Vision Mamba-based UNet++ Approach for Effective Medical Image Segmentataion

Rupak Chakraborty[1], *Shashi Kant Gupta*[2],
[1] Postdoctoral Researcher, LINCOLN UNIVERSITY COLLEGE, Malaysia; [2]Adjunct Professor, LINCOLN UNIVERSITY COLLEGE, Malaysia
Email ID pdf.rupak@lincoln.edu.my, rupak.jis@gmail.com

Abstract: In the field of medical image segmentation, UNet and UNet++ approaches got attention. Though the performance of modified architecture of those models was promising, they got stuck in some areas like inaccurate boundary delineation, sensitivity to image quality variance etc. So, researchers moved towards Vision Transformer (ViT) architecture where self-attention mechanisms were applied on a set of patches of images. But these pre-trained transformer-based models suffer from high complexity and heavy power support. To overcome those issues, State Space Model (SSM)-based Vision Mamba (VM) architecture has been introduced. Incorporation of bidirectional sequence feature of SSM to encoder-decoder based UNet architecture is the recent research interest. Inspired by existing literature, one Vision Mamba UNet++ (VMUNet++) has been proposed for effective segmentation. The nested 'skip-connections' of UNet++ has been chosen here to reduce the semantic gap between encoding and decoding features. The proposed model will be tested to public datasets like ISIC18, Synapse, SegPC-2021 and the outcome of the model will be compared with existing recent architectures to show the efficacy of the model.
Keywords: UNet++; ViT; SSM; VM-UNet++.

Introduction

Accurate segmentation of MRI images is a crucial task in neuroimaging and cancer diagnostics. Traditional manual annotation is time-consuming and inaccurate. Recent developments in convolutional neural networks (CNNs) and transformer-based models have undoubtedly improved segmentation performance. However, existing models often lack interpretability and robustness to noise, scanner variability, and limited data conditions. This paper addresses these gaps and proposes a hybrid model combining CNNs for spatial detail and Vision Transformers for global context understanding. The research also explores explainable AI to visualize decision-making regions and uncertainty estimation to quantify model confidence. The evolution of proposed architecture starting from CNN-based U-Net model has been presented in next section.

Related work

This section will start to investigate the work done on medical image segmentation using U-Net++, Transformer-based and SSM-based models for accurate segmentation.

U-Net++ based models for image segmentation

UNet++ is a nested U-Net architecture which helps to reduce the semantic gap between encoding and decoding features of U-Net. Here nested skip-pathways are used to gain knowledge from both low- and high-level features at decoder end from encoder side which ultimate results in the detailed and comprehensive understanding of the image. Here, optimizers deal with an easier learning task as they need to handle the almost similar and semantical feature maps [1-2]. Xu et. al. [3] proposed an algorithm based on UNet++ where augmentation and dropout steps were added to segment low-grade glioma MRI image. Li. et. al. [4] added threshold parameters to identify and remove less important decoder blocks to compress the network architecture. In addition, UNet++ architecture has been modified by incorporating the Residual-Attention mechanism16 to improve the network degradation problem and focus on the area of the target for the segmentation task. Li et. al. [5] proposed a Channel-Attention-based UNet++ architecture where channel and attention modules had been applied to prevent the loses of eigenvalues in the long-distance skip connection process. Some limitations of this architecture have been noticed like susceptibility to class imbalance, sensitivity to image quality variations, and potential for inaccurate boundary delineation. So, researchers considered the transformer-based architecture in medical image segmentation domain and compared the outcomes.

Transformer-based models for image segmentation

After continuous success of UNet and UNet++ models on medical image segmentation, researchers chose Vision Transformer (ViT) architecture to remove the limitations like susceptibility to class imbalance, sensitivity to image quality variations, and potential for inaccurate boundary delineation of earlier used models. In ViT, images are represented as a set of patches which are non-overlapping blocks of image. These blocks are formed by vector embeddings of pixel information. ViT is built on transformer architecture where self-attention mechanisms are applied on patches to form relationships between blocks. Cao et. al. [6] proposed a Swin-Transformer-based U-shaped architecture with encoder-docoder where skip-connections helped to learn local-global semantic features. Chen et. al proposed TransUNet [7] architecture where tokenized image patches were encoded by the Transformer for extracting global contexts whereas the decoders were used to upsample the encoded features. One effective transformer-based approach known as MissFormer [8] had been proposed to explore global dependencies for better feature discrimination of 2D Medical Image. Literature survey continued and noticed that ViT-based models suffer with some drawbacks and need some upgradation. Next subsection pays attention to it.

SSM-based Model for Image Segmentation

State Space Model (SSM)-based Vision Mamba (VM) architecture draws attention to researchers in medical image field. SSMs are used to handle sequences by modeling the hidden states over time26. Vision Mamba extends these models to visual data, incorporating bidirectional sequence modeling. The main advantages of vision Mamba over ViT architecture are unlike ViT, Mamba is more computationally efficient, suitable for high-resolution images, perform well with limited computational resources and work as bidirectional sequence modeling. Ruan et. al. proposed VM-UNet architecture combined with SSM and isual State Space (VSS) block to gain extensive contextual information with a smaller number of convolution layers [9]. Researchers like Wu et. al. even did some improvement over this architecture

where the extended the 2D-selective-scan (SS2D) of SSM and proposed a high-order Vision Mamba UNet (H-vmunet) model for better segmentation [10]. this VM-UNet had been further modified to VM-UNet++ where nested skip-connections of U-Net++ model had been applied to up-sample the decoding results more accurately. This approach is now-a-days an interest of research, so inspired from this literature we are proposing an improved VM-UNet++ model for accurate and improved segmentation over state-of-the-art approaches.

Methodology

From the literature review section, it has been shown that VM-based architecture may overcome the drawback the transformer-based models. Combining VM architecture with UNet++ may produce the best segmentation results. So, in the next subsection VM-UNet++ architecture along with XAI has been discussed.

VSS Block with SS2D Architecture

The architecture steps of SS2D have been shown in the following diagram Figure 1.

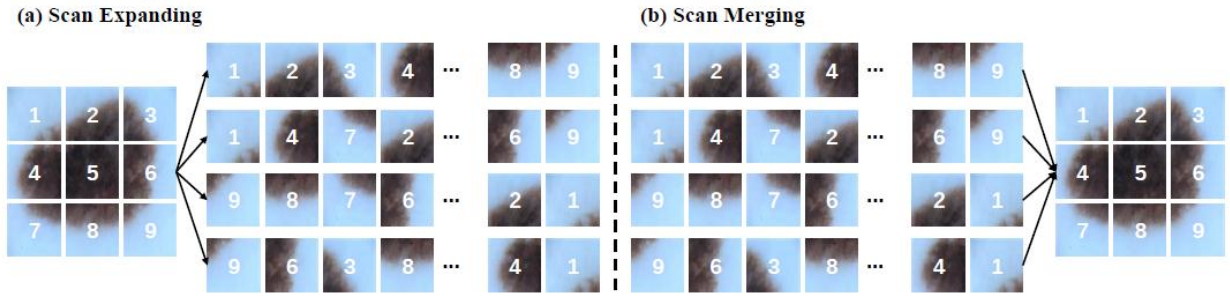

Figure 1 Proposed SS2D architecture with VSS block

Proposed VM-UNet++ model with XAI for visualizing results

VM-UNet passes through the steps like skip connections, Patch Embedding layer, an encoder, a decoder, a Final Projection layer. The Patch Embedding layer is responsible for dividing the input image into non-overlapping patches of fixed size. The resultant embedded image is normalized using Layer Normalization before feeding it into the encoder for extraction of features. Similarly, the decoder is organized into four stages where in last three stages, a patch expanding operation is carried out to decrease the number of feature channels so that the height and width can be increased. At the decoder end, patch expanding is carried out using the concept of 'skip-connections' of U-Net architecture. Now in continuation with that, instead of single skip connection, we are going to use nested skip-connections of U-Net++ model that will help to up-sample the decoding results more accurately.

Databases for conducting research

The below mentioned publicly available datasets below have been chosen for the experiment.

- **ISIC 2017 Dataset: -** https://challenge.isic-archive.com/data/#2017
- **ISIC 2018 Dataset:** https://challenge.isic-archive.com/data/#2018

This dataset contains 2594 images and 12970 corresponding ground truth response masks (5 for each image).

- **CVC-ClinicDB :-**https://www.kaggle.com/datasets/balraj98/cvcclinicdb
- **Synapse Dataset: -**https://www.synapse.org/Synapse:syn3193805/wiki/217753

Review and Analyze Results

- The result of the recent research experiment has been presented here on ISIC 2018 dataset.
- The result shows the effectiveness of VM-UNet++ architecture in terms of evaluation parameters. As a limitation, it can be claimed that researchers yet to provide sufficient number of research papers based on Vision Mamba UNet++ architecture.
- XAI integration with VM-UNet++ is still under working phase.

Table 1 Results of different approaches on ISIC2018 Dataset

Methods	mIoU%	DSC%	SE%	SP%	Acc%
VM-UNet++	**80.27**	**89.38**	**88.59**	**97.67**	**96.62**
VM-UNet	79.88	88.98	88.25	97.11	96.12
AttUNet	79.29	88.71	87.88	96.78	95.79
ResUNet	78.96	87.88	87.47	96.22	95.06
MultiRUNet	78.56	87.66	86.99	95.66	94.39
TransUNet	77.97	86.55	86.66	95.01	94.02
MissFormer	77.63	86.30	85.76	94.76	93.59
UCTransNet	76.99	85.77	85.42	93.88	93.20

Statistical results in Table 1 have shown that the proposed VM-UNet++ architecture overcome the challenges of existing state-of-art transformer based-approaches on ISIC 2018 dataset.

Conclusion

This paper surveys various segmentation techniques of medical data. Starting from UNet++ architecture to its advanced versions have been surveyed here. The drawback of UNet++ model has been overcome by replacing transformer-based approaches. Survey continued and found some major drawbacks of transformer-based models like computational cost, complexities, high resource power etc. So, 2D-Selective-Scan algorithm-based VSS block has been found which effectively did segmentation with the help of Vision Mamba (VM) architecture. The working steps of VM-UNet++ have been discussed in the proposed methodology. After applying the proposed method on ISIC 2018 dataset, the statistical and visual both the outcomes have been presented in this paper. The future direction states that the success of the proposed approach will be tested on known popular public datasets like Synapse, SegPC-2021 and CVC-ClinicDB to check the effectiveness of the approach.

References

1. Z. Zhou, M.M. Rahman Siddiquee, N. Tajbakhsh, and J. Liang, "Unet++: A nested u-net architecture for medical image segmentation", *In International workshop on deep learning in medical image analysis Cham: Springer International Publishing*, pp. 3-11, 2018. https://doi.org/10.1007/978-3-030-00889-5_1
2. T. Jahnavi, and D. Vasundhara, "Segmentation of medical images using u-net++", *In 4th International Conference on Advances in Computing, Communication Control and Networking (ICAC3N*), (IEEE, 2022), pp. 801–807, 2022. https://doi.org./10.1109/ICAC3N56670.2022.10074438

3. D. Xu, X. Zhou, X. Niu, and J. Wang, “Automatic segmentation of low-grade glioma in MRI image based on UNet++ model”, *In Journal of Physics: Conference Series* Vol. 1693, No. 1, pp. 012135, 2020. https://doi.org./10.1088/1742-6596/1693/1/012135
4. Z. Li, H. Zhang, Z. Li, and Z. Ren, “Residual-attention unet++: a nested residual-attention u-net for medical image segmentation”, *Appl. Sci.* vol.12, pp. 7149, 2022. https://doi.org/10.3390/app12147149
5. B. Li, F. Wu, S. Liu, J. Tang, G. Li, M. Zhong and X. Guan, “CA-Unet++: An improved structure for medical CT scanning based on the Unet++ Architecture” *International Journal of Intelligent Systems*, vol.37, no.11, pp.8814-8832, 2022. https://doi.org/10.1002/int.22969Digital Object Identifier (DOI)
6. H. Cao, Y. Wang, J. Chen, D. Jiang, X. Zhang, Q. Tian, and M. Wang, “Swin-unet: Unet-like pure transformer for medical image segmentation”, *In European conference on computer vision). Cham: Springer Nature Switzerland*, pp. 205-218, 2022. https://doi.org/10.1007/978-3-031-25066-8_9
7. J. Chen, Y. Lu, Q. Yu, X. Luo, E. Adeli, Y. Wang, L. Lu, A.L. Yuille, and Y. Zhou, “Transunet: Transformers make strong encoders for medical image segmentation”, *arXiv preprint arXiv:2102.04306*, 2021. https://doi.org/10.48550/arXiv.2102.04306
8. X. Huang, Z. Deng, D. Li, X. Yuan, and Y. Fu, “Missformer: An effective transformer for 2d medical image segmentation”, *IEEE transactions on medical imaging*, vol.42, no.5, pp.1484-1494, 2022. https://doi.org/10.1109/TMI.2022.3230943
9. J. Ruan, J. Li, and S. Xiang, “Vm-unet: Vision mamba unet for medical image segmentation”, *ACM Transactions on Multimedia Computing, Communications and Applications*, 2024. https://doi.org/10.1145/3767748
10. R. Wu, Y. Liu, P. Liang, and Q. Chang, “H-vmunet: High-order vision mamba unet for medical image segmentation”, Neurocomputing, vol. 624, pp.129447,2024. https://doi.org/10.1016/j.neucom.2025.129447