International Journal of Computational Intelligence Systems and Applications. Volume 13, Issue 1, (2026) 9-15
http://www.cayley-nielson.com/ijcisa/

Retinal Image Segmentation Using U-Net Model

Cui Dongyan
School of Artificial Intelligence, North China University of Science and Technology, Tangshan, Hebei, China
E-mail: *cdy_xxz@163.com*

Abstract

With the continuous development and application of deep learning technology in the field of computer vision, deep learning network models can now perform precise and rapid analysis and processing of retinal images, providing ophthalmologists with more accurate diagnostic support. This paper employs the U-Net network model to establish a suitable retinal model. After training, the U-Net model can segment different structures in input retinal images, clearly revealing vascular changes, thereby enabling automated retinal disease diagnosis and analysis and helping doctors develop effective treatment plans. Experiments demonstrate that the U-Net model achieves relatively good performance in retinal image segmentation tasks, enhancing both the efficiency and accuracy of medical diagnosis.

Keywords: Deep Learning; U-Net Model; Retinal Images; Medical Image Segmentation.

1. Introduction

With the acceleration of technological advancements, medical image processing technology has become increasingly important in the healthcare industry. Medical image segmentation technology has witnessed decades of technical progress. Initially, doctors had to manually label images one by one—a method that was inefficient and prone to errors. However, with the rapid development of computer technology, automatic and intelligent segmentation techniques have gradually gained popularity. Traditional segmentation methods primarily include grayscale thresholding, edge detection, and data-driven approaches. Although these techniques have been widely adopted, they still face challenges such as noise interference and poor adaptability to different image patterns. Deep learning has become widely used in medical image processing and has achieved significant progress in certain application areas [1-2]. Deep learning techniques can be employed for the segmentation and detection of medical images, thereby assisting doctors in accurate diagnosis and treatment [3-4].

In 2020, Ibtehaz et al. proposed several improved U-Net models, achieving remarkable performance enhancements [5].

Isensee et al. introduced nnU-Net, a robust and adaptive framework for medical image segmentation. This deep learning-based method automatically configures itself for any new task, including preprocessing, network architecture, training, and postprocessing [6].

In 2024, Peng et al. proposed DP-U-Net++ for the segmentation of colorectal adenoma microscopic images. By utilizing deep feature fusion, pre-trained encoders, and positional attention, DP-U-Net++ effectively integrates multi-level features, reconstructs highly correlated features, and preserves low-level features for output refinement [7].

The U-Net model is a deep learning architecture designed for image segmentation. It has demonstrated excellent performance in tasks such as medical image segmentation and has been widely applied in various

image segmentation applications. Therefore, this study focuses on applying the U-Net model to retinal image segmentation, with the aim of achieving improved segmentation results.

2. U-Net Network Model

The U-Net model is primarily used for pixel-level classification, where the output corresponds to the category of each pixel. It has achieved outstanding performance in tasks such as medical image segmentation and has been extensively adopted in various image segmentation applications. The U-Net network model is illustrated in Figure 1.

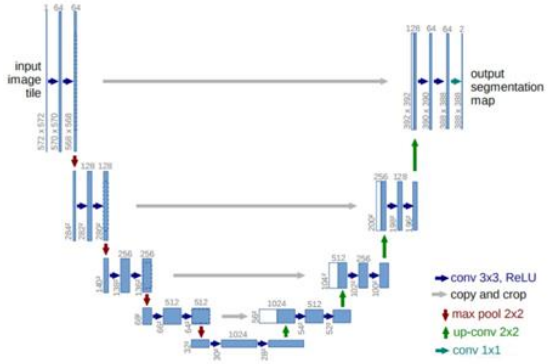


Fig.1 U-Net Network Model

2.1. Composition of the U-Net Network Model

The U-Net architecture can be summarized into three main modules: the feature enhancement module, the result prediction module, and the backbone extraction module. Each component serves a specific function and collaborates to achieve effective image segmentation.

2.1.1. Backbone Extraction Section

In the U-Net architecture, the classic VGG architecture serves as the foundation, allowing for the extraction of important features through a series of convolution and max-pooling operations. This extraction process consists of two parts: an encoder and a decoder, ensuring deep learning and feature extraction of the input data.

1) Encoder

The encoder structure typically comprises multiple layers of convolutional processing units. Each unit consists of a convolutional layer with filters, followed by a batch normalization layer for data standardization, and an activation layer (most commonly the ReLU activation function) to introduce non-linear processing. These components work together to identify and learn basic features in the image. These convolutional blocks help the network learn low-level features of the image. After each convolutional block, pooling layers (e.g., max pooling or average pooling) are typically added for downsampling, gradually increasing the receptive field and reducing spatial dimensions. The encoder is usually divided into multiple stages, each reducing the feature dimensionality while increasing the level of abstraction, enabling the network to learn hierarchical features of the image.

2) Decoder

The decoder is a network architecture that includes transposed convolutional layers and skip connections. In this process, the decoder fuses the corresponding feature layers of the encoder with the

upsampled feature layers of the decoder through skip connections, thereby aiding in the precise restoration of image details. At each decoding stage, the feature dimensions are progressively expanded, while more detailed information is extracted from the skip connections of the encoder, helping the network effectively reconstruct image details.

By combining the encoder and decoder, U-Net achieves accurate image segmentation, outputting category labels or probability values for each pixel. The design of the backbone extraction module in U-Net enables the network to perform excellently with images of varying sizes and complexities, achieving remarkable results in many image segmentation tasks.

2.1.2. Feature Enhancement Module

To effectively improve the accuracy of image segmentation, the feature extraction stage of the architecture can be optimized. First, five basic feature layers are obtained from the main part of the network. These feature layers are then upsampled and concatenated to integrate information from different layers, forming an effective feature map.

2.1.3. Prediction Module

The input is an image, which is processed through the encoder to produce a series of feature maps. These feature maps are then upsampled in the decoder and restored to the original image size using skip connections, generating category predictions for each pixel. This step essentially involves classifying each pixel in the image individually.

2.2. Loss Function of the U-Net Model

The main components of the U-Net loss function include the softmax function, the weighted cross-entropy loss function, and the weight calculation function. The softmax activation function non-linearly transforms the input features and weights of each pixel, and the number of output values generated after softmax processing equals the number of categories in the labels. The cross-entropy loss function measures the difference between two probability distributions. The weight calculation function adjusts the importance of specific regions in the image.

2.2.1. Softmax Activation Function

The softmax activation mechanism non-linearly combines the features of input pixels with their corresponding weights and converts them into a specific output format. This transformation generates as many output values as the total number of categories. After application, it converts the output of each pixel into a set of positive numbers that sum to 1, essentially forming a probability distribution. This allows us to evaluate the confidence level of each pixel belonging to each category. The formula is shown in (1).

$$Softmax(z_i) = \frac{\exp(z_i)}{\sum_j \exp(z_j)} \tag{1}$$

where z_i and z_j are elements in Z.

2.2.2. Cross-Entropy Loss Function

The cross-entropy loss function is primarily used to measure the difference between two probability

distributions. The formula is shown in (2):

$$L = -\sum_{c=1}^{M} y_c \log(p_c) \quad (2)$$

In this formula, y_c represents the true distribution of the sample, taking values of either 0 or 1, while p_c represents the predicted distribution of the sample. The formula is shown in (3):

$$E = \sum_{x=\Omega} \omega(x) \log(p_{\delta(x)}(x) \quad (3)$$

Here, p is the output value processed by softmax; $p_{\delta(x)}(x)$is the activation value output for the relative category at point x as given by the corresponding label.

2.2.3 Weight Calculation Function

$\omega(x)$ represents a weight map. $\omega_c(x)$ is the segmentation ground truth, and the segmentation value balances the frequency of pixels of different classes in the training dataset; d_1 is the distance to the nearest selected boundary, and d_2 is the distance to the second nearest selected boundary. The formula is shown in (4):

$$\omega(x) = \omega_c(x) + \omega_0 * \exp(-\frac{(d_1(x)+d_2(2))^2}{2\sigma^2}) \quad (4)$$

3. Data Processing

3.1. Data Collection

The data used in this experiment consists of the DRIVE retinal image dataset. The dataset contains 20 images, each with dimensions of 565×584 pixels. The processed results (including ground truth, predicted results, and original image masks) amount to a total of 60 images. The dataset structure is shown in Figure 2.

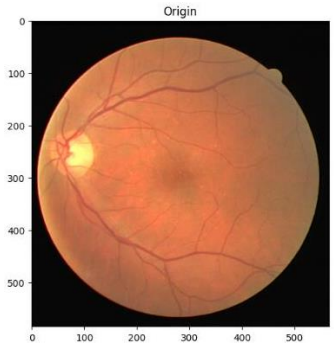


Fig.2 Dataset Structure Diagram

3.2. Data Processing

In the initial stages of data analysis or model construction, data preprocessing plays a crucial role. The primary tasks at this phase involve cleaning, transforming, and adjusting data to ensure its quality and enhance usability. Particularly in removing noise, filling missing values, and filtering anomalous data, data preprocessing is essential. This not only improves data accuracy but also enhances the performance of the final analysis or model.

In the context of image processing, especially in image segmentation, preprocessing often refers to image denoising. This is because noise in the original images can adversely affect segmentation accuracy.

4. Experimental Process and Result Analysis

4.1. Model Training

The typical steps for training a U-Net model are as follows:

1) Prepare the training dataset: Use the collected retinal images as training data.

2) Build the U-Net model: Create the U-Net model architecture and set appropriate parameters.

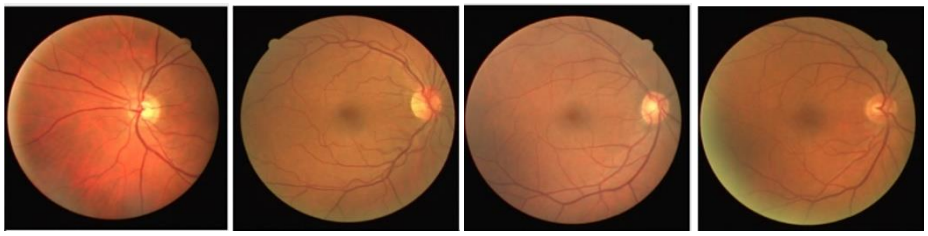
3) Compile the model: Select a suitable optimizer and loss function, and compile the U-Net model.

4) Model training: Input the training data into the U-Net model to optimize the model parameters, ensuring accurate prediction of image labels.

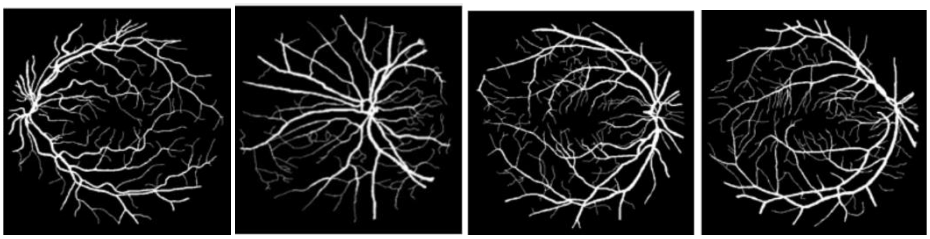
5) Model evaluation: Use the validation set to evaluate the performance of the trained model, including key metrics such as model accuracy and loss values.

6) Prediction process: Apply the trained U-Net model to analyze new images and generate segmentation results.

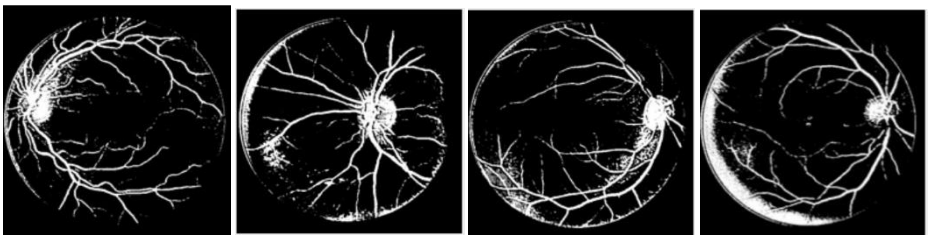
4.2. Experimental Results and Analysis



(a) Original Image



(b) Ground Truth



(c) Predicted Result

(d) Original Image with Mask

Fig.3 Experimental Results

The experimental results are shown in Figure 3 above. Figure 3 shows the results obtained through the U-Net network model. Specifically, Figure 3(a) is the original image, Figure 3(b) is the ground truth mask, Figure 3(c) is the prediction result, and Figure 3(d) is the original image overlaid with the prediction mask.

By comparing the original image with the prediction result, we can clearly observe the distribution of retinal blood vessels. When comparing the prediction result with the ground truth mask, it is evident that the ground truth lacks the bright spots present in the original image but displays the retinal blood vessels more distinctly. The ground truth allows for clear observation of vascular changes and the complex network of retinal blood vessels.

In contrast, the prediction result is generated without a dedicated denoising process. Furthermore, the nature of the loss function also contributes to the differences observed between the ground truth and the prediction. The mask is used to filter specific areas, enabling better integration with the Canvas element for visualization. This helps the network model present its output more effectively.

To generate the prediction mask compatible with Canvas, the model's output is transformed during conversion: values of 0 remain 0, while values of 1 are converted to 255. Overlaying this mask with the original image produces Figure 3(d). Compared to the original image, the version with the prediction mask appears clearer and enhances visual interpretation.

Figure 4 below shows the training progress curve.

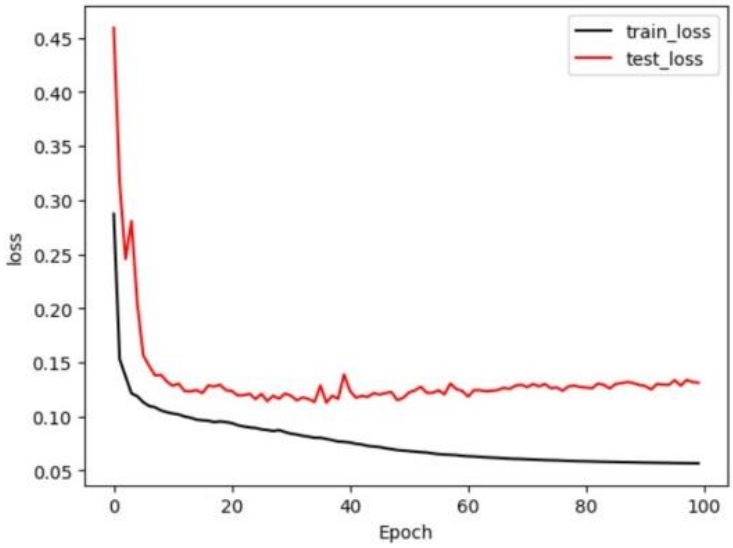


Fig.4 Training Progress Curve

As shown in Figure 4, as the number of training epochs increases, the overall trends of both train_loss and test_loss show a decreasing pattern. However, the train_loss curve exhibits relatively stable descent without significant peaks, while the test_loss curve demonstrates more pronounced fluctuations with noticeable minor peaks during its decline. This volatility in test_loss is indicative of overfitting during the training process.

5. Conclusion

The application of deep learning for processing and analyzing retinal images enables more effective automated diagnosis and assessment of retinal conditions. Deep learning algorithms can efficiently extract features from retinal images, leading to more accurate disease diagnosis and demonstrating better precision

and robustness in practical applications. After training, the U-Net model can segment different structures in input retinal images, thereby assisting physicians in more precisely locating and analyzing lesions during the diagnosis and treatment of retinal diseases. Multiple studies have confirmed that the U-Net model achieves relatively strong performance in retinal image segmentation tasks, ultimately enhancing both the efficiency and accuracy of medical diagnostics.

References

[1] Hassanzadeh Tahereh, Shamonin Denis P., Li Yanli, et al. A deep learning-based comparative MRI model to detect inflammatory changes in rheumatoid arthritis, Biomedical Signal Processing and Control, 2023, 16(2):101-114.

[2] Ebied Mostafa, Elmisery F. A., El Hag Noha A., et al. A Proposed Deep-Learning-Based Framework for Medical Image Communication,Storage and Diagnosis, Wireless Personal Communications, 2023, 131(4):2331-2369.

[3] Vukadinovic Milos, Kwan Alan C, Yuan Victoria, et al. Deep learning-enabled analysis of medical images identifies cardiac sphericity as an early marker of cardiomyopathy and related outcomes, Med (New York, N.Y.), 2023, 4(4):252-262.

[4] Chunmei Liu, Junfeng Qu, Mohammad N.A. Rana, et al. A Survey of Deep Learning Based Methods in Medical Image Processing, Current Signal Transduction Therapy，.2021, 16(2):101-114.

[5] Ibtehaz N, Rahman M S.MultiResUNet: Rethinking the U-Net Architecture for Multimodal Biomedical Image Segmentation. Neural Netvorks, 2020, 121:74-87.

[6] Isensee F, Jaeger P F, Kohl S A A, et al. NNU-Net: A Self-Configuring Method for Deep Learning-Based Biomedical Image Segmentation. Nature Methods, 2021, 18(2): 203-211.

[7] Peng Z, Peng K, Liu C, et al. DP-U-Net++: Inter-Layer Feature Fusion for Colorectal Gland Image Segmentation. International Journal of Machine Learning and Cybernetics, 2024:1-15.