2021 43rd Annual International Conference of the
IEEE Engineering in Medicine & Biology Society (EMBC)
Oct 31 - Nov 4, 2021. Virtual Conference

UCATR: Based on CNN and Transformer Encoding and Cross-Attention Decoding for Lesion Segmentation of Acute Ischemic Stroke in Non-contrast Computed Tomography Images

Chun Luo[1], Jing Zhang[1*], Xinglin Chen[1], Yinhao Tang[1], Xiechuan Weng[2] and Fan Xu[3]

Abstract— **The acute ischemic stroke (AIS) impacts extensively all over the world, the early diagnosis can provide valuable property information of disease. However, it's difficult for our human eyes to distinguish the fine pathological changes. Here we introduce self-attention mechanisms and propose UCATR, an NCCT image segmentation network for AIS lesions. It uses the advantages of Transformer to effectively learn the global context features of the image, and is based on convolutional neural network (CNN) and Transformer as the encoder, adding Multi-Head Cross-Attention (MHCA) modules to the decoder to achieve high-precision spatial information recovery. This method is experimentally verified on the NCCT dataset of AIS provided by Chengdu Medical College in China to obtain that the Dice similarity coefficient of lesion segmentation is 73.58%, which is better than U-Net, Attention U-Net and TransUNet. Furthermore, we conduct ablation study on the MHCA module at three different positions in the decoder to prove its efficiency.**

I. INTRODUCTION

Stroke is a lethal and disabling acute disease worldwide [1]. The majority of strokes are ischemic, and more than 80% of cases are due to the large artery atherosclerotic, cardioembolic and small vessel occlusion caused by thromboembolism [2]. Stroke can be classified into acute phase, subacute phase and chronic phase [3]. The core symptoms of acute ischemic stroke (AIS) contain aphasia, hemianopia, and loss of sensation, which could develop into chronic conditions (such as dementia, hemiplegia, etc.). Definitely, it is necessary to be diagnosed in a fast and accurate manner.

It is critical to use Computed Tomography (CT) and Magnetic Resonance Imaging (MRI) to evaluate AIS [4]. Our human eyes are difficult to distinguish the difference between lesion sites and healthy brain tissues at the early stage of AIS [5]. We are biased to recognize that CT is not as sensitive as MRI for ischemic stroke detection; nevertheless, it has recorded the detail changes of all scanned brain slices [6]. Consequently, it is necessary to extract valuable information from CT scan slices. We use the automatic segmentation algorithm, which can accurately and appreciably evaluate the lesion [7].

As deep learning develops, the convolutional neural network (CNN) has played a significant role in various visual-related tasks, especially U-Net [8] has been widely used in biomedical image segmentation globally. U-Net has a completely symmetrical encoder-decoder structure, and transfers the feature maps extracted by the encoder to the decoder through skip-connections. Some individuals have proposed a 3D multi-scale CNN automatic NCCT stroke segmentation method that requires the connected component analysis and the automatic hole filling method [9]. In addition, Hulin Kuang et al. [10] proposed a multi-task learning method EIS-Net, which includes T-CNN with a triple encoder and a decoder. To abstract and enhance the features of the image, it is designed with a comparison disparity block. Furthermore, a multi-level attention gate module is used during the process of recalibration of the decoder's the features.

However, the receptive field of convolution operations in CNN is limited by the size of the convolution kernel, which is lack of the long-distance dependency [11]. Transformer, which is very popular in natural language processing, has been proved to be effective in learning global context features in computer vision [12]. Therefore, someone proposed the TransUNet to use a CNN encoder to obtain local features, and then merge Transformer into a hybrid encoder in the U-Net down-sampling path to obtain global context features [13]. Moreover, some people have proposed U-Transformer, a Multi-head Self-Attention (MSA) module is used to obtain long range structural information from the image after the down-sampling path, and in the up-sampling path Multi-Head Cross-Attention (MHCA) modules combine high-level feature maps with rich semantic features and high-resolution feature maps connected by skip-connections to suppress irrelevant areas or noisy areas of high-resolution feature maps [14].

The contribution of our work is summarized as follows. Inspired by TransUNet and U-Transformer, we introduce self-attention mechanisms and propose UCATR, an NCCT image segmentation method for AIS lesions from sequence to sequence as illustrated in Fig. 1. On the basis of TransUNet, we use MHCA modules to fuse the feature maps of the CNN encoder in the up-sampling path of UCATR. Compared with TransUNet, it can achieve fine spatial recovery. Experiments have demonstrated that the accuracy of the current design facilitates NCCT image segmentation of AIS lesions.

*This research was supported partly by the National Science Foundation of China (No. 61405028, 61905036) and the Fundamental Research Funds for the Central Universities (University of Electronic Science and Technology of China) (No. ZYGX2019J053).

[1]School of Optoelectronic Science and Engineering, University of Electronic Science and Technology of China, Chengdu, China, 611731.

[2]Beijing Institute of Basic Medical Sciences, Beijng, China, 100850.

[3]Department of Public Health, Chengdu Medical College, Sichuan 610500.

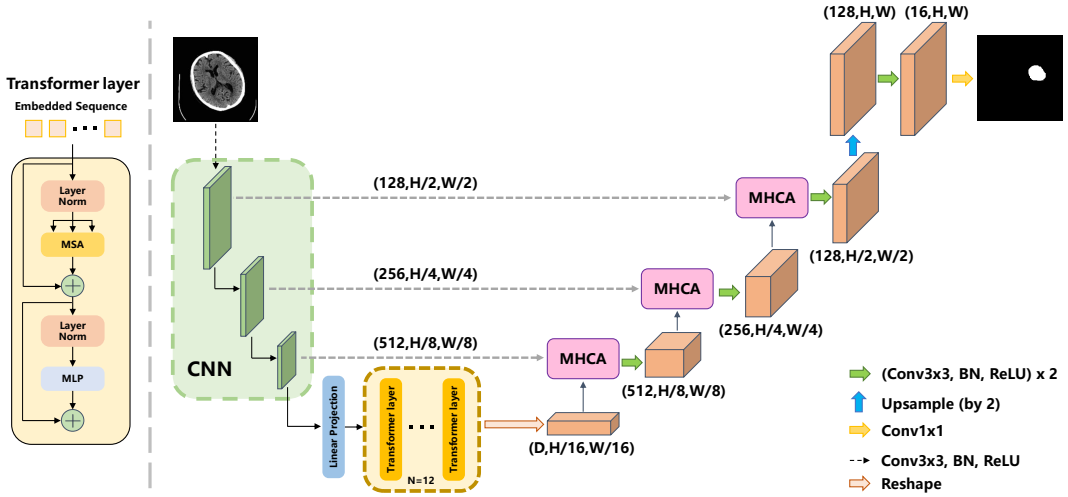

Fig. 1. The architecture of UCATR, which consists of CNN-Transformer Encoder and Cross Attention Decoder. From the CNN encoder, we obtain feature maps of different layers, and filter out their irrelevant information by skip-connections and the MHCA module to achieve a more refined restoration of spatial information in the decoder.

II. METHOD

Our proposed method UCATR is composed of CNN-Transformer Encoder and Cross Attention Decoder, and the feature maps in the CNN encoder are fused with the feature maps to be upsampled in the decoder through skip-connections and MHCA modules. ResNet50 is used as the backbone of the CNN encoder. The details of our proposed method are as follows:

A. CNN-Transformer Encoder

The spatial resolution of the dataset image used in this method is $x \in \mathbb{R}^{H\times W\times C}$, and the number of channels is C. CNN first extracts the features of the input image and generates the feature maps as the input of Transformer. The advantage of this is that in the up-sampling path the MHCA module can use the intermediate high-resolution image to achieve fine spatial information recovery.

With reference to the approach of Vision Transformer [12], we first convert the output feature map x of CNN encoder into a series of 2D patches$\{\mathbf{x}_p^i \in \mathbb{R}^{P^2\cdot C} \mid i = 1,...,N\}$, each of which has a size of $P\times P$, and an image has a total of $N = \frac{HW}{P^2}$ patches, where P is typically set to 16. Afterwards, the patches are passed into a linear embedding layer with output dimension D. To utilize the spatial information of the patch, the network could learn a specific position embedding and add it to the patch embedding to retain the position information: $\mathbf{z}_0 = \left[\mathbf{x}_p^1\mathbf{E};\mathbf{x}_p^2\mathbf{E};\cdots;\mathbf{x}_p^N\mathbf{E}\right] + \mathbf{E}_{pos}$, where $\mathbf{E} \in \mathbb{R}^{\left(P^2\cdot C\right)\times D}$ is patch embedding projection, and $\mathbf{E}_{pos} \in \mathbb{R}^{N\times D}$ is the position embedding.

$\mathbf{z}_0$ denotes the input of Transformer encoder, which Transformer encoder is composed of L layers Multi-head Self-Attention (MSA) and Multi-Layer Perception module (MLP) according to

$$\mathbf{z}_i' = \mathrm{MSA}\left(\mathrm{LN}\left(\mathbf{z}_{i-1}\right)\right) + \mathbf{z}_{i-1} \tag{1}$$

$$\mathbf{z}_i = \mathrm{MLP}\left(\mathrm{LN}\left(\mathbf{z}_i'\right)\right) + \mathbf{z}_i' \tag{2}$$

Where $\mathrm{LN}(\cdot)$ represents layer normalization and i is the intermediate layer identifier ranging from 1 to $L = 12$ total layers. An MSA layer consists of m parallel self-attention(SA) heads. Defining three learnable matrices $\mathbf{M}_q/\mathbf{M}_k/\mathbf{M}_v \in \mathbb{R}^{D\times D_h}$, the SA block calculates the similarity between two elements through the query($Q = \mathbf{z}_i\mathbf{M}_q$) and key($K = \mathbf{z}_i\mathbf{M}_k$) of the input sequence $\mathbf{z}_i$. Its calculation formula is below:

$$\mathrm{SA}(\mathbf{z}_i) = \mathrm{Softmax}\left(\frac{QK^\top}{\sqrt{D_h}}\right)V \tag{3}$$

Where $V = \mathbf{z}_i\mathbf{M}_v$ is the values of the input sequence and D_h is generally set to $\frac{D}{m}$. MSA is an extension of SA, and its formula is shown as follows:

$$\mathrm{MSA}(\mathbf{z}_i) = [\mathrm{SA}_1(\mathbf{z}_i);\mathrm{SA}_2(\mathbf{z}_i);\ldots;\mathrm{SA}_n(\mathbf{z}_i)]\,\mathbf{M}_{msa} \tag{4}$$

Where $\mathbf{M}_{msa} \in \mathbb{R}^{D\times D}$ denotes the learnable weight matrices of SA. MLP consists of two linear layers with a GELU activation function.

After CNN-Transformer Encoder we obtain the encoded sequence $\mathbf{z}^L \in \mathbb{R}^{\frac{HW}{P^2}\times D}$.

B. Cross Attention Decoder

The output $\mathbf{z}^L$ of CNN-Transformer Encoder reshapes from $\frac{HW}{P^2}\times D$ to $\frac{H}{P}\times\frac{W}{P}\times D$. Although the feature maps skip-connected from CNN hold high-resolution information, they are short of the rich semantic information of feature maps in deeper layers of the network. Therefore, we use the MHCA module as shown in Fig. 2, tending to suppress the unrelated or noisy areas in the feature map connected by skip-connections and highlight the significant areas. The V obtained from the low-level feature map C and the Q and K obtained from the high-level feature map T are input into the MSA layer for calculation, and the result is readjusted to the calculated weight value between 0 and 1 through a sigmoid activation function to obtain the tensor Z. This is equivalent to the low-level feature map C using the richer semantic

features of the high-level feature map T to obtain a filter Z for reducing irrelevant or noisy areas and enhancing important features. Therefore, after the Hadamard product of Z and the low-level feature map C, the irrelevant or noisy areas are filtered out, and more refined spatial information recovery is achieved. After that, the result is connected with the high-level feature map T. After passing through multiple MHCA modules, the image is restored to the original resolution. The detailed Cross Attention Decoder could be found in Fig. 1.

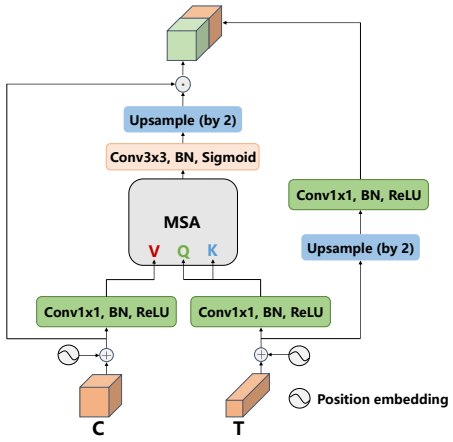

Fig. 2. The structure of the MHCA module. From the high-level feature map T, we could obtain Q and K. The input of the MSA layer is Q, K and V which are acquired from the feature map C. In addition, the result of the MSA layer passes a sigmoid activation function, and then calculates the Hadamard product with the feature map C. The final result is stitched with the up-sampled feature map T.

III. EXPERIMENT AND RESULT DISCUSSION

A. Dataset

Our experiment is conducted on the NCCT image dataset of AIS provided by Chengdu Medical College, China. It is composed of the data from 11 patients, with an average of 95 slices per patient, and 293 label images are obtained by manual annotation by 3 professional doctors. The dataset is divided into the training set, validation set, and test set on the scale of 8:1:1. The training set is rotated randomly within the range of [-90°, 90°], and flipped randomly with a probability of 0.5 for the augmentation. All the experimental procedures involving human are approved by the Institution's Ethical Review Committee.

B. Metrics

To evaluate the accuracy of segmentation, we select the Dice similarity coefficient (DSC) and Sensitivity (SEN) between the manually segmented test image and the corresponding automatic segmentation as evaluation indicators. The formats are as follows:

$$DSC = \frac{2TP}{2TP+FP+FN} \tag{5}$$

$$SEN = \frac{TP}{TP+FN} \tag{6}$$

Where TP and FP are the sizes of true or false positive areas, and FN is the size of false negative areas.

C. Training setup

The SGD optimizer is used in the experiment for 110 epochs, with a learning rate of 1e-4, the momentum of 0.9, and the weight decay of 1e-3. Besides, the default batch size is 2. And we train the model by adding cross-entropy loss function and Dice loss function ($Dice\ loss = 1 - DSC$) to reduce the adverse effects of category imbalance. All experiments are performed on an Nvidia RTX2080Ti GPU.

TABLE I
QUANTITATIVE COMPARISON OF DIFFERENT SEGMENTATION METHODS

Method	Dice(%)	Sensitivity(%)
U-Net	49.38	50.00
Attention U-Net	62.50	61.66
TransUNet	70.62	69.39
UCATR	**73.58**	**73.12**
UCA	71.39	73.37

D. Result discussion

UCATR Compare with Different Methods: As can be seen from the Table I, UCATR has improved DSC from 49.38% to 73.58% compared with U-Net. SEN also improves from 50.00% to 73.12%. In addition, compared with Attention U-Net, both DSC and SEN have been greatly improved, indicating that UCATR is better than some of the current commonly used segmentation methods. Secondly, The MHCA module placed in the up-sampling path can achieve more refined spatial information recovery, so UCATR compared with TransUNet, DSC increases from 70.62% to 73.58%, SEN increases from 69.39% to 73.12%, which proves the effectiveness of Cross Attention Decoder. To test the efficiency of the Transformer encoder, we remove the Transformer of UCATR and mark it as "UCA". It exhibits that UCA is 2.19% lower than UCATR for DSC. This is because of the lack of Transformer to learn global features and provide long-distance structural information. Fig. 3 shows the qualitative segmentation comparison between U-Net, Attention U-Net, TransUNet and UCATR. We can observe that UCATR performs better on the segmentation of small target lesions. For example, the other methods in the first row have over-segmentation or under-segmentation. This represents that the combination of Transformer and MHCA is better for the segmentation of small target lesions.

TABLE II
THE RESULTS OF ABLATION STUDY WITH MODULES PLACED IN DIFFERENT POSITIONS

Method	Bottom	Middle	Top	Dice(%)	Sensitivity(%)
1	✓			72.26	78.16
2		✓		73.02	74.21
3			✓	68.83	67.74
4	✓	✓		72.37	76.00
5		✓	✓	73.01	76.04
6	✓		✓	72.18	73.93
7	✓	✓	✓	**73.58**	**73.12**

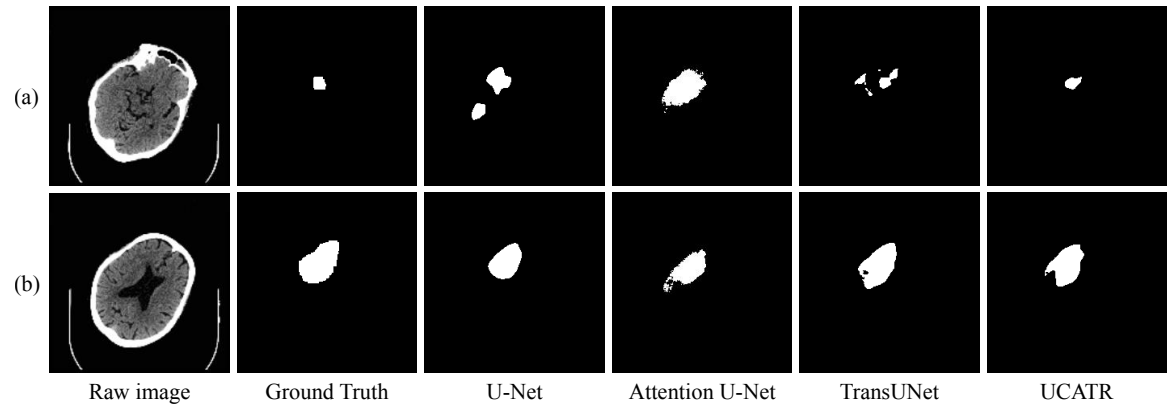

Fig. 3. The comparison of the results of the segmentation of two NCCT images (a) and (b) of AIS with ground truth by four methods. And it could be observed that the effect of UCATR is the best.

Ablation Study: The results of the ablation study for the MHCA module are displayed in Table II. The method marked with ✓ in the three different skip-connections (Bottom, Middle, Top) means that the MHCA module is placed at the corresponding skip-connection. It shows that if the MHCA module is separately placed at the top skip-connection, it is 4.75% lower than UCATR for DSC and 5.38% lower than UCATR for SEN. We speculate that this is because feature maps transmitted from CNN through the top skip-connection has more original information and less semantic features than feature maps transmitted through the bottom and middle skip-connections, and the semantic features of high-level feature maps used in the MHCA module at the top skip-connection to form the filter are not as rich as the semantic features of high-level feature maps at the bottom and middle skip-connections. Moreover, the MHCA modules in UCATR are placed at the three skip-connections, which could have a higher DSC than other placement schemes, indicating that this could better fuse low-level features with high-level features. Furthermore, compared to TransUNet, on condition that one MHCA module is added, the DSC rises by at least 1.56%, which proves the effectiveness of the MHCA module.

IV. CONCLUSIONS

In this paper, we present a network that introduces a self-attention mechanism, using a mixture of CNN and Transformer as the encoder and using the MHCA module to achieve more refined decoding. The proposed method is varified on the ischemic stroke NCCT dataset provided by Chengdu Medical College, China. The DSC of this method is 73.56% according to the experimental result, which is superior to U-Net, Attention U-Net, and TransUNet. The ablation study for the MHCA module at three different positions in the decoder also proves its usefulness. This effectively helps the doctor to quickly delineate the lesion.

REFERENCES

[1] N. J. Kassebaum, A. Bertozzi-Villa, M. S. Coggeshall, K. A. Shackelford, C. Steiner, K. R. Heuton, D. Gonzalez-Medina, R. Barber, C. Huynh, D. Dicker *et al.*, "Global, regional, and national levels and causes of maternal mortality during 1990–2013: a systematic analysis for the global burden of disease study 2013," *The Lancet*, vol. 384, no. 9947, pp. 980–1004, 2014.

[2] M. Ariesen, S. Claus, G. Rinkel, and A. Algra, "Risk factors for intracerebral hemorrhage in the general population: a systematic review," *Stroke*, vol. 34, no. 8, pp. 2060–2065, 2003.

[3] R. G. González, J. A. Hirsch, W. Koroshetz, M. H. Lev, and P. W. Schaefer, *Acute ischemic stroke*. Springer, 2011.

[4] R. T. Higashida and A. J. Furlan, "Trial design and reporting standards for intra-arterial cerebral thrombolysis for acute ischemic stroke," *Stroke*, vol. 34, no. 8, pp. e109–e137, 2003.

[5] P. Mikhail, M. G. D. Le, and G. Mair, "Computational image analysis of nonenhanced computed tomography for acute ischaemic stroke: A systematic review," *Journal of Stroke and Cerebrovascular Diseases*, vol. 29, no. 5, p. 104715, 2020.

[6] D. P. Davis, T. Robertson, and S. G. Imbesi, "Diffusion-weighted magnetic resonance imaging versus computed tomography in the diagnosis of acute ischemic stroke," *The Journal of emergency medicine*, vol. 31, no. 3, pp. 269–277, 2006.

[7] O. Maier, B. H. Menze, J. von der Gablentz, L. Häni, M. P. Heinrich, M. Liebrand, S. Winzeck, A. Basit, P. Bentley, L. Chen *et al.*, "Isles 2015-a public evaluation benchmark for ischemic stroke lesion segmentation from multispectral mri," *Medical image analysis*, vol. 35, pp. 250–269, 2017.

[8] 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*. Springer, 2015, pp. 234–241.

[9] A. Tuladhar, S. Schimert, D. Rajashekar, H. C. Kniep, J. Fiehler, and N. D. Forkert, "Automatic segmentation of stroke lesions in non-contrast computed tomography datasets with convolutional neural networks," *IEEE Access*, vol. 8, pp. 94 871–94 879, 2020.

[10] H. Kuang, B. K. Menon, S. I. Sohn, and W. Qiu, "Eis-net: Segmenting early infarct and scoring aspects simultaneously on non-contrast ct of patients with acute ischemic stroke," *Medical Image Analysis*, vol. 70, p. 101984, 2021.

[11] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7794–7803.

[12] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[13] 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.

[14] O. Petit, N. Thome, C. Rambour, and L. Soler, "U-net transformer: Self and cross attention for medical image segmentation," *arXiv preprint arXiv:2103.06104*, 2021.