Deep Attention Unet: A Network Model with Global Feature Perception Ability

Jia-Cheng Li[1,2]
[1] University of Chinese Academy of Sciences
[2] Institute of Computing Technology Chinese Academy of Sciences
lijiacheng222@mails.ucas.ac.cn

Abstract

Remote sensing image segmentation is a specific task of remote sensing image interpretation. A good remote sensing image segmentation algorithm can provide guidance for environmental protection, agricultural production, and urban construction. This paper proposes a new type of UNet image segmentation algorithm based on channel self attention mechanism and residual connection called . In my experiment, the new network model improved mIOU by 2.48% compared to traditional UNet on the FoodNet dataset. The image segmentation algorithm proposed in this article enhances the internal connections between different items in the image, thus achieving better segmentation results for remote sensing images with occlusion.

1. Introduction

Many elements in remote sensing images have some connections, such as rivers and buildings often appearing together, and roads and vehicles often appearing together. Deep learning[8] has shown amazing results in many fields. In the field of remote sensing image segmentation, UNet has a very important position[11]. UNet was mainly used for medical image segmentation when it was proposed. After several years of development, it has also been widely used in the field of remote sensing, and has many different variants, which have been applied to different types of tasks.

The dataset I am using is the FloodNet Dataset[10], which is a multi category segmented remote sensing image dataset. The data is collected with a small UAS platform, DJI Mavic Pro quadcopters, after Hurricane Harvey. The whole dataset has 2343 images.

The classic Unet network is divided into two parts: encoder and decoder, and the encoding and decoding processes are connected through residual connections to avoid information loss during the encoding process. In Unet, the way the network obtains global features is convolution, and the network structure I designed adds a self attention module to the residual connection on the last layer of the encoder[12]. The encoder obtains global features through multiple convolutional layers, and in the final layer of the encoder, the deepest global feature representation is obtained. Then, through the self attention module, the network's understanding ability of global information is improved to improve the network's representation accuracy.

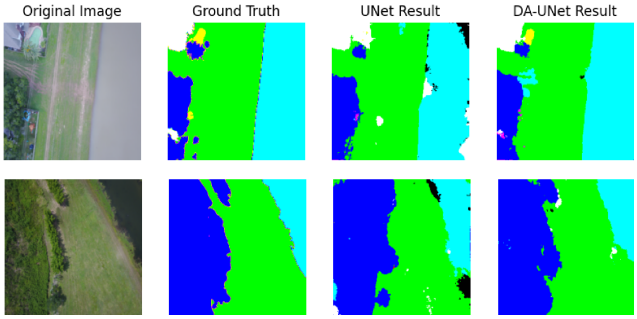

Figure 1. Different segmentation results between DA-UNet and classical UNet on FoodNet dataset. DA-UNet with a sub-attention module added to the deep residual module can more easily obtain global information and show better segmentation ability

However, directly replacing the deepest residual part of the self-attention module will lose the ability of the model to extract local information. Therefore, the residual connection[5] is added to the self-attention part, so that the self-attention module can take into account the global and local information of the image.

Since UNet was proposed in 2015, there have been many variants, and since the attention mechanism has performed prominent in many fields in recent years, there have been many works that integrate UNET with attention mechanism. The most relevant ones to my work are attention-UNet and P-UNet. However, due to the different prediction speed and focus, they are mainly applied to medical images, and are not suitable for remote sensing. The new UNet network combining self-attention mechanism and residual is more suitable for the field of remote sensing images.

In summary, the contributions of our work are three-fold:

- The self-attention mechanism is introduced into the classical UNet network, which makes it easier to ob-

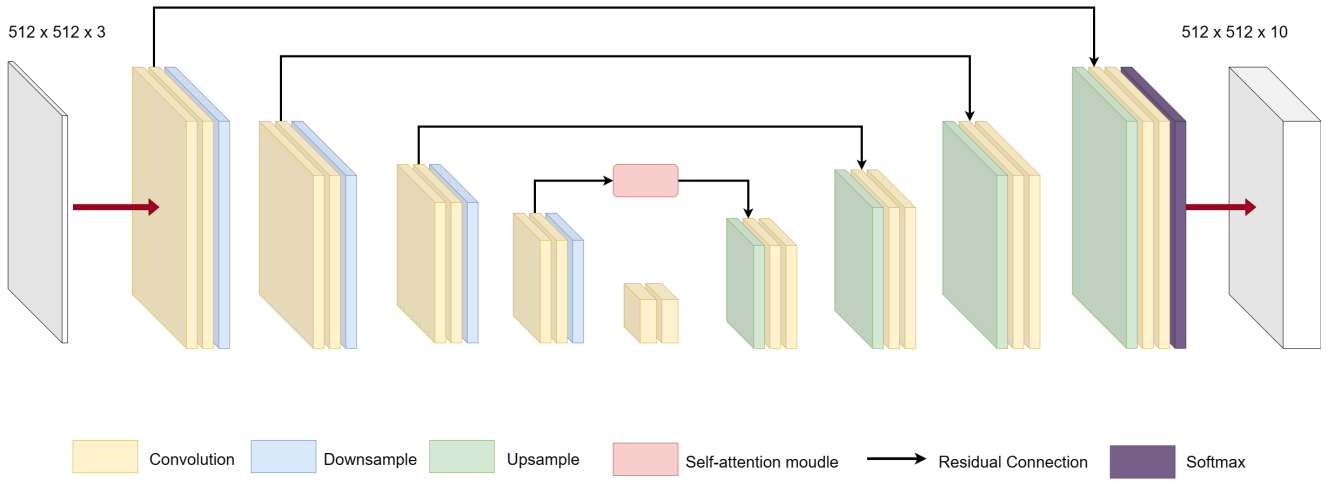

Figure 2. The network structure of DA-UNet consists of three parts: encoder, decoder and self-attention module. The encoder consists of four convolutional layers plus a downsampling layer. The decoder consists of four upsampling layers and convolutional layers. In the four feature passes from encoder to decoder, the deepest connection is replaced by the self-attention module.

tain global features and reflect them in the segmentation results.

- The residual module was added in parallel with the self-attention module, so that the network could obtain global information without losing local detail information.
- Applying the UNet network structure with residual and self-attention module to the field of remote sensing images, which may contribute to urban planning and road design.

2. Related Work

In recent years, the application of attention mechanism to the field of computer vision has gradually emerged[4, 2].There are many variants of UNet[15, 13, 6, 3]. Related works to the algorithm proposed in this paper are Parallel Attention Based UNet(P-UNet) by Xiaohu Zhang from Sun Yat-sen University[14] and Attention U-Net by Ozan Oktay et al. from Imperial College London[9].

The decoder part of classical U-Net has four upsampling modules, and P-UNet adds a channel attention module after the second upsampling module. P-UNet is designed to be applied to crack detection, which needs to pay more attention to the detail characteristics of the image, and does not consider the global characteristics of the image too much. The details of image segmentation of UNet are obtained by directly connecting the residual from the encoder with different resolutions to the decoder part, so P-UNet does not modify the residual connection part.

The model proposed in this paper is applied to remote sensing images, which needs to take into account both global features and local detail features. Therefore, the self-attention module is applied to the residual part of the deepest layer, and on this basis, the residual connection is added to retain the extraction of local detail features. Since the deepest encoder analyzes the deep meaning of the image, the self-attention module is used here to help the network understand the internal information of the image.

Attention U-Net is used primarily for medical images. It introduces an attention mechanism in the decoder, using an attention module that realigns the output features of the encoder before splicing the features at each resolution of the encoder with the corresponding features in the decoder, focusing attention on salient features useful for a particular task (such as relevant tissues or organs) and suppressing irrelevant areas in the input image. However, there are many obvious differences between remote sensing images and medical images.

In medical image segmentation, in order to obtain accurate segmentation results, we need to pay more attention to some contents. However, in remote sensing images, image segmentation requires multiple categories with equal weights between categories, so Attention Unet is not suitable for multi-category segmentation of remote sensing images. At the same time, attention-unet adds an Attention module to the features of each resolution, and the complexity of the attention module is O(N). In the field of medical images, more attention is paid to the time cost of segmentation than in the field of remote sensing, so this structure is not suitable for the segmentation of remote sensing images.

3. Model Structrue

The structure of my proposed network is shown in Figure 2. The network model structure contains three parts: an encoder part composed of a convolutional layer and a downsampling layer, a decoder part composed of an upsampling layer and a convolutional layer, and a self-attention module added to the deepest residual connection.

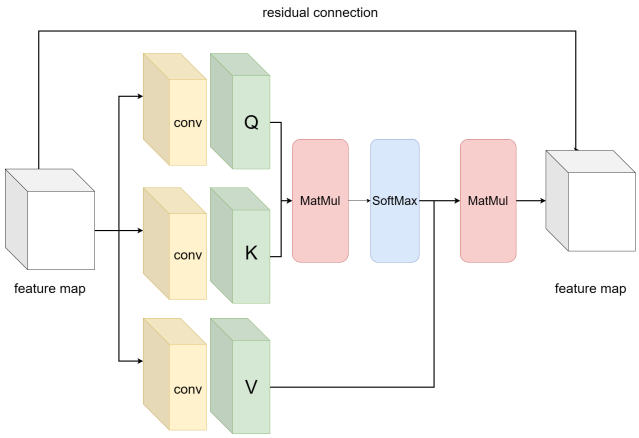

Figure 3. The self-attention module passes the input feature map through three different convolutional layers to calculate Query,Key and Value, calculates the energy tensor through Key and Query, and obtains the attention tensor through Sofmax. Finally, the Value is multiplied with the attention tensor to obtain the tensor containing the dependency information between elements. Finally, the output tensor is residually connected to the input tensor.

The encoder part is the left half of the network structure, and the decoder part of the network is composed of four convolutional layers + downsampling layers, which are used to extract feature information in different dimensions. In the decoder part, it is composed of four upsampling layers + convolutional layers, which are used to translate the features extracted by the encoder part. The final output has the same length and width as the input image, and the dimension is the number of segmentation categories. The cross-entropy loss function is used as the loss function.

At the same time, the network uses several direct connections to connect features from the encoder to the mapping decoder of the corresponding feature for feature fusion between layers, and also to prevent the loss of information during the downsampling process.

On the direct connection structure from encoder to decoder in the deepest layer, a self-attention module combined with residual connection is added. Attention modules are often used in deep learning to process sequence or image data. The main role of this module is to interact each element of an input sequence or image, such as a word or pixel, with other elements in order to better capture dependencies between elements.

$$Q, K, V = Conv_{1,2,3}(input) \tag{1}$$

This specific self-attention module employs three convolutional layers and a softmax layer. First, the input tensor x goes through three different convolutional layers to generate three different "query", "key" and "value" tensors. The "query" and "key" tensors are computed by matrix multiplication to obtain an "energy" tensor and a softmax operation to obtain an "attention" tensor. Finally, the "value" tensor is multiplied and added with the "attention" tensor to generate a new tensor.

$$Attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_k}})V \tag{2}$$

The input tensor of the self-attention module has the same dimension as the output tensor. At the same time, a residual connection is added to the self-attention module in this paper. The residual connection can effectively prevent the gradient disappearance and gradient explosion in the process of model training. At the same time, it also plays a role in preventing the loss of detail information in the network model.

$$DA\ Block(input) = input + Attention(Q, K, V) \tag{3}$$

4. Experiments

The experiment is divided into two parts in total: training with the traditional UNet network on the FoodNet dataset, and training with my DA-UNet network and recording the loss and mIOU.Due to time constraints, the experiments of P-UNet and attention-UNet were not completed.All experiments were performed on an NVIDIA Tesla A100 graphics card with 80G memory.

mIOU refers to the average intersection and union ratio between the output image of the network and the label image. m refers to the number of segmentation categories, which is 10 in this experiment.

$$mIOU = \frac{1}{k+1}\sum_{i=0}^{k}\frac{TP}{FN+FP+TP} \tag{4}$$

4.1. Segmentation experiments on pure Unet

In the experiments on UNet, the cross-entropy loss function is used as the loss function, the optimizer is Adam optimizer[7], the learning rate is 0.001, and 300 epochs are trained.

$$loss = -\frac{1}{batchSize}\sum_{i=1}^{10}[yln\hat{y} + (1-y)ln(1-\hat{y})] \tag{5}$$

Considering the different size of the image, I resize the input image and label into a 512*512 size image. Although this will lose some information of the image itself, the images obtained by the two compared networks are the same.

Some tricks are used during the training process to improve the model's ability to understand global and local information. We use 256*256 images for the first 100 epochs to train the model to obtain the global information of the

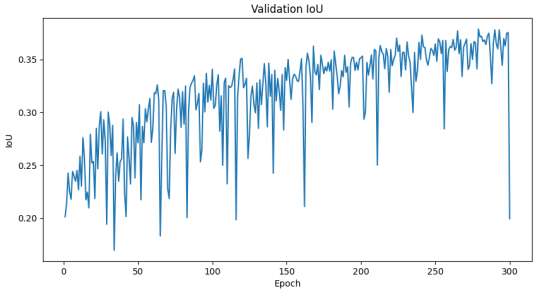

Figure 4. mIOU on val of UNet

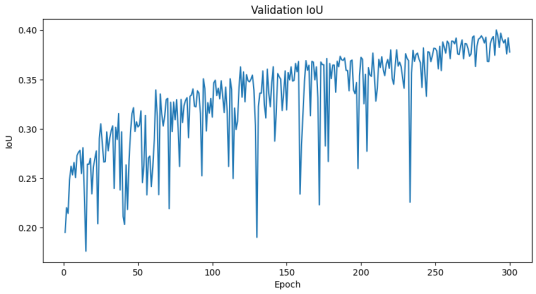

Figure 5. mIOU on val of DA-UNet without res

image, and 512*512 images for the last 200 epochs to train the model to understand the detailed information.

In Figure 4, the performance of mIOU of the UNet network model on the validation set is shown.As you can see from the figure, the mIOU fluctuates a lot, and the final best mIOU on the validation set is about 37.55%. The UNet network structure in this experiment directly adopts the structure of four layers of downsampling plus four layers of upsampling.

4.2. Segmentation experiments on DA-Unet

In the experiments of DA-UNet, two experiments are divided: no residual connection is added on the self-attention module, and residual connection is added on the self-attention module.

Like the UNet experiment, we used the cross-entropy loss function, the Adam optimizer, and a learning rate of 0.001. Figure 5 illustrates the results of DA-UNet without adding residual connections for 300 rounds. The figure shows that during the training process of DA-UNet without adding residual connections, the mIOU vibration amplitude is also very large in the first half, but gradually becomes stable in the last part. Here, the reason is that if the residual connection is not added on the attention module, the model will lose part of its attention on the local features, resulting in poor segmentation on the details and large mIOU amplitude on the validation set.

Finally, on DA-Unet without adding residual connections, the final mIOU on the validation set is 40.03%

Although the UNet model with self-attention module has

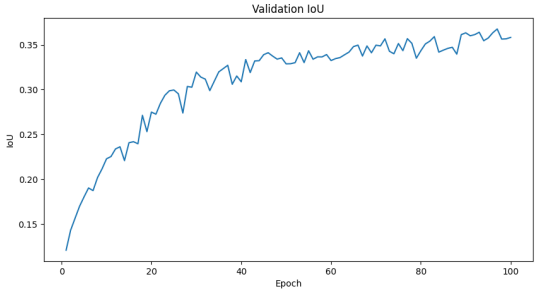

Figure 6. mIOU on val of DA-UNet with res

shown better results than the classic UNet model on remote sensing image segmentation tasks, the evaluation indicators during training are also very large. This situation also occurred despite my attempts to reduce the learning rate.

After analysis, it is caused by the insufficient attention weight of the model for the local detail information of the image. When I used the DA-UNet model without adding residual connection to segment the image, I found that the detail processing of the segmentation results was not excellent, but the overall segmentation situation was good.

In order to enhance DA-UNet's attention weight for local details of the image, I add a residual connection to the outermost layer of the self-attention module, so that the model takes into account both global and local information of the image.

In this experiment, except for the structure of the network, the rest of the parameters are completely consistent with the previously introduced experiments. Due to time constraints, only 100 epochs were trained in this experiment. Same as the predicted results, the fluctuation amplitude of mIOU of DA-UNet with residual connections is reduced very much during training, as shown in Figure 6.

Figure 7 shows the segmentation performance of UNet

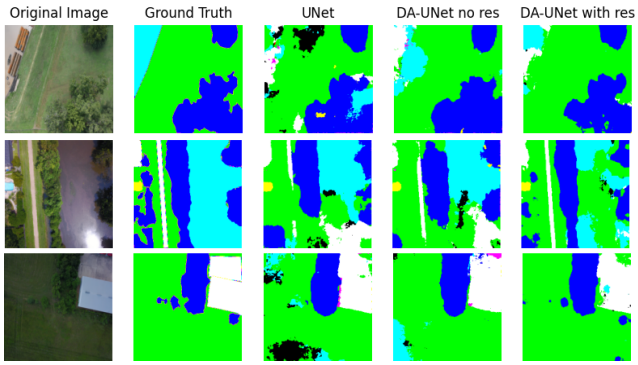

Figure 7. Performance of UNet, DA-UNet without residual connection and DA-UNet with residual connection for segmentation on FoodNet dataset. After adding the self-attention module, the segmentation results of DA-UNet fuse global information, but the detail performance is poor. After adding the residual connection, DA-UNet has better global and detail performance.

trained for 300 epochs and DA-UNet without residual connection on remote sensing images, while showing the performance results of DA-UNet with residual connection added trained for 100 epochs.

It can be found that, although the number of training epochs and skills are not as complete as other experiments, the performance of DA-UNet with residual connections on remote sensing images exceeds that of classical UNet and DA-UNet without residual connections.

5. Conclusions and Future Work

When segmenting a remote sensing image, the neural network needs to consider the relationship between the elements inside the image for the following reasons:

- There may be some dependencies between adjacent elements, for example, roads may be adjacent to elements such as buildings or trees, so if the global features of the image are not taken into account, such dependencies may not be captured, resulting in inaccurate segmentation results.
- Adjacent pixels may have similar features or attributes. If the local detail features in the image are not considered, similar regions may be divided into different parts, or different regions may be divided into the same part, resulting in inaccurate segmentation results.

When I used UNet to segment the FoodNet dataset, I found that the segmentation effect was very poor. Not only the insufficient analysis of local features led to many subtle segmentation error blocks in the segmented images, but also the insufficient ability to capture global features led to many large segmentation errors.

After discovering this problem, this paper proposes a UNet network using self-attention module and residual connection. The self-attention module is applied to the deepest encoder to decoder connection to obtain information between the deep meaning of the image. The residual connection ensures that the local detail features of the picture are preserved, so that the model can take into account the global features and local features of the picture.

I call this neural network that combines global and local features DA-UNet, which stands for UNet with Deep Attention.

However, it can be seen that whether using the classical UNet network or the improved DA-UNet, our segmentation effect on the FoodNet dataset is not excellent in general, and the evaluation index mIOU does not exceed 0.45.This happens because of two things:

- the problem of labeling the dataset. After my inspection, the FoodNet dataset has many labeling errors, including: some categories are not labeled, some label images are upside down, some labels are upside down, and some categories are even labeled incorrectly.
- the training degree of the neural network is insufficient. Due to the time constraints, more tricks were not used during model training, and the model could theoretically show better performance.

Nevertheless, DA-UNet shows superior results on the FoodNet dataset compared to traditional UNet. The convergence speed and final accuracy of DA-UNet after adding residual connection are also more excellent.

In my future work, I will build on this work, use more accurate datasets, and explore the best performance of the model in more experiments. At the same time, the self-attention module used in this paper does not use multi-head, and the multi-head attention module can be tried in the later experiments to obtain more global information.

The segmentation model in this paper is used for segmentation of still images, and a video dataset will be tried in a follow-up work if it goes well.Recently, there have been some breakthroughs in the field of image segmentation[1]. In the future work, I will improve the work of this paper in the process of continuous learning.

References

[1] Nikhila Ravi Hanzi Mao Chloe Rolland Laura Gustafson Tete Xiao Spencer Whitehead Alexander C. Berg Wan-Yen Lo Piotr Dollár Ross Girshick Alexander Kirillov, Eric Mintun. Segment anything. *arXiv preprint arXiv:2304.02643*, 2023. 5

[2] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16*, pages 213–229. Springer, 2020. 2

[3] Jose Dolz, Ismail Ben Ayed, and Christian Desrosiers. Dense multi-path u-net for ischemic stroke lesion segmentation in multiple image modalities. In *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries: 4th International Workshop, BrainLes 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 16, 2018, Revised Selected Papers, Part I 4*, pages 271–282. Springer, 2019. 2

[4] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 2

[5] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 1

[6] Qiangguo Jin, Zhaopeng Meng, Tuan D Pham, Qi Chen, Leyi Wei, and Ran Su. Dunet: A deformable network

for retinal vessel segmentation. *Knowledge-Based Systems*, 178:149–162, 2019. 2

[7] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 3

[8] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553):436–444, 2015. 1

[9] Ozan Oktay, Jo Schlemper, Loic Le Folgoc, Matthew Lee, Mattias Heinrich, Kazunari Misawa, Kensaku Mori, Steven McDonagh, Nils Y Hammerla, Bernhard Kainz, et al. Attention u-net: Learning where to look for the pancreas. *arXiv preprint arXiv:1804.03999*, 2018. 2

[10] Maryam Rahnemoonfar, Tashnim Chowdhury, Argho Sarkar, Debvrat Varshney, Masoud Yari, and Robin Murphy. Floodnet: A high resolution aerial imagery dataset for post flood scene understanding. *arXiv preprint arXiv:2012.02951*, 2020. 1

[11] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. pages 234–241, 2015. 1

[12] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 1

[13] Jiawei Zhang, Yanchun Zhang, Yuzhen Jin, Jilan Xu, and Xiaowei Xu. Mdu-net: Multi-scale densely connected u-net for biomedical image segmentation. *Health Information Science and Systems*, 11(1):13, 2023. 2

[14] Xiaohu Zhang and Haifeng Huang. P-unet: Parallel attention based unet for crack detection. In *2022 7th International Conference on Signal and Image Processing (ICSIP)*, pages 311–315. IEEE, 2022. 2

[15] Zongwei Zhou, Md Mahfuzur Rahman Siddiquee, Nima Tajbakhsh, and Jianming Liang. Unet++: A nested u-net architecture for medical image segmentation. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support: 4th International Workshop, DLMIA 2018, and 8th International Workshop, ML-CDS 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 20, 2018, Proceedings 4*, pages 3–11. Springer, 2018. 2