A Medical Image Semantics Segmentation Method Based on Image Pre-processing and Image Transformer

Zhaopeng Li*

School of Mathematics and Statistics, University of Melbourne, Melbourne, 3052, Australia

Abstract. Semantics segmentation is a task aiming at classifying each pixel of an image into a category. Because of its ability to comprehend and interpret the context of images, it plays a vital role in the medical area. Existing models in this field pay more attention to editing U-Net's structure and convolutional layers. In this paper, an image pre-processing technique to enrich the potential of images and an image transformer involved in network TrUNet are proposed to tackle these issues. In particular, the histogram equalisation method was introduced to improve the quality of each image, and image augmentation methods including flipping and rotation were used to increase the size of the image dataset. Then TrUNet cooperates vision transformer to U-Net to improve the model's ability to recognize global context. Extensive experiments show that this method outperforms U-Net on the DRIVE dataset. Its validation loss curve is smoother and decreases more decently. Its accuracy increases more significantly in early epochs and is higher than that of U-Net. It has higher precision and F1-score compared with U-Net.

1 Introduction

Semantics segmentation is a popular task in computer vision that focuses on segmenting image pixels into groups based on pixels presented in the images [1]. This technique plays an important role in the medical field since accurate segmentation results provide huge amounts of shape information and are the basis for other applications such as computer-aided diagnosis, detection of brain tumours, and liver segmentation [2-4].

With the introduction of U-Net in medical segmentation in 2015, U-Net has gained many attractions [5]. In 2017, Dong et al. applied U-Net on a magnetic resonance imaging (MRI) dataset which can achieve a dice similarity score of 0.86 compared to previous methods of only 0.67 [6]. Due to its success, many types of research have been conducted and produced variants focused on redesigning encoder-decoder architecture, convolutional layers and skip connections to make the model better learn the local context better.

However, these researchers omitted the power of global context in the medical images and spent too much effort on model structures. Nevertheless, compared with massive research

* Corresponding author: zhaopengl@student.unimelb.edu.au

done on deep learning models, few types of research have been done on the impact of image pre-processing methods on medical image segmentation.

In this paper, a two-stage algorithm is proposed. This paper aims to effectively improve the performance of the traditional U-Net model.

First, the image pre-processing method histogram equalisation was introduced which strengthens the key features of images and applies data augmentation methods to improve the quantity of data.

Second, TrUnet is a network in which a vision transformer is utilized in U-Net to improve the model's capabilities in recognizing the long-range features of images.

Extensive experiments have been done on the DRIVE dataset to evaluate the performance of TrUnet which outperforms traditional U-Net in both accuracy and speed.

2 Method

The traditional U-Net model extracts local features through convolutional layers. In a task like semantic segmentation, global features are necessary as well. Moreover, in order to deal with data scarcity, it is important to employ image processing methods to enhance data quality and data quantity. Therefore, a two-stage model is proposed.

2.1 Stage one: Image preprocessing stage

2.1.1 Histogram equalisation

In this stage, image processing methods are employed to improve the image in two ways. At first, all images will be passed through histogram equalisation to improve the quality of each pixel. Histogram equalisation is a schema widely used in image processing which aims at making images have a uniform distribution of intensities [7]. By applying the equalisation method, each pixel will be assigned a new intensity value based on the formula 1. In the formula, r_k is the intensity with value k, L is the total number of possible intensities and $p_r(r_j)$ is the probability of intensity with value r. As a result, the intensity histogram was flattened.

$$s_k = T(r_k) = (L-1)\sum_{j=0}^{k} p_r(r_j) \tag{1}$$

2.1.2 Image augmentation

The equalized images were then exploited for augmentation. Since labelled data is very expensive to produce in the medical domain, especially in segmentation tasks [8]. It is wise to create additional images without breaking the main content. Geometric transformation is a commonly used augmentation strategy in medical image processing [9]. At this step, rotation together with both horizontal and vertical flipping was executed and finally quadrupled the size of the training dataset.

2.2 Stage two: Transformer involved U-Net network

2.2.1 Vision transformer

Transformer is a technique in natural language processing that implements an attention mechanism and achieves successful results [10]. Vision transformers as an extension in the image field also have strengths in computational usage efficiency, multi-task competency,

and a strong ability in recognising the global context of images [11]. An image is split into a series of equal-sized image patches. In this experiment, each image is split into patches with size 16*16 and passed through the convolutional neural network with 768 output channels.

After flattening sequential patches, a positional encoding matrix was added to include spatial information. The embedded patches were then sequentially passed to 12 transformer encoders each containing two blocks: multi-head attention block and multi-layer perceptron block.

2.2.2 TrUNet

U-Net applied an encoder-decoder structure composed of sequential multiple 2D convolution, batch normalisation, and activation (RELU) steps. In this structure, max pooling is employed to encode and up-sampling strategy to decode the encoded images. Additionally, there were skip connections to stack encoder outputs to decoder inputs at the same dimension.

In the proposed TrUNet, outputs from transformers after transformation were also stacked together in skip connection steps to incorporate the information of long-range context of the original image. These transformations each consist of designed 2D convolution with specific kernel and stride size, batch norm, and RELU activation function. They were exploited to transform output to keep the shape matches the decoder inputs (Fig. 1).

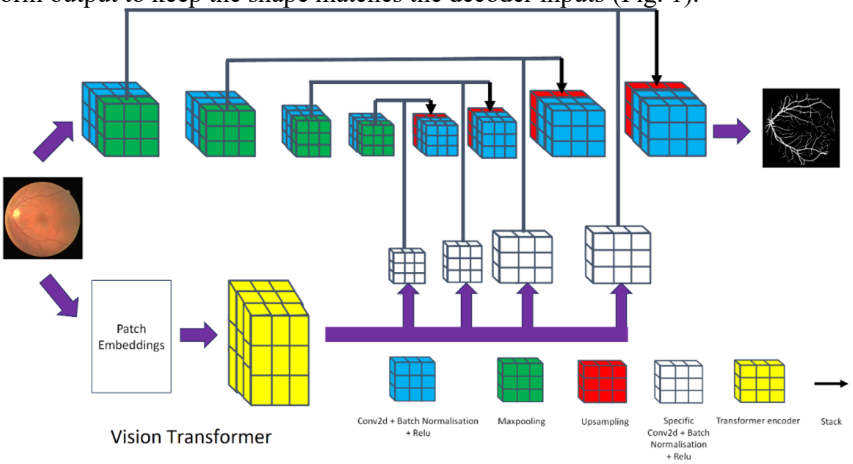

Fig. 1. The structure of TrUNet (Picture credit: Original)

3 Experiment

3.1 Dataset

The dataset used for this experiment is The Digital Retinal Images for Vessel Extraction (DRIVE) dataset. This dataset contains 40 colored retina images half of them are training sets and the other half are testing sets. Among training samples, 25% of them were used as validation sets.

3.2 Evaluation metrices

Evaluation metrics for comparing the performance of models consist of validation loss, accuracy, precision, recall, and F1-score.

Accuracy is the percentage of pixels classified as correct. It follows the following formula 2 where TP is the number of pixels correctly classified as true, TN is the number of pixels classified as negative, FP is the number of pixels incorrectly classified as true and FN is the number of pixels incorrectly classified as negative.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (2)$$

Precision is the proportion of all classified as positive samples that are positive.

$$Precision = \frac{TP}{TP+FP} \quad (3)$$

Recall refers to the proportion of all real positive samples which are classified as positive.

$$Recall = \frac{TP}{TP+FN} \quad (4)$$

Although precision and recall are meaningful as they reflect the quality and quantity of the model, it is hard to balance these two. Thus, the F1-score as an evaluation metric that combines precision and recall is applied.

$$F1 - score = \frac{2*Precision*Recall}{Precision+Recall} \quad (5)$$

3.3 Experiment settings

3.3.1 Optimizer

The optimizer chosen is the Adaptive Moment Estimation (ADAM) optimizer which utilizes two techniques: momentum and root mean square propagation [12]. The former makes it faster to reach the global minimum and the latter makes it possible to adaptively learn parameters.

3.3.2 Loss function

Binary cross entropy (BCE) loss function is a traditional loss function widely involved in segmentation tasks. Its stability produces smooth curves making model training faster. In the formula below, N represents the total number of instances, y_i is the true label of instance I and $\hat{p}_i$ is the probability of predicting the correct label.

$$BCE\ loss = -\frac{1}{N}\sum_{i=0}^{N} y_i * log\ \log(\hat{p}_i) + (1-y_i) * \log(1-\hat{p}_i) \quad (6)$$

The dice loss function is another famous loss function used in segmentation because of its potential to deal with class imbalance. In formula 7, N is the total number of pixels, y_i and $\hat{y}_i$ each represent the ith pixel of the true image and predicted image.

$$Dice\ loss = 1 - \sum_{i=0}^{N} \frac{2*|y_i \cap \hat{y}_i|}{|y_i|+|\hat{y}_i|} \quad (7)$$

When combining these loss functions, it can guarantee both a nice gradient during the training process contributed by BCE loss as well as handling class imbalance solved by dice loss [13].

$$BCE\ Dice\ loss = 1 - \sum_{i=0}^{N} \frac{2*|y_i \cap \hat{y}_i|}{|y_i|+|\hat{y}_i|} - \frac{1}{N}\sum_{i=0}^{N} y_i * log\ log\ (\hat{p}_i) + (1-y_i) * log\ log\ (1-\hat{p}_i) \quad (8)$$

4 Results

4.1 Data-preprocessing

Fig 2 describes the comparison of change of validation loss across epochs between the method using image pre-processing and the one without. In Fig 2, the vertical axis represents the validation loss and the horizontal axis represents the epochs number.

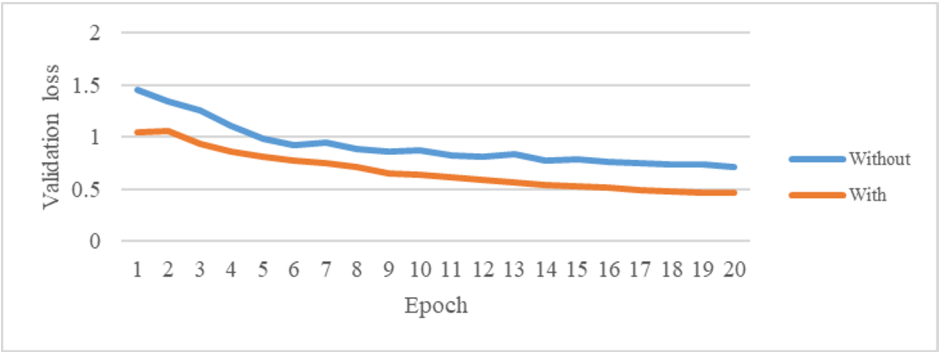

Fig. 2. The comparison of validation loss between using image preprocessing and without using it (Picture credit: Original)

According to Fig 2 validation loss descends more significantly when using pre-processing methods than the one without. With using the preprocessing technique, the curve is smoother and decreases more decently. They both showed that the preprocessing technique is very helpful in improving the model performance.

4.2 Model

Fig 3 depicts the accuracy change of the UNet model and TrUNet model along the epochs. The y-axis of Fig 3 is the accuracy score and the x-axis is the epoch number.

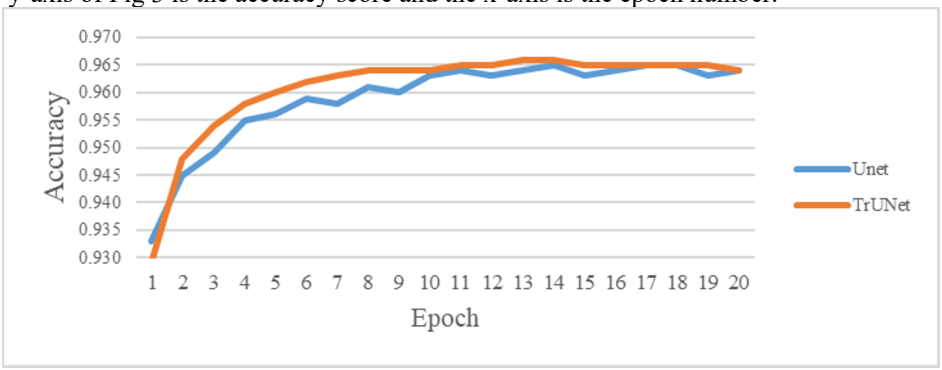

Fig. 3. The accuracy plot between Unet and TrUNet (Picture credit: Original)

According to the image, TrUNet achieves higher accuracy than UNet along epochs. Moreover, during earlier epochs, the accuracy of the TrUNet model grows faster than the original UNet which shows that the concatenated embeddings can accelerate the training process.

Table 1. The metrics between U-Net and TrUNet

Model	Precision	Recall	F1-score
U-Net	0.790	0.800	0.793
TrUNet	0.814	0.797	0.803

Based on the content of Table 1, TrUNet can achieve a higher F1-score of 0.803 compared with 0.793 achieved by U-Net. When comparing precision and recall, TrUNet has a much higher precision score than the U-Net and TrUNet has a very closed recall score compared with the U-Net. They all show that TrUNet performs better than U-Net.

5 Discussion

Even though the proposed TrUNet performs better than U-Net in performance, it has two drawbacks comparatively. First of all, it introduced the transformer technique which contains many parameters and undoubtedly increases the size of the model. This also reflects the need for more GPU space to train the model during training. Another drawback is TrUNet has a lower Recall value than U-Net. It demonstrates that while having more information about each pixel, TrUNet becomes stricter in predicting pixels to positive which eventually reduces the TP value and recall score.

For computer vision-related machine learning tasks in the medical field, data scarcity is a big problem. It requires professors with experience and huge amounts of time to obtain a labelled image. Since data preprocessing has proved to have a significant effect on the final result and only simple geometry transformation methods were applied in this experiment, more professional image augmentation techniques including using CutMix technique which replaces one region of the image with a patch from another or cooperating Segment Anything (SAM) technique in image augmentation can be applied [14, 15]. Furthermore, there are many different datasets available which aim at solving different medical problems. Since in this experiment, only the DRIVE dataset is tested which lacks universality, various medical image datasets can be employed to test the model's performance in different medical scenarios in the future. Moreover, stacking outputs of vision transformers directly to the input of decoders in U-Net seems to have some positive but limited impact on model performance. Cooperating embeddings to the network in other ways can be tried in the future. Changes can be made to the transformer. Patch sizes other than 16 such as 32 can be tested in the future to check whether patch size could be another factor that affects the overall performance.

6 Conclusion

In this paper, a two-stage strategy was proposed for semantics segmentation in medical images. The first stage is to employ image processing methods to improve the quality and quantity of data. Histogram equalisation and geometric data augmentation methods were employed at this stage. In the second stage, TrUNet is proposed which involves using a vision transformer to encode images and incorporate them into U-Net in the skip connection step to improve the model's capability in using global context to produce more accurate results. After extensive experiments on the DRIVE dataset, the strategy has proved its superiority in performance compared with U-Net. However, TrUNet also has drawbacks in increased model size and reduction in recall value. For future research, advanced augmentation techniques, versatile datasets, and different transformer cooperating methods can be implemented.

References

1. R. Yang, Y. A. Yu, Artificial convolutional neural network in object detection and semantic segmentation for medical imaging analysis. Front. Oncol. 11, 638182 (2021).
2. B. Van Ginneken, C. M. Schaefer-Prokop, M. Prokop, Computer-aided diagnosis: how to move from the laboratory to the clinic. Radiology 261(3), 719-732 (2011).
3. R. Hashemzehi, S. J. S. Mahdavi, M. Kheirabadi, S. R. Kamel, Detection of brain tumors from MRI images base on deep learning using hybrid model CNN and NADE. Biocybern. Biomed. Eng. 40(3), 1225-1232 (2020).

4. G. Chlebus, A. Schenk, J. H. Moltz, B. van Ginneken, H. K. Hahn, H. Meine, Automatic liver tumor segmentation in CT with fully convolutional neural networks and object-based postprocessing. Sci. Rep. 8(1), 15497 (2018).
5. O. Ronneberger, P. Fischer, T. Brox, U-net: Convolutional networks for biomedical image segmentation. In: Med. Image Comput. Comput.-Assist. Interv. MICCAI 2015, Part III, 234-241 (Springer, 2015).
6. H. Dong, G. Yang, F. Liu, Y. Mo, Y. Guo, Automatic brain tumor detection and segmentation using U-Net based fully convolutional networks. In: Med. Image Understand. Anal. MIUA 2017, 506-517 (Springer, 2017).
7. S. S. Bagade, V. K. Shandilya, Use of histogram equalization in image processing for image enhancement. Int. J. Softw. Eng. Res. Pract. 1(2), 6-10 (2011).
8. G. Bortsova, F. Dubost, L. Hogeweg, I. Katramados, M. De Bruijne, Semi-supervised medical image segmentation via learning consistency under transformations. In: Med. Image Comput. Comput.-Assist. Interv. MICCAI 2019, Part VI, 810-818 (Springer, 2019).
9. P. Chlap, H. Min, N. Vandenberg, J. Dowling, L. Holloway, A. Haworth, A review of medical image data augmentation techniques for deep learning applications. J. Med. Imaging Radiat. Oncol. 65(5), 545-563 (2021).
10. A. Vaswani, Attention is all you need. Adv. Neural Inf. Process. Syst. (2017).
11. A. Dosovitskiy, An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020).
12. D. P. Kingma, Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014).
13. S. A. Taghanaki, Y. Zheng, S. K. Zhou, B. Georgescu, P. Sharma, D. Xu, D. Comaniciu, G. Hamarneh, Combo loss: Handling input and output imbalance in multi-organ segmentation. Comput. Med. Imaging Graph. 75, 24-33 (2019).
14. S. Yun, D. Han, S. J. Oh, S. Chun, J. Choe, Y. Yoo, Cutmix: Regularization strategy to train strong classifiers with localizable features. In: Proc. IEEE/CVF Int. Conf. Comput. Vis., 6023-6032 (2019).
15. A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W. Y. Lo, P. Dollár, Segment anything. In: Proc. IEEE/CVF Int. Conf. Comput. Vis., 4015-4026 (2023).