# LSTM Multi-modal UNet for Brain Tumor Segmentation

Fan Xu
School of Computer Science and Engineering
Southeast University
Nanjing, China
e-mail: seufanxu@qq.com

Haoyu Ma
School of Computer Science and Engineering
Southeast University
Nanjing, China
e-mail: howiema@seu.edu.cn

Junxiao Sun
School of Computer Science and Engineering
Southeast University
Nanjing, China
e-mail: 934377127@qq.com

Rui Wu
School of Computer Science and Engineering
Southeast University
Nanjing, China
e-mail: RhysWu@outlook.com

Xu Liu
College of Software Engineering Southeast University
Nanjing, China
e-mail: 213150176@seu.edu.cn

Youyong Kong*
School of Computer Science and Engineering
Southeast University
Nanjing, China
e-mail: kongyouyong@seu.edu.cn

***Abstract*—Deep learning models such as convolutional neural network has been widely used in 3D biomedical image segmentation. However, most of them neither consider the correlations between different modalities, nor fully exploit depth information. To better leverage the multi-modalities and depth information, we proposed an architecture for brain tumor segmentation in multi- modal magnetic resonance images (MRI), named LSTM multi- modal UNet. Experiments results on BRATS-2015 show that our method outperforms the state-of-the-art biomedical segmentation approaches.**

***Keywords-multi modalities images; image segmentation; UNet; LSTM; deep learning***

## I. INTRODUCTION

Magnetic resonance imaging (MRI) has been widely used in the study of the structure and function of the human brain in recent years. Thanks to it, significant achievements have been made in the areas of brain disease analysis, diagnosis, treatment and recognizes research, especially for brain tumors. To make better diagnosis and treatment of tumors, segmentation of tumors based on brain MRI is critical.

There are four sequences of images in MRI: T2-weighted fluid attenuated inversion recovery (FLAIR), T1-weighted (T1), T1-weighted contrast-enhanced (T1c), T2-weighted (T2). These four images are often referred to as one modality of MRI, respectively, which can play different roles in the segmentation of tumors. For example, the whole tumor segmentation is better performed with Flair, and the segmentation of tumor core can be better segmented under T2. One of the typical processing methods is early fusion, which combines the modalities on low-level features. This early fusion is based on the assumption that relation between different modalities is simple which is actually complicated [1]. To better learn about the multi-modal information, other researchers put forward late fusion strategy, where each modality is merged with others in a deep layer after an independent CNN. This late fusion strategy is superior to early fusion in brain segmentation [2]. In addition, some studies investigated that the complexity between different modalities cannot be easily modeled by a single layer [1]. A CNN that incorporates dense connections not only between pairs of layers within a single modality, but also between layers across different modalities, can account for the non-linearity in multi-modal data modeling, was proposed in [3].

Recently, with the development of convolutional neural networks (CNN), deep learning has made remarkable achievements in the area of segmentation for brain tumors. some methods use fully convolutional network(FCN) [4] to segment 3D biomedical images. U-Net [5] based on FCN has also shown a good performance in this problem.

However, the segmentation of 3D images often leads to the problem of slow training, because of the large size of 3D images. Patch-based methods [6] take a small region of images into a network and predict the result for each central pixel. Some methods [7] utilize 2D segmentation to 3D biomedical data, which means applying 2D segmentation to each slice of 3D images and concatenating each result to get 3D segmentation. However, those approaches ignore the global structure information or the sequential information between continuous slices. Considering learning the whole information adequately, we utilize convolutional LSTM [8] to better exploit the relationship between consecutive slices.

978-1-7281-2325-7/19/$31.00 ©2019 IEEE

In this paper, we propose a new method, LSTM multi-modal UNet, as shows in Fig 1. Simply put, we utilize 2DUnet to segment the slices of MRI images, and LSTM [8] to learn the sequential information of consecutive slices. Dense connections are used to obtain not only the full characteristics of each modality, but also the complex relation between them.

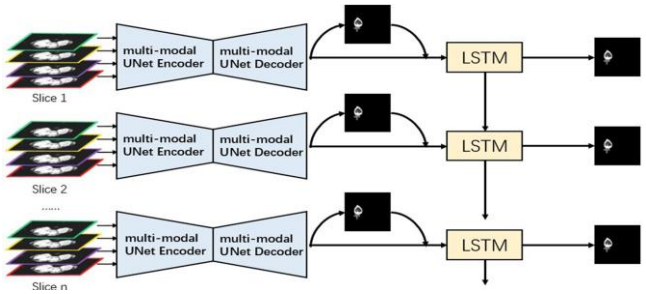


Figure 1. Our proposed LSTM multi-modal U-Net.

## II. METHODS

Our method, LSTM multi-modal UNet, is composed of two parts, 1) multi-modal UNet and 2) convolution LSTM [8]. Multi-modal UNet includes hyper-dense encoder and decoder to fully exploit multi-modal data. Convolutional LSTM further exploits the sequential information between consecutive slices.

### A. Multi-modal UNet

The proposed multi-modal UNet architecture follows the structure of IVD-Net [3].

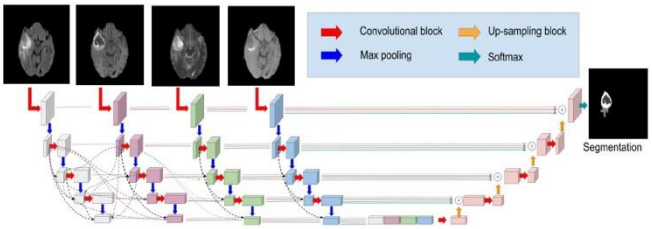


Figure 2. Multi-modal UNet.

*1) Encoder and Decoder:* We adopt the architecture of UNet [5] as our basic encoder and decoder structure. This well-known model is composed of a contracting path and an expansive path. While the former contains multiple convolutions for down sampling to generate high level features, the later has several deconvolution layers to up-sample the features to generate a pixel-wise segmentation. Furthermore, skip-connections, which concatenate the cropped feature maps from the contracting path, are used to transfer information during the compression process. Nevertheless, UNet ignores depth information as it is a 2D segmentation method and not well utilize multi-modal data as it is a pre-fusion strategy.

*2) Multiple encoding paths:* To utility the structure of UNet and achieve the dense connectivity pattern, we use an architecture with multiple UNet encoding paths, each path processing one modality image respectively. The goal of multiple encoding paths is to better account for the complex relationships between multi-modal data and avoid early fusion, which limits the learning capabilities of the network.

*3) Hyper Dense connectivity:* In order to exploit multi- modal information at various levels thoroughly, we adopt the hyper-dense connections methods [9] in the multiple UNet encoding paths network.

Let $x_l$ denote the output of the $l^{th}$ layer, and $H_l$ be a mapping function, which corresponds to the convolution block we proposed. In previous CNNs, the output of the $l^{th}$ layer is typically obtained from the output of the previous layer $x_{l-1}$ as

$$x_l = H_l(x_{l-1}). \tag{1}$$

In a densely-connected network, all feature outputs are concatenated in a feed-forward manner

$$x_l = H_l([x_{l-1}, x_{l-2}, ..., x_0]). \tag{2}$$

where [...] denotes a concatenation operation.

In the present work, the outputs from previous layers in different encoding paths are also concatenated to form the input to the subsequent layers. This structure generates a much more powerful feature representation than early or late fusion strategies in a multi-modal context, as the network has the ability of learning more-complex relationships between the different modalities. For simplicity, assume we have only two modalities.

Let $x_l^1$ and $x_l^2$ denote the outputs of the $l^{th}$ layer in encoding path 1 and 2 respectively. Then the output of the $l^{th}$ layer in a given path $p$ can be defined as

$$x_l^p = H_l^p([x_{l-1}^1, x_{l-1}^2, x_{l-2}^1, x_{l-2}^2, ..., x_0^1, x_0^2]). \tag{3}$$

Furthermore, recent works have found that shuffling and interleaving feature maps in a CNN can improve its performance, since it works as a strong regularizer. Thus we concatenate feature maps in a different sequence for each branch and layer, where the output of the $l^{th}$ layer now becomes

$$x_l^p = H_l^p(F_l^p[x_{l-1}^1, x_{l-1}^2, x_{l-2}^1, x_{l-2}^2, ..., x_0^1, x_0^2]). \tag{4}$$

With $F^p$ being a function that change the order of input feature maps.

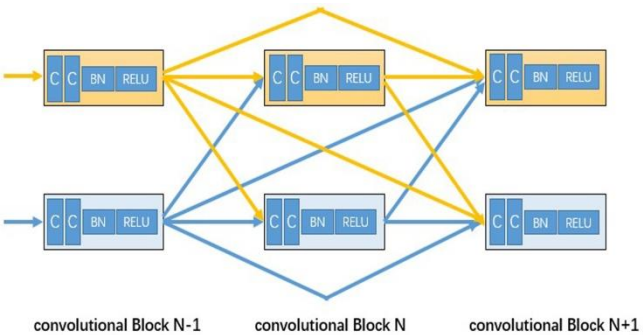


Figure 3. Detail of dense connection.

### B. Slice Sequence Learning

We proposed an end-to-end slice sequence learning model to exploit the sequential dependencies. Image depths are regarded as a sequence of slices. We use convolutional LSTM [8] to model the relationship between slices.

*1) Convolutional LSTM:* Different from traditional LSTM method, convolutional LSTM replace the matrix multiplication by a convolution operator *, which reserves the spatial information for long-term sequences. The overall network is defined as following:

$$i_t = \sigma(x_t * W_{xi} + h_{t-1} * W_{hi} + b_i), \quad (5)$$

$$f_t = \sigma(x_t * W_{xf} + h_{t-1} * W_{hf} + b_f), \quad (6)$$

$$c_t = c_{t-1} \circ f_t + i_t \circ tanh(x_t * W_{xc} + h_{t-1} * W_{hc} + b_c), \quad (7)$$

$$o_t = \sigma(x_t * W_{xo} + h_{t-1} * W_{ho} + b_o), \quad (8)$$

$$h_t = o_t \circ tanh(c_t). \quad (9)$$

*2) Late-LSTM:* As our multi-modal UNet is composed of two parts: encoding path and decoding path, slice dependencies. We add the convLSTM block after the decoding path, named LSTM multi-modal UNet. Since the multi-modal UNet has already fully utility different modalities information, convLSTM network is only supposed to capture sequence dependencies. Furthermore, we add supervise after the encoding path to avoid gradient disappearance.

### C. Relations to Existing Approaches

In 2017, Kuan-Lun also proposed a network with convL- STM and cross-modality convolution [10] to exploit sequence and multi-modal information. However, the encoding path of Kuan-Lun's network follows the strategy of late-fusion. Be- sides, they add the convLSTM block after the encoding path. This strategy need more parameters since the size of channels for the cross-modality convolution is larger. Apparently, our network is different to architecture of the IVD-Net [3]. since IVD-Net is a 2D segmentation method, sequential relationship cannot be utilized. IVD-Net adopts the inception module, which highly increase the parameters of model and memory consumption.

## III. Experiments

We conduct experiments on 3D biomedical image segmen- tation problems and compare our LSTM multi-modal UNet with traditional methods to demonstrate its utility.

### A. Data Set

BRATS-2015 [10] is one of the most challenging 3D segmentation problems since the size and shape of tumor varies. The training dataset comprises of 220 subjects with high grade gliomas (HGG) and 54 subjects with low grade gliomas (LGG). The size of each MRI image is 155 x 240 x240. We randomly choose 224 subjects for training and 50 for testing from all 274 subjects. All brain in the dataset have the same orientation and four modalities (Flair, T1, T1c, T2) are registered. All brain contains five labels: 0 for non-tumor, 1 for necrosis, 2 for edema, 3 for non-enhancing tumor and 4 for enhancing tumor.

### B. Evaluation Metrics

The evaluation criteria are the intersection over union ($IoU$) for each label. $IoU$ is a commonly-used measurement for segmentation performance that calculates the ratio of the area of intersection to the area of unions.

$$IoU = \frac{A \cap B}{A \cup B - A \cap B}. \quad (10)$$

Meanwhile, we separate the tumor structure into three regions as illustrated in BRATS-2015 online judge system.

- Complete score: it means whole tumor areas and mea- sures labels 1, 2, 3, 4.
- Core score: it only considers tumor core region and evaluate label 1, 3, 4.
- Enhancing score: it takes the enhancing core structure into account and evaluate label 4. (HGG only)

There are three kinds of evaluation criteria: Dice similarity coefficient (DSC), Positive Predicted Value (PPV) and Sensi- tivity.

$$Dice = \frac{P_1 \cap T_1}{(P_1 + T_1)/2}. \quad (11)$$

$$PPV = \frac{P_1 \cap T_1}{P_1}. \quad (12)$$

$$Sensitivity = \frac{P_1 \cap T_1}{T_1}. \quad (13)$$

### C. Implement Details

*1) Baseline:* We implement our experiments with Pytorch [11]. Several models are used to demonstrate the advantages of our network. We select UNet with early fusion strategy as baselines. All modalities are merged to build an image with 4 channels. The detail information is in the Table 1.

TABLE I. Detail Information of Network Channels

| | Name | Feat maps(input) | Feat maps(output) |
|---|---|---|---|
| Encoding | Conv layer 1 | 4×240×240 | 64×240×240 |
| | Max pooling 1 | 64×240×240 | 64×120×120 |
| | Conv layer 2 | 64×120×120 | 128×120×120 |
| | Max pooling 2 | 128×120×120 | 128×60×60 |
| | Conv layer | 128×60×60 | 256×60×60 |
| | Max pooling 3 | 256×60×60 | 256×30×30 |
| | ... | | |
| Multi-modal UNet | | | |
| Encoding (each mod) | Conv layer1 | 1×240×240 | 32×240×240 |
| | Max pooling 1 | 32×240×240 | 32×120×120 |
| | Conv layer 2 | 32×120×120 | 64×120×120 |
| | Max pooling 2 | 64×120×120 | 64×60×60 |
| | Conv layer 3 | 64×60×60 | 128×60×60 |
| | Max pooing 3 | 128×60×60 | 128×30×30 |
| | ... | | |

For each convolutional block in UNet, we first use convo- lution layers with kernel size 3x3 to generate feature maps, which are followed by a batch normalization layer and an element-wise rectified-linear non-linearity (ReLU). Since the distributions of tumor and non-tumor tissues are unbalanced, batch normalization layer is vital to training our network. Then a max pooling layer with size 2 and stride 2 is used to down- sample feature maps. All numbers of input and output channels are based on the standard UNet as shows in Table 1.

We also implemented 2D multi-modal UNet as contrast. 2D multi-modal UNet has multiple UNet encoding paths and one decoding path. To prove that more parameters are not the reason for performance improvement of our model, we use half the number of channels for standard UNet.

*2) Our network:* For our proposed architecture, we use the same number of channels of 2D multi-modal UNet as shows in Table 1. As for sequence learning, we set sequence length to 4 initially.

*3) Training:* Since the label distribution of BRATS-2015 dataset is extremely imbalanced. Thus it is easily for modal to converges into local minimum. We use the strategy of median frequency balancing [10] to solve this problem. The weight to each class Wc in the cross-entropy loss function is defined as:

$$W_c = \frac{MedianFreq}{Freq(c)}. \tag{14}$$

where $Freq(c)$ is the frequency of class $c$ in all training pixels. Furthermore, we only sample slices with tumor tissues in training.

We employed Adam [12] optimizer to train all architectures in our experiment, with $\beta_1$=0.9 and $\beta_2$=0.99. Initial learning rate is set to $10^{-4}$. 32 images were used in each mini-batch for all 2D based models and 16 3D volumes for all sequence based network. All MRI images were normalized between 0 and 1 and no data augmentation was employed to improve the performance of all networks.

## IV. Results

The quantitative results of our experiments are reported in Table 2 and Table 3. Also we show the number of parameters of network and model size. We find that our LSTM multi- modal UNet brings a boost in performance compared to the standard U-Net with less model parameters. This result demonstrates that the correlation between each modality and sequence could highly help improve the accuracy of segmentation results based on our network. Furthermore, we provide the qualitative evaluation of our proposed model in Fig. 4.

TABLE II. IOU Results for Each Class

| class | UNet | Multi-modal UNet |
|---|---|---|
| label 0 | 0.9912 | **0.9921** |
| label 1 | **0.1910** | 0.1716 |
| label 2 | 0.3986 | **0.3987** |
| label 3 | 0.1582 | **0.1801** |
| label 4 | **0.3568** | 0.2967 |

TABLE III. Evaluation Criteria of BRATS-2015

| | Network | Complete | Core | Enhancing |
|---|---|---|---|---|
| Dice | UNet | 0.7171 | 0.5989 | **0.5022** |
| | ours | **0.7309** | **0.6235** | 0.4254 |
| Sensitivity | UNet | 0.6116 | 0.5480 | 0.6524 |
| | ours | **0.6376** | **0.5975** | **0.7163** |
| PPV | UNet | **0.9169** | **0.7358** | **0.4871** |
| | ours | 0.8979 | 0.7264 | 0.3860 |

TABLE IV. Compare of Network Size

| | Number of parameters | model size |
|---|---|---|
| U-Net | 34530437 | 138.2MB |
| Ours | 28713450 | 115.6MB |

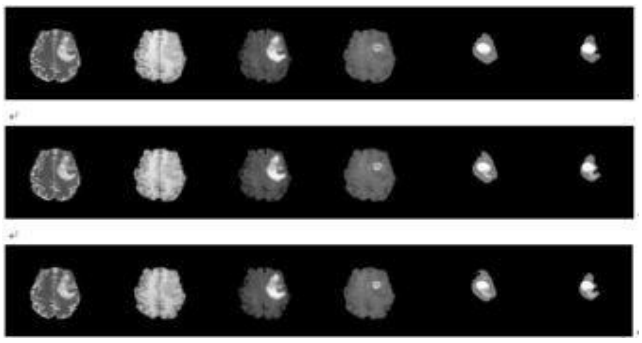

Figure 4. Visualization of each modal image, predict results and ground truth.

## V. Conclusion

In this paper, we proposed a network which combines multi- modal UNet and LSTM to fully exploit the relationships between modalities and the correlation between sequences, named LSTM multi-modal UNet. We trained our model based on BRATS-2015 and it showed a better performance than traditional UNet.

In the future, we will try to learn from UNet++ to reduce the complexity of the model and improve the segmentation speed of a single image. In addition, we also consider ways to improve the algorithm by weakly supervised learning.


## Acknowledgment

This work was supported by the Student Research Training Program from Southeast University under grant 201809006.


## References


[1] N. Srivastava and R. Salakhutdinov, "Multimodal learning with deep boltzmann machines," Journal of Machine Learning Research, vol. 15, pp. 2949–2980, 2014.

[2] D. Nie, L. Wang, Y. Gao, and D. Sken, "Fully convolutional networks for multi-modality isointense infant brain image segmentation," in 13th International Symposium on Biomedical Imaging (ISBI), 2016. IEEE, 2016, pp. 1342–1345.

[3] Dolz J, Desrosiers C, Ayed I B. IVD-Net: Intervertebral Disc Localiza- tion and Segmentation in MRI with a Multi-modal UNet[J]. 2018.

[4] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In CVPR, 2015.

[5] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In MICCAI, 2015.

[6] M. Havaei, A. Davy, D. Warde-Farley, A. Biard, A. Courville, Y. Bengio, C. Pal, P.-M. Jodoin, and H. Larochelle. Brain tumor segmentation with deep neural networks. Medical Image Analysis, 2016.

[7] H. Chen, X. Qi, L. Yu, and P.-A. Heng. Dcan: Deep con- touraware networks for accurate gland segmentation. arXiv preprint arXiv:1604.02677, 2016.

[8] Akilan, Thangarajah, et al. A 3D CNN-LSTM-Based Image-to-Image Foreground Segmentation. IEEE Transactions on Intelligent Transporta- tion Systems, 2019.

[9] Dolz J, Gopinath K, Yuan J, et al. HyperDense-Net: A hyper-densely connected CNN for multi-modal image segmentation[J]. IEEE transac- tions on medical imaging, 2018.

[10] Tseng K L, Lin Y L, Hsu W, et al. Joint Sequence Learning and Cross-Modality Convolution for 3D Biomedical Segmentation[C]// IEEE Conference on Computer Vision & Pattern Recognition. 2017.

[11] Ketkar N. Introduction to PyTorch[M]// Deep Learning with Python. 2017.

[12] Kingma D P, Ba J. Adam: A method for stochastic optimization[J]. arXiv preprint arXiv:1412.6980, 2014.