Attention-DPU: Dual-path UNet with an attention mechanism for ultrasound image segmentation

Pengli Wei, Mengjun Tong*
Information Engineering, Zhejiang A&F University, Hangzhou, 310000, China
tzhuimeng@163.com, tmj@zafu.edu.cn
*Corresponding author

Abstract: With the continuous advancement in computer vision, image segmentation has achieved fruitful results in many applications such as medical image processing. In recent years, UNet and Dual Path Network (DPN) have achieved promising results in medical image segmentation. UNet cannot effectively obtain new features and reuse features. Also, DPN cannot effectively transfer the contour information of shallow blocks to the subsequent deep blocks. This paper proposes a dual path U-shaped network with the attention mechanism (Attention-DPU); taking advantage of the two networks. In the proposed network, the ordinary convolutional layer is replaced by the micro block with a dual path. Also, the attention mechanism is adopted to improve the efficiency and accuracy of segmentation.

1. Introduction
In modern medicine, ultrasound imaging plays an increasingly important role in clinical diagnosis. Image segmentation is an essential task to analyze ultrasound images. However, irregular, changeable, and noisy ultrasound images often induce poor segmentation results. Recently, image segmentation has been rapidly advanced since the emergence of a fully convolutional network (FCN) [1]. Inspired by the FCN model, Ronneberger et al. [2] proposed a fully convolutional network with symmetric structure, UNet, for medical image segmentation. UNet is composed of two paths: encoder and decoder. In the UNet, semantic features and spatial features are combined by fusing deep and shallow information. Also, the up-sampling is realized by deconvolution. However, UNet cannot effectively obtain new features and reuse features in the training process. Chen et al. [3] proposed a dual-path network (DPN), which has the advantages of ResNet [4] and DenseNet [5]. The DPN not only reused features but also explored new features so that the model can be fully trained. However, in DPN, the contour information of shallow blocks cannot be effectively transmitted to the subsequent deep blocks, which leads to the failure of accurate localization in ultrasonic image segmentation. In this paper, the two networks are combined in a U-shaped dual-path structure to take the advantages of each network while avoiding the above problems. Furthermore, the attention mechanism [6] is employed to connect deep layers to shallow layers. The segmentation accuracy is improved due to the enhanced ability of local response. The experimental results on the ultrasound images of the brachial plexus validate the performance of the proposed model.
2. The proposed Attention-DPU

2.1. Network structure of Attention-DPU

Figure 1 shows the structure of the proposed Attention-DPU network. The network architecture is based on the structure of the U-Net, consisting of two parts: encoder and decoder, where a 3×3×3 convolution is used as an initial Conv. A typical structure of a convolutional neural network (CNN), i.e., alternating convolutional and pooling layers, is used in the down-sampling process of the contraction path on the left. This process is mainly responsible for feature extraction and gradually reducing the spatial dimensions of the features. The expansion path on the right performs the up-sampling process opposite to the contraction path. In this process, the segmented parts are accurately located, and the corresponding spatial dimension is gradually restored. The expansion path consists of the up-sampling layer and the following 2x2 convolution layer. The features from contraction and expansion paths are concatenated with a skip connection. Such cross-layer feature fusion of deep and shallow features well combines the semantic and spatial features.

![Fig.1 The structure of attentional DPU](image)

2.2. Micro-block structure in DPN-U-Net model

The micro-block structure of Attention-DPU is shown in Figure 2. Arrows in Figure 2 indicate the Batch Normalization (BN) and the Rectified Linear Unit (ReLU). In the proposed model, the micro-block structure divides the convolutional feature maps into two paths: one is ResNet, and the other is DenseNet. In the residual connection, input features are summed to add the input feature to the element-based output feature. Also, the feature is reused through the residual branch channel. In the dense connection path, the concatenation method is employed to merge the features, thereby connecting the input features.
on the axis of the channel. Thus, the channel can receive information from the previous layer and explore new features through the dense connection path.

![Diagram](image)

Fig.2 The micro-block structure of Attention-DPU

2.3. Attention mechanism

Inspired by the human visual attention system, the attention mechanism in computer vision enables the model to distinguish regions as a useful feature selection function. In this paper, the attention mechanism is based on the attention-gate signals. The structure diagram of the gate signal is shown in Figure 3.

![Diagram](image)

Fig.3 The Attention mechanism structure of Attention-DPU

1) The attention-gate has two inputs: the up-sampling feature (g) of the decoder and the corresponding depth feature (f) in the encoder. The first input (g) is used as a gating signal to enhance the learning of the second input (f). All in all, the gating signal (g) can select more useful features from the coding features (f) and pass them to the upper decoder.

2) After convolution operation \((W_g, W_f)\) and normalization, the two inputs \((b_g, b_f)\) are combined pixel by pixel.

3) Then, the combined result is sent to the rectified linear unit \((RELU, \sigma_1(x) = \max(0, x))\) for activation.

4) After activation, the feature will be convolved \((W_g)\) and normalized \((b_g)\) again.
5) The sigmoid function \( \sigma_2(x) = \frac{1}{1+e^{-x}} \) is selected to train the convergence of the parameters in the gate signal and obtain the attention coefficient \((\alpha)\).
6) Finally, the output is obtained by the pixel-wise multiplication of the encoder features by coefficients.

The process of selecting features of the attention gate can be expressed as:

\[
F = \sigma_1 \left( (W^T_x \times f + b_f) + (W^T_s \times g + b_s) \right)
\]

\[
\alpha = \sigma_2 \left( W^T_x \times F + b_s \right)
\]

\[
output = f \times \alpha
\]

2.4. Loss function
Defining a proper loss function is essential to increase the convergence speed and learning effect of the model. To this end, the loss function is defined as follows:

\[
Loss = L_{BCE} + L_{DICE}
\]

where \( L_{BCE} \) represents the cost function of binary cross entropy. If the sample label is expressed as targets, the model segmentation result is expressed as logits. The expression is as follows:

\[
L_{BCE} = \text{targets} \times - \log(\text{sigmoid(logits)}) \times \text{weight} + (1 - \text{targets}) - \log(1 - \text{sigmoid(logits)})
\]

The loss of similarity coefficient, \( L_{DICE} \), can be used to measure accuracy and recall rate comprehensively. It is defined as the following formula:

\[
L_{DICE} = 1 - 2 \frac{W|T|Y| + S}{W|T| + |Y| + S}
\]

where \( T \) is the corresponding image label, \( Y \) is the result of model segmentation, \( W \) is the weight value, and \( S \) is set to 1 to avoid the denominator is 0.

3. Experiments
3.1. Dataset and pre-processing
In this paper, a series of brachial plexus ultrasound images were used for the experiment. The dataset consists of 5,000 sets of samples \((580 \times 420)\), and each set of samples had the corresponding manually labeled ground-truth. The processing of the high resolution of ultrasonic images requires a huge amount of memory. In order to improve the efficiency of the processing, this paper used pre-processing operation to divide the image with a width of 580 and a height of 420 into non-overlapping square images with lower resolution \((256 \times 256)\) for training according to certain step size.

3.2. Model evaluation indexes
In order to evaluate the performance of the proposed model, four evaluation indexes are used: Pixel accuracy (PA), Mean Intersection Over Union (MIOU), PR curve, and Dice coefficient. They are defined in the following sub-sections.

1) PA
   The PA is defined as follows:

   \[
   R_{PA} = \frac{P_r}{P_s} \times 100\%
   \]

   where \( P_r \) is the number of pixels correctly predicted and \( P_s \) is the total number of pixels.

2) MIOU
   MIOU precision is defined as follows:

   \[
   R_{MIOU} = \frac{P_t}{P_t + P_{fp} + P_{fn}} \times 100\%
   \]
where $P_T$ represents the number of correctly predicted pixels, $P_{TF}$ represents the number of pixels of the labeled brachial plexus predicted as the background, and $P_{FP}$ represents the number of pixels of the background predicted as the brachial plexus.

3) P-R curve:
In the P-R curve, the proportion of real positive cases in the predicted positive cases (precision) is taken as the horizontal axis, while the proportion of the predicted positive cases in real positive cases (recall) is taken as the vertical axis.

4) Dice coefficient
In the experiment, the Dice coefficient was used to calculate the similarity between the segmented image and the label image of the model. The Dice is defined as follows:

$$\text{Dice} = \frac{2|Y_{pred} \cap Y_{true}|}{|Y_{pred}| + |Y_{true}|}$$

where $Y_{true}$ is the sample label and $Y_{pred}$ is the segmentation result of each network. $|Y_{pred} \cap Y_{true}|$ is the overlap between the prediction result and the sample label, and $|Y_{pred}| + |Y_{true}|$ is the total amount of the prediction result and the label. The Dice coefficient is 1 when the prediction result is completely consistent with the label in the experiment, and the Dice coefficient is 0 when the prediction result is completely inconsistent with the sample label. Therefore, the higher the similarity between the prediction result and the sample label, the larger the Dice coefficient, the better the segmentation effect.

3.3 Experimental results and analysis
The performance of the proposed Attention-DPU model is verified on the brachial plexus ultrasound images, compared with the UNet model and DPN model. Table 1 compares PA and MIOU for the compared segmentation network models. As shown in Table 1, DPU provides better results than UNet and DPN in terms of the four indexes. DPU’s test set accuracy is about 3% better than DPN and about 7% better than UNet. In addition DPU’s test set iou about 1% higher than DPN and about 6% higher than UNet.

Table 1. Comparisons of Intersection-Over-Union(IOU) and recognition rate of each algorithm

| Data type       | Input   | UNet  | DPN  | Article method(DPU) |
|-----------------|---------|-------|------|---------------------|
| Ultrasound Image| 256×256 Training set Accuracy | 88.09 | 93.41 | 96.78               |
|                 | 256×256 Test set Accuracy     | 86.34 | 92.03 | 94.91               |
|                 | 256×256 Training set IOU       | 82.31 | 87.01 | 88.42               |
|                 | 256×256 Test set IOU           | 79.96 | 84.93 | 86.29               |

Figure 4 depicts the P-R curves for each network.
As shown in Figure 4, the PR-curve of our model is closer to the (1,1) coordinate point than the other two models, and basically wraps around the other two networks. So the network model in this paper performs better than the other two models.
Table 2 shows the similarity values between the effect map and the label map of six groups of randomly selected sample data. The effect maps were obtained by segmentation of the UNet model, DPN model, and the proposed Attention-DPU model. As shown in Table 2, the segmentation map of the proposed model is the most similar to the ground-truth label map. Moreover, for the similarity calculation of a single image, the Dice coefficient of the proposed model is higher than that of the other two compared networks.

Table 2. Compare the segmentation results of UNet, DPN and the method in this paper

| Test samples | label | UNet   | DPN   | Article method (DPU) |
|--------------|-------|--------|-------|----------------------|
| Sample #1    | ![Sample #1](image1.png) | ![Sample #1](image2.png) | ![Sample #1](image3.png) | ![Sample #1](image4.png) |
|              | Dice = 0.7395 | Dice = 0.7630 | Dice = 0.9359 |
| Sample #2    | ![Sample #2](image1.png) | ![Sample #2](image2.png) | ![Sample #2](image3.png) | ![Sample #2](image4.png) |
|              | Dice = 0.6943 | Dice = 0.8436 | Dice = 0.9122 |
| Sample #3    | ![Sample #3](image1.png) | ![Sample #3](image2.png) | ![Sample #3](image3.png) | ![Sample #3](image4.png) |
|              | Dice = 0.7824 | Dice = 0.9207 | Dice = 0.8968 |
| Sample #4    | ![Sample #4](image1.png) | ![Sample #4](image2.png) | ![Sample #4](image3.png) | ![Sample #4](image4.png) |
|              | Dice = 0.8061 | Dice = 0.8661 | Dice = 0.9272 |
4. Conclusion
In this paper, a novel dual-path UNet with attention mechanism (so-called Attention-DPU) is proposed for ultrasound image segmentation. It combines the characteristics of DPN and UNet in a dual-path structure. Moreover, the attention mechanism is adopted to improve the accuracy and efficiency of segmentation. In order to verify the effectiveness and advantages of the proposed Attention-DPU model, it is compared with the DPN model and UNet model in terms of four evaluation indexes. The experimental results show that the proposed Attention-DPU model achieves outperforming results for each evaluation standard.

References
[1] Long J, Shelhamer E, Darrell T. Fully convolutional networks for semantic segmentation. IEEE Trans Pattern Anal MachIntelli 2014;39(4):640–51.
[2] Ronneberger, O.; Fischer, P.; Brox, T. U-net: Convolutional networks for biomedical image segmentation. In Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, Munich, Germany, 5–9 October 2015
[3] Chen Y, Li J, Xiao H, et al. Dual Path Networks[C]. neural information processing
[4] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," pp. 770-778, 2015.
[5] Ram Deepak Gottapu, Cihan H Dagli. DenseNet for Anatomical Brain Segmentation. 2018, 140:179-185.
[6] Oktay, Ozan, Jo Schlemper, Loïc Le Folgoc, Matthew C. H. Lee, Mattias P. Heinrich, Kazunari Misawa, Kensaku Mori, Steven G. McDonagh, Nils Y. Hammerla, Bernhard Kainz, Ben Glocker and Daniel Rueckert. “Attention U-Net: Learning Where to Look for the Pancreas.” ArXiv abs/1804.03999 (2018): n. pag.