MDA-NET: MULTI-DIMENSIONAL ATTENTION-BASED NEURAL NETWORK FOR 3D IMAGE SEGMENTATION

Rutu Gandhi

Institute for Artificial Intelligence
University of Georgia, Athens, GA, USA

Yi Hong[*]

Dept. of Computer Science and Engineering
Shanghai Jiao Tong University, Shanghai, China

ABSTRACT

Segmenting an entire 3D image often has high computational complexity and requires large memory consumption; by contrast, performing volumetric segmentation in a slice-by-slice manner is efficient but does not fully leverage the 3D data. To address this challenge, we propose a multi-dimensional attention network (MDA-Net) to efficiently integrate slice-wise, spatial, and channel-wise attention into a U-Net based network, which results in high segmentation accuracy with a low computational cost. We evaluate our model on the MICCAI iSeg and IBSR datasets, and the experimental results demonstrate consistent improvements over existing methods.

***Index Terms*—** Attention network, 3D image segmentation, Squeeze and excitation block

1. INTRODUCTION

Image segmentation is a fundamental task in image understanding, which distinguishes regions of interest from image background for further analysis. Recently, deep segmentation networks, e.g., fully convolutional networks (FCN) [1], U-Net [2], tackle the 2D image segmentation problem and outperform conventional approaches. However, segmenting 3D image volume like brain MRI scans is still challenging, especially in medical image analysis. Models extended from FCNs and U-Nets have been proposed to handle the 3D image segmentation, e.g., V-Net [3]. Due to the high-dimensional nature of image data, most existing models have a high demand for computational resources, especially the GPU memory, and often have a large number of parameters to estimate.

In this paper, we propose an economical solution for segmenting 3D image volume, which roots on a 2D network and handles 2.5D data by augmenting a 2D image slice with an additional image condensed from the third dimension with attention. To integrate the information across slices of an image volume, we propose a compression technique based on the squeeze and excitation (SE) technique [4] to concisely abstract multiple neighboring slices of a volume into a single slice. Apart from the augmented input for the 2D backbone network, we also augment the network with spatial and channel-wise attention by upgrading the concurrent SE block [5]. The resulting network benefits from both the low computation cost of the 2D network and enriched information from image volumes and learned features with attention. Figure 1 depicts the overall architecture of our multi-dimensional attention network (MDA-Net) for segmenting 3D images. The MDA-Net is an end-to-end solution and can automatically learn how to compress volumetric image information and extract useful features in an attention scheme.

Our MDA-Net mimics the process of the manual segmentation for a 3D image. When handling an image volume, we often select one main view to sequentially segment 2D slices and check the third dimension across slices occasionally to obtain additional information. To integrate the information among image slices, we condense the ordered slice difference computed with respect to the current main slice. This compression step allows collecting extra information, i.e., image residuals, to assist the segmentation. Another benefit of using our model is the increased data samples. If we have p slices in one view, we convert one 3D image into at most p samples. Each sample contains a 2D slice and another slice compressed from the original image. This compression is achieved by using slice-wise attention. We also have the spatial and channel-wise attention used in the segmentation network, resulting in our multi-dimensional attention network. The attention is automatically estimated via modified squeeze and excitation (MSE) blocks, which improve segmentation performance over the original SE block and concurrent scSE block.

The main contributions of this paper are as follows:

- Modified SE (MSE) Block: We upgrade the channel attention mechanism in the concurrent scSE and replace the sigmoid function with softmax to ensure the weights for measuring the attention are normalized.
- Slice-wise condensing module: We propose a compression module that uses slice-wise attention to extract residual information in the third dimension.
- Multi-Dimension Attention Network (MDA-Net): We propose an efficient 3D image segmentation network, which fully leverages the 3D data with the balanced computational cost.

[*] yi.hong@sjtu.edu.cn

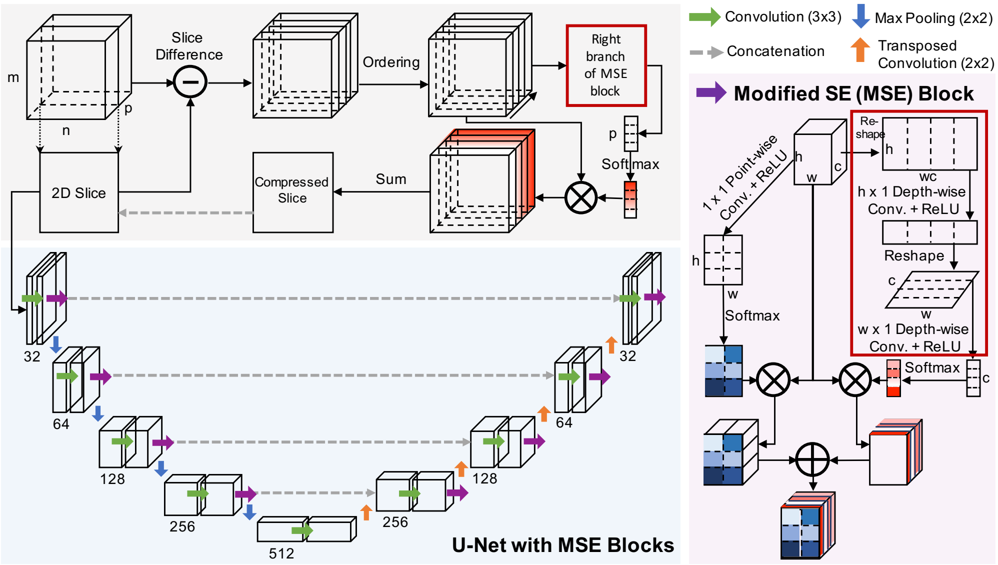

Fig. 1. Overview of our proposed multi-dimensional attention network (MDA-Net) for 3D image segmentation. It consists of a third-dimensional compression and an attention-augmented U-Net with modified squeeze and excitation (MSE) blocks.

We evaluate our MDA-Net on the MICCAI iSeg dataset [6] and the IBSR dataset [7] through segmenting 3D brain scans. The segmentation results on both datasets show the improvement over previous methods with five-fold cross-validation.

Related Work. The 3D variants of the U-Net [3, 8] were proposed to handle volumetric images. Compared to a 2D U-Net working slice by slice, its 3D version fully uses the data in all dimensions. However, 3D models face two main challenges in segmenting the entire high-dimensional image volume. We often have limited computational resources, especially limited GPU memory, to handle the whole image volume. Compared to a greatly increased number of model parameters when switching from a 2D network to 3D, we have a reduced number of data samples since a 2D slice sequence becomes one sample in a 3D U-Net. Existing approaches to address these challenges include downsampling the 3D images to fit in memory [8], assembling multiple 2D networks for accepting different image views [9], working on 3D image patches [10], modifying the existing segmentation architectures [11, 12] or combining 2D slices with 3D patches [13].

Unlike previous approaches, we segment an image volume in the slice-to-slice fashion while integrating the residuals across slices. A work related to ours is the volumetric attention Mask-RCNN [14], which considers three adjacent slices when calculating attention. Our attention model is built based on the SE block, which is not limited to three slices and provides the first-order statistic across slices.

2. MDA-NET

To provide an economical solution for 3D image segmentation, we adhere to treat the 3D data as a sequence of 2D image slices, which allows us to work in a lower-dimensional space with relatively low demand in computational resources. For a specific 2D slice, we augment it with the first-order information in the third dimension, which is achieved by condensing ordered image differences in its neighborhood into one slice. Our model in Fig. 1 has an image dimension reduction component using an ordering-based image difference compression. The resulting slice is concatenated with the associated 2D slice as inputs for an attention-augmented 2D U-Net. We stack the 2D segmentation masks back to form the segmentation mask for an image volume.

2.1. Slice-wise Compression

Assume we segment the i-th image slice, e.g., $I^{(i)}_{m\times n}$, of an image volume $I_{m\times n\times p}$, $i = 1, \cdots, p$. Besides the 2D slice, i.e., $I^{(i)}_{m\times n}$, we prepare an additional slice $\bar{I}^{(i)}_{m\times n}$ that associates with $I^{(i)}_{m\times n}$ and contains information across the slices. The concatenation, $I^{(i)}_{m\times n}\cup\bar{I}^{(i)}_{m\times n}$, is the input of our attention-augmented U-Net described in Section 2.2. This new presentation allows us to take full use of the data in all three dimensions when segmenting one 2D slice of the 3D image volume.

A typical way to compress the data from 3D to 2D is by performing a weighted average over the third dimension. Instead of directly compressing the original image volume, we choose to compress the difference images $I^{(j+i)}_{m\times n} - I^{(i)}_{m\times n}$, $j \in [-r, \cdots, -1, 1, \cdots, r]$. Here, $2r$ (e.g., $r = 5$ chosen experimentally) images in the 2D slice neighborhood are selected. A weighted average of these difference images is the condensed slice $\bar{I}^{(i)}_{m\times n}$, and the weights are estimated using the right branch of our MSE block, as described in the next paragraph. As the center image $I^{(j)}_{m\times n}$ changes, the difference images will change accordingly, and their contributions in the estimated condensed image, which are measured by their weights, vary as well. However, once the network is trained, the weights for the difference images are fixed. To have the weights consistently associated with the difference images, we order them based on the sum of their absolute values. That is, the compressed slice $\bar{I}^{(j)}_{m\times n}$ is the weighted average of the ordered difference images, as shown in Fig. 1.

To estimate the weights for the ordered difference images, we follow the spatial squeeze and channel excitation idea [4] but with some modifications on channel-wise SE block. Firstly, the original SE block squeezes the spatial domain using the global average pooling. That is, the SE block summarizes each channel using its average over pixels with equal weights. The foreground's pixels are often sparse in the difference images, and nearly-zero pixels dominate in the background. Therefore, we need the flexibility in spatial squeeze and adopt convolution filters to average over pixels with learned weights. A simple way to get a weight for each difference image (or a channel used later) is to use a filter of size $m \times n$. To reduce the number of parameters, we decompose this 2D filter into two 1D filters, i.e., one with size $m \times 1$ and the other with size $n \times 1$, and apply depth-wise convolutions. In this way, we can reduce the number of parameters from mn to $m + n$. Since we expect to summarize the information in each channel independently, we need to reshape the input of each depth-wise convolution accordingly, as shown in Fig. 1. As a result, we obtain a vector $z \in \mathbb{R}_p$ as the unnormalized weights for difference images. Also, instead of using Sigmoid in [4], which normalizes each weight independently into the range [0, 1], we choose the Softmax function, which counts the correlation among weights and enforces their sum to be 1. The normalized weights are used to rescale the ordered difference images, and then a weighted average results in the compressed slice $\bar{I}^{(i)}_{m\times n}$.

2.2. 2D U-Net with MSE Blocks

After compressing the residual information in the third dimension, we represent a 3D volume as a pair of 2D slices, i.e., the current 2D slice $I^{(j)}_{m\times n}$ and the slice-wise compression $\bar{I}^{(j)}_{m\times n}$. A 2D U-Net takes our 2D slice pairs as a two-channel input for segmenting the current 2D image. Using a similar approach in the slice-wise compression, which offers us the third dimension's attention, we add MSE blocks to a plain 2D UNet, which provides both spatial and channel-wise attention on its input and extracted feature maps.

Like the concurrent SE block in [5], we add both channel squeeze and spatial excitation branch (sSE) and spatial squeeze and channel excitation branch (cSE), as shown in Fig. 1. But, different from [5], we use the Softmax operator instead of Sigmoid to normalize the spatial or channel-wise weights in both branches. We consider the attention depends on the space and the channels, and the separated weights allow us to treat spatial pixels or channels differently. In particular, the sSE branch uses a pixel-wise convolution to summarize the input feature maps across channels, which is then normalized over the spatial domain using Softmax to ensure the weights' sum is 1. The cSE branch is the same as we discussed in Section 2.1. In particular, given an input feature map $U_{H\times W\times C}$ ($H \times W$ are the spatial size of the feature map and C is the number of channels), we apply depth-wise convolution with an $H \times 1$ filter followed by a $W \times 1$ filter. The resulting $C \times 1$ vector is normalized using Softmax to rescale the channels before taking their average. Reshaping the feature maps from 3D to 2D or from 1D to 2D is required before applying the depth-wise convolutions. The addition of these two branches gives us the output of the MSE block. Each MSE block is used after the convolution pair at each resolution of the U-Net. Worth to mention that, different from the concurrent SE block, we do not use fully-connected layers, resulting in a drop in the number of model parameters.

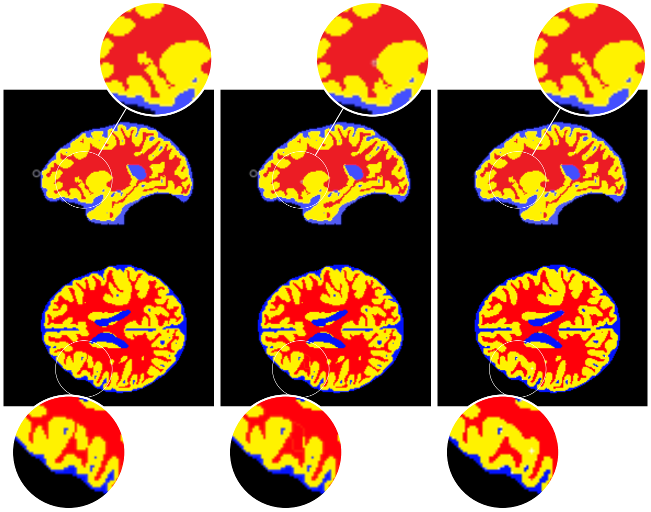

Fig. 2. Qualitative iSeg sample results (from left to right: Ground-Truth, cscSE-UNet, and our MDA-Net. **Blue**: cerebrospinal fluid, yellow: gray matter, **red**: white matter.

3. EXPERIMENTS

3.1. Datasets and Experimental Settings

iSeg Dataset. We evaluate our model on the dataset provided by the MICCAI iSeg challenge [6], which aims to segment the brain MRI scans into the cerebrospinal fluid (CSF), white matter (WM), gray matter (GM), and the background. The dataset consists of T1-weighted brain MRI scans collected from 10 6-month infants. Each scan has an image dimension of $144 \times 256 \times 192$ with a spatial resolution of $1 \times 1 \times 1mm^3$.
IBSR Dataset. We use the IBSR dataset [7] to further demonstrate the performance improvement of our model over other

	Sagittal (%)			Axial (%)			Coronal (%)		
	CSF	GM	WM	CSF	GM	WM	CSF	GM	WM
Plain UNet [2]	91.41±1.12	81.54±1.37	75.13±1.19	95.22±1.01	92.33±1.18	89.51±2.16	92.48±2.34	84.36±1.16	80.24±2.35
cscSE-UNet [5]	92.71±1.74	85.56±1.14	79.14±1.11	96.47±1.01	92.38±1.18	89.83±2.13	92.11±1.45	87.12±1.09	82.15±1.42
MSE-UNet (Ours)	93.11±1.34	87.21±1.54	81.13±2.34	96.27±1.01	92.62±1.18	89.89±2.13	93.91±2.34	88.25±2.16	84.23±2.65
MDA-Net (Ours)	**93.46±1.05**	**87.23±2.64**	**82.79±2.55**	**96.81±1.15**	**94.32±2.13**	**90.01±1.24**	**94.28±2.01**	**88.44±1.45**	**85.23±1.67**

Table 1. Segmentation comparison (measured in Dice score) among different approaches applied on the MICCAI iSeg dataset.

	Sagittal (%)			Axial (%)			Coronal (%)		
	CSF	GM	WM	CSF	GM	WM	CSF	GM	WM
Plain UNet [2]	90.81±1.85	95.91±2.24	92.92±1.55	91.24±1.15	90.33±2.16	88.38±1.84	91.45±2.08	87.30±1.46	89.22±2.67
cscSE-UNet [5]	92.41±1.15	96.06±2.14	93.61±2.55	91.42±2.15	91.17±2.10	88.81±1.04	92.15±1.01	88.12±1.95	89.82±1.27
MSE-UNet (Ours)	94.01±1.95	96.96±2.14	94.79±3.55	92.23±1.25	92.65±2.14	89.82±1.64	93.93±2.01	89.21±1.15	90.58±1.60
MDA-Net (Ours)	**95.12±1.00**	**97.27±2.14**	**95.04±1.15**	**94.89±1.05**	**95.31±2.73**	**92.01±2.24**	**94.52±2.31**	**91.45±1.35**	**92.22±1.67**

Table 2. Segmentation comparison (measured in Dice score) among different approaches applied on the IBSR dataset.

approaches. The MR images provided in this dataset are T1-weighted three-dimensional coronal brain scans preprocessed with a positional normalization and a skull-stripping step. It contains 18 subjects with an image dimension of $512 \times 128 \times 256$ and a spatial resolution of 1.5mm in each dimension.

Ablation Study. The backbone network of our method is the 2D U-Net [2], which is augmented by our MSE blocks. In the ablation study, we compare our MSE-UNet with both the plain UNet and the concurrent scSE-UNet (cscSE-UNet) [5]. These three models take only a single 2D slice of a 3D image scan to perform the 3D segmentation task in a slice-by-slice manner. Our whole model (MDA-UNet) is also reported to demonstrate the performance gain by including the slice-wise compression component into the MSE-UNet.

We examine three views for both datasets, i.e., sagittal, axial, and coronal, respectively. Take the iSeg dataset for instance, we have 1152 samples (including a 2D slice and a sequence of difference images) in the sagittal view, 2048 samples in the axial view, and 1536 samples in the coronal view.

We implement our model using Keras and the Tensorflow backend. We use Adam optimizer with a learning rate of 5e-5, the dropout of 0.3, and the L2 regularizer. The Dice score is used during training and evaluation. The maximum number of epochs is 300. All models were trained on one NVIDIA GeForce 1080 8GB GPU. We use five-fold cross-validation on the subjects for both datasets.

3.2. Experimental Results

Tables 1 and 2 report the segmentation results for iSeg and IBSR datasets, respectively. The brain segmentation performance for both datasets is steadily improved from plain U-Net, to cscSE-UNet, to MSE-UNet, and then to MDA-Net for each image view. The improvement of MSE-UNet over cscSE-UNet demonstrates the effectiveness of using MSE-Blocks in U-Net, and the improvement of MDA-Net over MSE-UNet demonstrates the effectiveness of using the slice-wise compression. Also, replacing Sigmoid with Softmax improves the stability and convergence of the network training. Figure 2 shows a qualitative comparison for the iSeg dataset.

	#Param.	Training time (per epoch)	Inference time (per patient)
Plain UNet	7.775M	30s	1.565s
cscSE-UNet	7.958M	59s	1.575s
MSE-UNet	7,823M	57s	2.285s
MDA-Net	7.825M	58s	2.795s

Table 3. Model comparison on the number of model parameters and the training and testing time for the iSeg dataset.

Table 3 reports the computational cost of the above four models. Compared to cscSE-UNet, our models have a reduced number of parameters and a reduced amount of training time but slightly increased inference time. Our maximum training time was around six hours, and the inference time is within seconds for a subject. We tried experiments with a 3D U-Net and a convolutional LSTM; however, we met memory difficulties on our machine, which motivates the MDA-Net.

4. CONCLUSION AND DISCUSSION

This paper investigated the 3D image segmentation problem and proposed an efficient solution using a multi-dimensional attention network. We tested the network on image volumes; however, it could be extended to handle the spatiotemporal data like videos or longitudinal images. We will apply our model to other segmentation tasks with different image types in the future. For multi-modality image scans, we could explore image-wise attention, that is, measuring the contributions of each modality to the segmentation task.

Compliance with Ethical Standards

This is a numerical simulation study for which no ethical approval was required.

Acknowledgements

No funding was received for conducting this study. The authors have no relevant financial or non-financial interests to disclose.

5. REFERENCES

[1] Jonathan Long, Evan Shelhamer, and Trevor Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3431–3440.

[2] Olaf Ronneberger, Philipp Fischer, and Thomas Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241.

[3] Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi, "V-net: Fully convolutional neural networks for volumetric medical image segmentation," in *2016 Fourth International Conference on 3D Vision (3DV)*. IEEE, 2016, pp. 565–571.

[4] Jie Hu, Li Shen, and Gang Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7132–7141.

[5] Abhijit Guha Roy, Nassir Navab, and Christian Wachinger, "Concurrent spatial and channel 'squeeze & excitation'in fully convolutional networks," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2018, pp. 421–429.

[6] Li Wang, Dong Nie, Guannan Li, Élodie Puybareau, Jose Dolz, Qian Zhang, Fan Wang, Jing Xia, Zhengwang Wu, Jia-Wei Chen, et al., "Benchmark on automatic six-month-old infant brain segmentation algorithms: the iseg-2017 challenge," *IEEE transactions on medical imaging*, vol. 38, no. 9, pp. 2219–2230, 2019.

[7] Torsten Rohlfing, "Image similarity and tissue overlaps as surrogates for image registration accuracy: widely used but unreliable," *IEEE transactions on medical imaging*, vol. 31, no. 2, pp. 153–163, 2011.

[8] Özgün Çiçek, Ahmed Abdulkadir, Soeren S Lienkamp, Thomas Brox, and Olaf Ronneberger, "3d u-net: learning dense volumetric segmentation from sparse annotation," in *International conference on medical image computing and computer-assisted intervention*. Springer, 2016, pp. 424–432.

[9] Grzegorz Chlebus, Andrea Schenk, Jan Hendrik Moltz, Bram van Ginneken, Horst Karl Hahn, and Hans Meine, "Automatic liver tumor segmentation in ct with fully convolutional neural networks and object-based post-processing," *Scientific reports*, vol. 8, no. 1, pp. 1–7, 2018.

[10] Zhijian Liu, Haotian Tang, Yujun Lin, and Song Han, "Point-voxel cnn for efficient 3d deep learning," in *Advances in Neural Information Processing Systems*, 2019, pp. 965–975.

[11] Christian Lucas, André Kemmling, Amir Madany Mamlouk, and Mattias P Heinrich, "Multi-scale neural network for automatic segmentation of ischemic strokes on acute perfusion images," in *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*. IEEE, 2018, pp. 1118–1121.

[12] Robin Brügger, Christian F Baumgartner, and Ender Konukoglu, "A partially reversible u-net for memory-efficient volumetric image segmentation," in *International conference on medical image computing and computer-assisted intervention*. Springer, 2019, pp. 429–437.

[13] Raunak Dey and Yi Hong, "Hybrid cascaded neural network for liver lesion segmentation," *arXiv preprint arXiv:1909.04797*, 2019.

[14] Xudong Wang, Shizhong Han, Yunqiang Chen, Dashan Gao, and Nuno Vasconcelos, "Volumetric attention for 3d medical image segmentation and detection," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2019, pp. 175–184.