolutional network with 3d majority voting. In *Deep Learning and Data Labeling for Medical Applications*, pages 111–120. Springer, 2016.

- [28] Y. Zhou, L. Xie, W. Shen, Y. Wang, E. K. Fishman, and A. L. Yuille. A fixed-point model for pancreas segmentation in abdominal ct scans. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 693–701. Springer, 2017.