The Evolution of U-Net Architectures in Medical Image Segmentation

Yanzhang Gu *

School of Computer Science and Technology , Donghua University, Shanghai, 201620, China

Abstract: As an important field of deep learning, image segmentation has developed rapidly in recent years. Medical image segmentation has always been an important application scenario of image segmentation, and many models with excellent performance have emerged. U-Net has become the focus of research and application recently because of its easy to understand structure and excellent performance. This paper focuses on the structure of U-Net itself and explains the reasons for U-Net's excellent performance as well as some defects. At the same time, this paper also introduces the improvements and adjustments made by different researchers to solve the problems encountered in the practical application of U-Net, including structural improvements, such as the adjustment of modules and the replacement of convolution methods, and non-structural improvements, such as the optimization of data sets and the improvement of loss functions. Finally, the prospects and suggestions for the future development and application of U-Net were put forward.

1. Introduction

In order to facilitate the diagnosis and treatment of diseases, medical imaging involves the use of a variety of imaging techniques to provide visual representations of the interior structures and functions of the body. The process of splitting an image into many different sections with unique features and identifying the target of interest is known as medical image segmentation. It is essential to clinical diagnosis, medicinal intervention, and quantitative analysis. Rapid advancements in medical imaging technology, including computed tomography (CT), magnetic resonance imaging (MRI), ultrasound, and X-rays, have made it possible for radiologists and internists to provide precise and timely diagnoses in recent decades.

Although these technologies have greatly improved the accuracy of manual diagnoses, they still heavily rely on the expertise of doctors and radiology specialists. Meanwhile, as the volume of medical image data increases, the workload for doctors has significantly risen, leading to an urgent need for automated image segmentation methods in clinical settings to assist physicians in quick diagnoses, enhance work efficiency, and reduce the burden on doctors—this is also a problem researchers have been striving to solve [1].

The U-Net architecture was introduced by Ronneberger and colleagues in 2015. This model is built upon the Fully Convolutional Network (FCN) framework, sharing a similar

* Corresponding author: 221310229@mail.dhu.edu.cn

structure that includes an encoder-decoder topology along with skip connections. This design enables more accurate segmentation even when trained on a limited dataset. Unlike FCN, U-Net features a symmetrical layout; the left side serves as the contraction path to gather contextual information, while the right side functions as the expansion path for precise localization and image size restoration. Each layer's output feature maps from the encoder are duplicated, cropped, and combined with corresponding decoder feature maps after deconvolution before being fed into subsequent layers for further upsampling. The U-Net model maintains numerous feature channels during this upsampling process, facilitating the effective transmission of context information to higher-resolution layers [2].

This paper discusses the application of U-net in the field of medical image segmentation, the difficulties encountered, and corresponding improvement. In particular, on the basis of the original module, the replacement of classical network modules such as residual module, Dense module, Inception module, Attention module, and convolution in submodules brings about the effect.

2. Overview of U-net's framework and challenges

A multi-channel feature map is represented by each of the blue boxes. On top of the box is a note that indicates the number of channels. The box's lower left border displays the x-y-size. Copies of feature maps are shown in white boxes. The various operations are shown by the arrows.

U-Net is a deep learning model specifically designed for image segmentation, focusing on the effective fusion of fine-grained features and contextual information to achieve high-quality segmentation results. The proposed architecture is particularly suitable for critical tasks that require precise localization and is highly relevant in the field of medical image analysis. By establishing skip connections between encoder and decoder paths, U-Net preserves important spatial information and significantly improves segmentation accuracy in applications such as tumor detection and organ delineation.

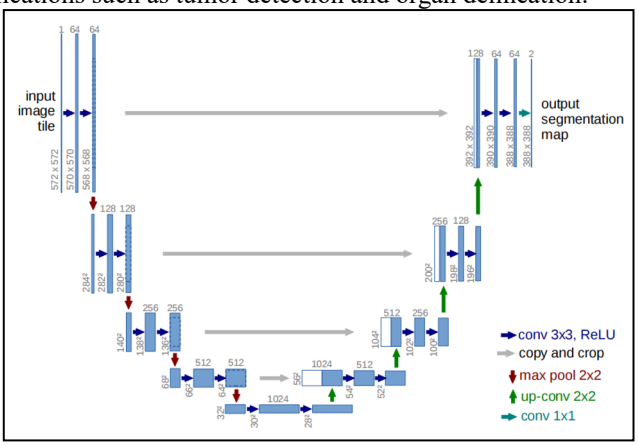

Fig. 1. U-net design (lowest resolution example for 32x32 pixels) (Photo/Picture credit : Original).

A contraction path and an extension path make up the U-Net design, as seen in Fig. 1. The shrinking approach does 3x3 convolutions repeatedly, activates ReLU and then applies 2x2 Max pooling to accomplish downsampling. The number of feature channels is doubled at each downsampling stage. The extended approach gradually decreases the feature channels by using a 2x2 convolution ("up convolution") after upsampling. In the shrinking route, this is paired with the matching feature maps that have been trimmed. The final result

maps features to the required number of categories using 1x1 convolutions. There are twenty-three convolutional layers in the network [3].

While in the field of biomedical image, the shortage exists in the amount of data for one specific single task. However, the amount of data is considered as the more important factor rather than the model that affects the performance [4]. Before U-net, Ciresan's model [5] took time and had a shortage of accuracy. Thus Ronneberger, the developer of U-Net tries to develop a new model that requires fewer training images and achieves more precise segmentation. Firstly, the pooling operator is replaced by an upsampling operator since pooling operators will only keep the features and drop other information that might be useful in the following steps. This replacement improves the accuracy of the model. Secondly, to solve the shortage of training images, elastic deformations are applied to the dataset without annotating deformations to achieve data augmentation. Because in biomedical segmentation, the most prevalent variable in tissue is deformation. Additionally, Dosovitskiy et al. [6]have demonstrated the significance of data augmentation for learning invariance in the scope of unsupervised feature learning. What's more, a weighted loss is applied to separate the touching objects of the same class better. Through these improvements, U-Net has achieved a great performance in 2017.

However, U-Net also has shortages. Firstly, medical image segmentation requires precise identification and delineation of the structure or tissue to which each pixel belongs, necessitating a deeper understanding of the complex relationships between local details and overall structures. However, the limited receptive field of convolutional kernels makes it difficult to capture long-range relationships and global information, thereby restricting further enhancements in model performance. Secondly, U-Net and its variants typically require substantial computational resources due to their deep network architecture and skip connections, especially when handling large images or performing three-dimensional image segmentation. For three-dimensional images, larger volume leads to greatly increased memory and computational demands and U-Net can only capture planar features while 3D images contain large numbers of features in a volumetric context. This high demand for resources limits the applicability of U-Net. Thirdly, for high-resolution images, U-Net may need to employ techniques such as patch processing or downsampling, which can lead to loss of detailed information due to increased complexity in processing [7].

3. Related improvements for U-Net

In order to solve problems like high demands for hardware and limits in performance of capturing long-range relationships and global information, researchers have made numerous attempts and improvements based on the fundamental architecture of U-Net, primarily including structural and non-structural improvements to accommodate a broader range of application needs and challenges.

3.1 Structural improvements

Structural improvements include improvements in encoding and decoding, skip connections, and overall structure. Most improvement works on the basis of the original module, adding the attention module, inception module, dense module, residual module, and other classic network modules. The residual module and dense module work similarly by outputting shallow layers' results to the deeper ones to avoid gradient vanishing and explosion. Guan et al. [8] proposed a new model FD-UNet by adding dense modules. Compared with the original model, this model is more compact, has enhanced artifact removal ability, and performs well in image quality improvement. Xiao et al. replaced the sub-module with a residual module to improve accuracy in the retinal vessel segmentation problem [9]. A

densely convolutional network (DenseNet) based on ResNet was proposed by Huang et al. By joining the output and input in parallel, the gradient disappearance can be effectively alleviated, and the reusability of features can be increased. [10]. The attention module assigns weight to the features and further and emphatically selects the effective features and suppresses the irrelevant features. Oktay et al. [11] proposed an Attention U-Net which uses a raster-based attention gate(AG) [12] to filter the output results and eliminate the noise and irrelevant information in the process of jump connection. Inception module improves segmentation accuracy by using convolution kernels of different scales for multi-scale feature learning. Ibtehaz et al. [13] were initiated by the Inception module, and on the basis of the original Inception module, the convolutional kernel was decomposed, and then the outputs of different convolutional cores were spliced together to obtain spatial information of different scales. The use of the Inception module enables the network to reduce the number of parameters and reduce the memory requirement before ensuring the receptive field. Jiang et al. [14] increased the network's breadth and depth by substituting the inception module for the original convolutional module in U-Net. In order to create receptive fields of varying sizes that may extract features of various scales and achieve improved segmentation accuracy, the inception module parallelly joins convolution kernels of various sizes.

Some researchers also proposed new models by changing convolution modes like deformable convolution and dilated convolutionIn order to increase segmentation accuracy, Jin et al. [15] suggested DUNet and substituted deformable convolutions for the original convolutions in the encoder and decoder. The introduction of deformable convolution brings the adjustment of receptive field size and improves the generalization ability. Chen et al. [16] introduced dilated convolution into the proposed DMFNet network. Dilated convolution allows the network could actively learn the most valuable information from different receptive fields, thereby improving the segmentation accuracy.

3.2 Non-structural improvement

Non-structural improvement is another reason that U-Net achieves excellent performance in medical image segmentation. These improvements usually involve data processing techniques, training strategies, loss function design, etc., aiming to improve the model performance and accelerate the training process. For example, Yang Xin et al. [17] adopted randomization of images in order to overcome the possible overfitting of neural networks. Enhancement techniques such as shearing, flipping, grayscale perturbation, and shape perturbation are used to increase the amount of data by simulating the situation encountered in real images. Wang et al. [18]designed a new mixed loss function to solve the imbalance problem in brain tumor MRI image samples. Using the dice coefficient alone as the loss function will make the features of smaller objects difficult to segment correctly. Focal loss reduces the weight of easy examples and makes the model focus more on learning difficult examples, which is suitable for highly imbalanced scenarios like brain tumors. As a result, a combination of the dice loss function and focus loss function makes the model perform better.

4. Conclusion

This article mainly discusses both structural improvement and non-structural improvement in U-Net. It is evident that the network structure is the primary focus of the U-Net model's advancement, with the most often inserted modules being the residual, dense, inception, and attention modules. These modules can increase segmentation accuracy, fully use context information, and enable the network to extract features efficiently. In addition, the

effectiveness of non-structural improvements to improve network performance has gradually attracted attention. While improving network structure, some researchers fully consider the comprehensive application of data augmentation and data normalization methods and propose some new hybrid loss functions to train the network in a sequential manner, thereby improving the performance of the network.

However, U-Net still faces many tasks. For example, there is not enough pertinent medical imaging data, which can easily result in overfitting. Various improvements have been made to the UNet infrastructure in recent years but a lack of truly groundbreaking improvements; The interpretability of related deep learning needs to be strengthened, and the promotion and application of computer-aided diagnosis and treatment systems need to be improved. By resolving these issues, U-Net should encourage ongoing innovation and discoveries in the field of medical image segmentation and raise the bar for medicine.

References

1. Y. LeCun, Y. Bengio, G. Hinton, Deep learning. Nature, **521**, 436-444 (2015).
2. H. Zhang, D. Qiu, Y. Feng, J. Liu, Improved U-Net models and its applications in medical image segmentation: A review. Laser & Optoelectronics Progress, **59**, 0200005 (2022)
3. O. Ronneberger, P. Fischer, T. Brox, U-Net: Convolutional networks for biomedical image segmentation. In: MICCAIx, 234–241 (2015)
4. A. Halevy, P. Norvig, F. Pereira, The unreasonable effectiveness of data. IEEE Intelligent Systems, **24**(2), 8-12 (2009)
5. D.C. Ciresan, L.M. Gambardella, A. Giusti, J. Schmidhuber, Deep neural networks segment neuronal membranes in electron microscopy images. In: NIPS , 2852–2860 (2012)
6. A. Dosovitskiy, J.T. Springenberg, M. Riedmiller, T. Brox, Discriminative unsupervised feature learning with convolutional neural networks. In: NIPS (2014)
7. Y. Yin, J. Ma, W. Zhang, A., From U-Net to transformer: Progress in the application of hybrid models in medical segmentation. Laser & Optoelectronics Progress, 62, 01 (2024)
8. S. Guan, A.A. Khan, S. Sikdar, P.V. Chitnis, Fully Dense U-Net for 2-D sparse photoacoustic tomography artifact removal. IEEE J. Biomed. Health Inform. **24**, 568-576 (2020)
9. X. Xiao, L. Shen, Z. Luo, S. Li, Weighted Res-U-Net for high-quality retina vessel segmentation. In: 2018 9th International Conference on Information Technology in Medicine and Education (ITME), 327-331 (2018)
10. H. Huang, L. Lin, R. Tong, et al., U-Net 3+: A full-scale connected U-Net for medical image segmentation. In: ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 1055-1059 (2020)
11. O. Oktay, J. Schlemper, L.L. Folgoc, et al., Attention U-Net: Learning where to look for the pancreas. arXiv preprint arXiv:1804.03999 (2018)
12. J. Zhang, Z. Jiang, J. Dong, et al., Attention gate ResU-Net for automatic MRI brain tumor segmentation. IEEE Access, **8**, 58533-58545 (2020)
13. N. Ibtehaz, M.S. Rahman, MultiResUNet: Rethinking the U-Net architecture for multimodal biomedical image segmentation. Neural Networks, **121**, 74-87 (2020)

14. Y. Jiang, M. Ye, P. Wang, D. Huang, X. Lu, MRF-IUNet: A multiresolution fusion brain tumor segmentation network based on improved Inception U-Net. Comput. Math. Methods Med. 1-8 (2022)
15. Q.G. Jin, Z.P. Meng, T.D. Pham, et al., DUNet: A deformable network for retinal vessel segmentation. Knowl. Based Syst., **178**, 149-156 (2019)
16. C. Chen, X.P. Liu, M. Ding, et al., 3D dilated multifiber network for real-time brain tumor segmentation in MRI. In: Shen D.G., Liu T.M., Peters T.M., et al. (Eds.), Medical Image Computing and Computer Assisted Intervention - MICCAI 2019, Lecture Notes in Computer Science (Cham: Springer, 2019), **11766**, 184-192 (2019)
17. X. Yang, X.Y. Li, X.T. Zhang, et al., Automatic segmentation method of organs threatened by radiotherapy for nasopharyngeal carcinoma based on adaptive U-Net network. J. Southern Med. Univ. **40**(11), 8 (2020)
18. S.Q. Wang, Research on MRI image segmentation of brain tumor based on improved U-Net. Master Thesis, Changchun University of Technology (2023)