Deeplabv3+ semantic segmentation model based on feature cross attention mechanism

Haibo Zeng¹, Siqi Peng², Dongxiang Li³

¹ School of Automation and Electronic Information, Xiangtan University, China
² School of Automation and Electronic Information, Xiangtan University, China
³ School of Automation and Electronic Information, Xiangtan University, China

*Haibo Zeng e-mail: Haibo Zeng.@xtu.edu.cn, 2926288575@qq.com

Abstract: Aiming at the problem that the deeplabv3+ model is not accurate in segmentation of the image target edge, the image feature fitting is slow, and the attention information cannot be effectively used. It is proposed to add a feature cross attention module (FCA) to the model. The cross-attention network is composed of two branches and a feature cross attention module. Among them, the shallow branch is used to extract low-level spatial information, and the deep branch is used to extract high-level context features to make important feature extraction more refined. This paper designs and realizes the connection between Feature Cross Attention module and Deeplabv3+ coding module, input the output features of the Deeplabv3+ encoding module into the feature cross attention module for convolution operation to realize the recalibration of the original features. The decoding module of Deeplabv3+ obtains spatial features and channel features from two branches respectively, and then merges the obtained features to obtain more important features. The improved model was validated by the Pascal Voc2012 data set, and the results showed that the ratio of average intersection and the average pixel accuracy were increased by 1.96% and 2.84%, respectively. The model added with FCA can effectively improve the shortcomings of the original model, can segment the target more finely, and better solve the problem of rough segmentation boundary.

1. Introduction

With the development of deep learning, the semantic segmentation technology based on convolutional neural networks is gradually mature. In 2015, Long et al. proposed full convolution networks for semantic segmentation (FCN) [1] for semantic segmentation of natural image processing. FCN is a pioneering work of deep learning segmentation network. It realizes the semantic segmentation for the first time in an end-to-end form. The retrained imagenet network is used in the segmentation problem, and deconvolution [3] is used for upsampling. It is proposed to use a jump structure to improve the roughness of upsampling, and it has achieved good results in natural image processing. Since then, FCN based image semantic segmentation algorithms have been proposed like mushrooms. However, FCN based algorithms have limited ability to segment small and complex objects, and the network segmentation graph is not precise enough. In order to solve these problems, some excellent semantic segmentation models that reuse low-level feature maps came into being, such as U-Net [7], Segnet [8]. These algorithms all use codec structures. Among them, the encoder uses the pooling layer to gradually reduce the spatial dimension of the input data. The decoder uses the deconvolution to wait for the network layer to gradually recover the target details and the corresponding spatial dimension. From the
encoder to the decoder, there is usually a direct information connection, to help the decoder better recover the target details and improve the segmentation accuracy of the network. Fisher et al. [9] first proposed the concept of hollow convolution, this article makes ResNet [10] keep the parameter amount unchanged and the convolutional layer view of each stage unchanged, the later convolutional layer can also maintain a larger feature maps size to facilitate the detection of small targets, improve the overall performance of the model, but the hollow convolution makes the calculation of the entire network larger. Pan et al. [11] proposed a novel Dense Pyramid Network (DPN) for semantic segmentation. The network separately extracts the feature map of each channel and performs channel switching operations to enhance the representation ability of the network. In 2018, Shen L [12] et al. proposed a simple and effective attention module for feedforward convolutional neural networks, followed by Jie Hu et al. proposed SEnet, and Sanghyun Woo et al. proposed CBAM [4] attention Mechanisms, they all use the attention mechanism to infer the important feature features of feature pixels. They all use the attention mechanism to speculate the important pixel features among the feature pixels. Recently, Deeplab[2,6,13,14] series of networks were proposed by Liang-Chieh Chen and the Google team, it is a model dedicated to processing semantic segmentation, and currently has 4 versions, they all have problems with less edge modification of the segmentation results, resulting in rough borders of some image segmentation, the relationship between the distance pixel categories cannot be fully utilized.

For these problems, this article addresses the problem of semantic segmentation of Deeplab series networks, we propose to introduce the feature cross-attention mechanism FCA [5] into the Deeplabv3+ model, and propose a Deeplabv3+ model based on the cross attention mechanism. Our model emphasizes meaningful feature information in the channel and spatial dimensions, and performs a convolution operation to redistribute the weight to distinguish the importance of the pixel feature. The more important the pixel feature, the greater the weight, then the image segmentation is obtained by the joint learning of the main branch and the cross-attention module. The attention mechanism is a simple and effective lightweight module. Adding this module hardly adds extra computation. After the introduction of Deeplabv3+, due to the selective attention to important information by the attention mechanism, the improved network area division is more accurate, and it can segment the ideal target area. It can accurately segment the edges of objects, effectively avoiding the problem of unreasonable semantic segmentation and labeling.

2. Methods

2.1. Deeplabv3+ model

The Deeplab series network was proposed by Liang-Chieh Chen and the Google team. It is a model specifically used to deal with semantic segmentation. Currently, four versions have been launched. Deeplabv1 [13] was rewritten based on the VGG16 network, first removing the last fully connected layer to achieve end-to-end output. Then the last two pooling layers are removed, because convolution itself has translation invariance, and pooling can further enhance this feature of the network, Because pooling itself is a process of fuzzy position, and semantic segmentation is an end-to-end problem, each pixel needs to be accurately classified, and it is very sensitive to the position of the pixel. Too much pooling is used, the size of the feature layer is too small, and the included features are too sparse, which is not conducive to semantic segmentation, so we have to remove some pooling. Secondly, the hollow convolution is added. Its advantage is that it increases the density of features and expands the receptive field. And use conditional random field CRF to improve the classification accuracy, but Deeplabv1 has poor processing ability for multiscale segmentation objects. In order to solve this problem, in Deeplab v2 [6], they feel that VGG16 has limited expressive power, and replaced with ResNet-101, which is more complex and expressive, an ASPP (Atrous Spatial Pyramid Pooling) structure is proposed. This structure uses the hole convolution operation with different sampling rates for the input feature map to sample in parallel, that is multiscale capture of image context information on the feature map. Deeplabv3[14] improves ASPP, uses hole convolution to deepen the network, and discards CRF, because the accuracy of the classification results has been improved to no longer need CRF, for the improvement
of ASPP, use 1×1 convolution, that is, when the rate increases, the degenerate form of 3×3 convolution, instead of 3×3 convolution, reduce the number of parameters, another point is to add image pooling, which can be called global pooling, to supplement global features. Deeplabv3+[2] modified the main network again on the original basis, and upgraded ResNet-101 to Xception. Three modifications were made on the basis of the original Xception, using a deeper network to replace all convolutional layers and pooling layers with deep separable convolutions, BN and ReLU are used after each 3×3 depth separable convolution. Deeplabv3+ is mainly implemented by an encoder and a decoder, the encoder is divided into deep separation convolution and ASPP layers, and the decoder merges low-level features and performs feature map recovery. The separable convolution is also discussed, which makes the proposed model faster and stronger, and significantly reduces the computational complexity of the proposed model.

2.2. Two attention branch modules

2.2.1. Channel attention module

The attention mechanism has been continuously used in various fields of deep learning in recent years, and it has performed well in image processing, speech recognition, and natural language processing. SEnet [15] learns feature weights according to loss through the network, give important feature maps large weights, invalid or unimportant feature maps small weights to train the model to achieve better results, and SEnet is embedded in some of the original classification networks with few added parameters and calculations. In this paper, the method of extracting the attention of feature channels is basically similar to SEnet. The feature extraction method of maxpool is added on the basis of SEnet. The final output is the result of adding the average pooling result and the maximum pooling result. The feature extraction method with avgpool is the same as the avgpool extraction method in Senet. In addition, when using these two pools, use a shared MLP to pay attention to inference to save the parameters, the two aggregated channel features are located in the same semantic embedding space, and we use the channel attention module as shown in Figure 1.

![Figure 1: Channel Attention Module (CA)](image)

Each channel of the channel attention module features represents a special detector. It is meaningful for channel attention to focus on what kind of features. In order to summarize the spatial characteristics, two methods of global average pooling and maximum pooling are used to respectively use different information. The operation process of the channel attention module is shown in formula 1.

\[
M_c = \sigma (MLP (AvgPool \ F)) + MLP (MaxPool \ F)) = \sigma (W_1(W_0(F_{avg})) + W_1(W_0(F_{max}))
\]  

Among them: MLP is the multisensing layer; \(\sigma\) is the sigmoid activation function. The input is a \(H \times W \times C\) feature \(F\), We first perform global average pooling and maximum pooling of a space to obtain two \(1 \times 1 \times C\) channel descriptions. Then send them to a two-layer neural network, the number of neurons in the first layer is \(C/r\), the activation function is Relu, and the number of neurons in the second layer is \(C\). This two-layer neural network is shared. Then add the two features and get a weight coefficient \(M_c\) through a Sigmoid activation function. Finally, the weighted coefficient is multiplied by the original feature \(F\) to obtain the scaled new feature.

2.2.2. Spatial attention module.

The spatial attention mechanism focuses on where the meaningful features are. Its model diagram is shown in Figure 2.
Similar to channel attention, given a feature $F$ of $H \times W \times C$, we first perform an average pooling and a maximum pooling of one channel dimension to obtain two $H \times W \times 1$ channel descriptions, and splice the two descriptions together according to the channels. Then go through a convolution layer with a convolution kernel size of $7 \times 7$, the activation function is Sigmoid, and the weight coefficient $M_s$ is obtained. Finally, multiplying the weight coefficient and the feature $F'$ is the new feature. Its operation process is shown in formula 2.

$$M_s(F) = \sigma(f^{7\times7}([\text{AvgPool}(F); \text{MaxPool}(F)]))$$

$$= \sigma(f^{7\times7}([F_{avg}^{s}; F_{max}^{s}]))$$

(2)

Among them: $\sigma$ is the sigmoid activation function; $f$ is the convolution layer; $[;]$ is the connection feature map in the channel dimension.

3. Improved model

3.1. Feature Cross Attention module

The feature cross-attention module uses the spatial attention module to extract shallow spatial information, and then uses the channel attention mechanism to capture contextual information. Its model diagram is shown in Figure 3. The output features of these two branches are different. Since the high-level features are mainly composed of category information, the channel attention module can be used to extract the high-level information. The low-level features correspond to more spatial information, which cannot be directly sampled and fused. You can use the spatial attention module to extract features from the low-level features. The high-level features of the channel attention module of the FCA we added are used to provide context information, while the low-level features extracted by the spatial attention module are used to refine pixel positioning. We first cascade the output features of the two branches, and perform convolution, batch normalization and ReLU unit processing on the cascaded features. Then the SA module’s fused features and the output of the spatial branch are used as inputs to help refine the positioning. The features of the SA module are normalized and S-shaped nonlinear convolution, and then multiplied by the fused features. The space of the context peak output is applied to the channel attention block, and the context features are compressed along the spatial dimension through global pooling and maximum pooling to obtain two vectors. Then the two vectors are shared to a fully connected layer and Sigmoid operator to generate the attention graph, and finally convolution, batch normalization and ReLu unit fusion.

$$\text{Figure 3: Feature Cross Attention Module (FCA)}$$

3.2. Network architecture

The DeepLabv3+ network is one of the most excellent semantic segmentation models at present, but there are some shortcomings in the models. In order to increase the ability to segment multi-scale targets,
DeepLabv3+ connects to the ASPP structure after extracting the network with hollow convolutional features. The structure consists of 3×3 hole convolutions with expansion ratios of 1, 6, 12, 18, respectively, and global average pooling operations. Excessive expansion rate cannot accurately extract the target features of the image edge, and at the same time it cannot completely simulate the relationship between the local features of the large-scale target, so that there is a hole phenomenon in the large-scale target segmentation. These results in the DeepLabv3+ network segmentation accuracy of remote sensing image edge targets and large-scale targets reduced. Therefore, this paper proposes a DeepLabv3+ model with a cross-attention mechanism to make up for the above deficiencies, the modified network model is shown in Figure 4.

In order to better mine the spatial and channel features in the decoder, we use the 1×1 convolution to extract shallow features after the hollow convolution of the original image, and input the extracted shallow features into the SA attention module to use get better shallow spatial information. After the original image is convoluted with holes and the spatial pyramid pooling operation is performed, the feature map is obtained. After the feature map is quadruple upsampled, the feature is input into the channel attention mechanism to obtain higher-level context channel information. First, the output features of the two branches are concatenated, and then 3×3 convolution, batch normalization, and replay units are applied to the cascaded features. Spatial branch features undergo 3×3 convolution, batch normalization and Sigmoid nonlinearity, and then multiplied by the fusion feature. The output of the spatial attention block and the context features of the context branch are applied to the channel attention block. By compressing the context features along the spatial dimension through the global pool and the maximum pool, two vectors are obtained. These two vectors are then applied to the shared fully connected layer and Sigmoid operator to generate attention maps. Next, the attention map is multiplied by the output feature from the spatial attention block and added to the fused feature. Then, the obtained features are subjected to 1×1 convolution and four times upsampling to obtain a feature map. This solves the problem of reduced accuracy of edge target and large-scale target segmentation. And the addition of modules is faster than the original network fitting. With almost no additional training parameters and overhead, our model can obtain more important spatial and channel feature maps.

4. Experiment

4.1. Data set and strategy
This experiment uses the public data set Pascal Voc2012 data set for verification. The segmentation set contains 21 categories, including 1464 training pictures, 1449 verification pictures and 1456 test pictures. It is an important database in image semantic segmentation tasks. We expand the data set with additional annotations and get 10582 training images. The algorithm uses the mIoU value to measure the performance of the algorithm. We use DeepLabv3+, DeepLabv3+ +spatial, DeepLabv3+ +channel, and this model DeepLabv3+ +FCA these four networks to carry out the experiment respectively, and use the average intersection ratio (mIoU), average pixel accuracy (mPA) and accuracy indicators for each
network. As an important criterion for measuring the accuracy of semantic segmentation algorithms, it reflects the degree of coincidence between the predicted value of the same category and the true value. The higher the IoU value, the higher the overlap between the measured value and the true value, and the more accurate the network prediction. mIoU (Mean Intersection over Union) is the mean value of IoU for all categories. Another metric is the PA (Pixel Accuracy) value, which reflects the overall degree of network prediction of the image. The larger the PA value, the better the network prediction. The experimental platform is the graphics card NVIDIA GeForce GTX1070Ti, the CPU is Intel i7-8700k, our batch size is 4, the initial learning rate is set to 0.0001, the learning rate uses a polynomial decay method, the total decay rate steps are 4000 steps, and the momentum is 0.9.

4.2. Attention Embedding Experiment

In order to verify the effectiveness of adding the FCA module in the Deeplabv3+ structure, the influence of two asynchrony lengths on the model is designed, the results are shown in Table 1. The model setting stride is 16 and the model recognition result mIoU added to the FCA module in the data test set is 1.97% higher than the original model. When stride is set to 8, the advantages of adding the FCA module in the network are also obvious. The segmentation results are better than the original network performance, and the segmentation result mIoU on the test set is 1.95% higher than the original model, the segmentation results of the two structures are better than the original network performance.

| Module       | Stride | mIoU/% |
|--------------|--------|--------|
| Deeplabv3+   | 16     | 72.25  |
| Deeplabv3+   | 8      | 72.39  |
| +FCA         | 16     | 74.22  |
| +FCA         | 8      | 74.34  |

In order to further verify the structural algorithm designed in this paper, Table 2 shows the comparison between the improved network and other classic networks we reproduced in the data set test in the mPA value, mIoU value and the accuracy of the entire network. It can be seen that the performance of the network with the FAC module added on the Pascal V oc2012 data set is better than the original model and our reproduced classic network model in all indicators. This article also verified the effect of the network segmentation on the Deeplabv3+ model with only the CA module or only the SA module, and found that adding SA or CA module alone is better than the original model. The recognition result of our model with FCA is 1.96% higher than the original Deeplabv3+ mIoU value, the mPA value is 2.84% higher, and the accuracy rate is also improved by 0.36%.

| Module  | mPA   | mIoU  | Accuracy |
|---------|-------|-------|----------|
| FCN-8S  | 79.83 | 70.23 | 89.14    |
| SegNet  | 80.24 | 70.89 | 89.68    |
| Deeplabv3 | 81.12 | 71.54 | 90.16    |
| Deeplabv3+ | 82.48 | 72.32 | 90.53    |
| +CA     | 84.85 | 73.62 | 90.72    |
| +SA     | 84.49 | 73.34 | 90.70    |
| +FCA    | 85.32 | 74.28 | 90.89    |

In order to verify the performance of each model in terms of time performance, a statistical analysis of the time spent testing a picture is performed, and the data set used is still verified with the Voc2012 data set. The results are shown in Table 3. It can be seen that after adding CA, SA, FCA modules, the average result of testing a picture hardly increases. It is enough to show that the attention module we added is a lightweight module with almost no additional time consumption and additional calculations.
Table 3: Comparison of average time of different model test results

| Module          | Time/ms |
|-----------------|---------|
| Deeplabv3+      | 632     |
| Deeplabv3+ +CA  | 637     |
| Deeplabv3+ +SA  | 635     |
| Deeplabv3+ +FCA | 641     |

Figure 5 shows the division result of whether to add FCA or not. It can be seen that the algorithm combined with FCA training is better than the algorithm without FCA on the overall image, edges and details. Appropriately adding FCA to the Deeplabv3+ network is simple and effective, and almost does not increase the cost. The joint study of the main branch and FCA parameters has improved the average cross combination ratio and the average pixel accuracy. The FCA module uses the features between different pixels to effectively enhance the same target features at the edge of the image, thereby accurately segmenting the edge target. After improvement, it can effectively compensate for the original network defects, and the edge target segmentation effect is better. When the segmentation target is too large, the local feature extraction is not coherent, which is easy to cause inconsistencies within the large-scale target class, and the large-scale target segmentation is incomplete. By adding FCA, the long term context dependent information between image positions can be effectively simulated, and different local feature information can be connected. The attention module can use the intraclass correlation of different channels to make the segmentation target consistent. It can be seen that our model is better than the original model in terms of overall, edge, and details. The network with the FCA model added to the edge details can learn to use the information in the target area and aggregate features from it. The feature refinement process of our model will eventually guide the network to rationally use the given features.

5. Conclusion and discussion

Our Deeplabv3+ model with the introduction of the FCA module can refine the characteristics of the attention mechanism into two different modules, in the case of keeping a small amount of calculation, to achieve better performance improvement, design a two branch network to improve context characteristics, and at the same time encode low-level spatial information, we propose that an FCA model can effectively ablate these two features, use channel features to provide global information, and use spatial features to refine the target edge with higher segmentation accuracy. Experiments in the Pascal Voc2012 dataset show that the Deeplabv3+ model with FCA has a higher mIoU, mPA equivalent value than the original model, and more detailed segmentation edges.

References
[1] Shelhamer Evan, Long Jonathan, Darrell Trevor. Fully Convolutional Networks for Semantic
Segmentation.[J]. IEEE transactions on pattern analysis and machine intelligence, 2017, 39(4).

[2] Chen L C, Zhu Y, Papandreou G, et al. Encoder-decoder with atrous separable convolution for semantic image segmentation [C]// Proceedings of the European conference on computer vision (ECCV). 2018: 801-818.

[3] H. Noh, S. Hong and B. Han, "Learning Deconvolution Network for Semantic Segmentation," 2015 IEEE International Conference on Computer Vision (ICCV), Santiago, 2015, pp. 1520-1528, doi: 10.1109/ICCV.2015.178.

[4] Woo S, Park J, Lee J Y, et al. CBam: Convolutional block attention module[C]//Proceedings of the European conference on computer vision (ECCV). 2018: 3-19.

[5] Liu M, Yin H. Cross Attention Network for Semantic Segmentation[C]//2019 IEEE International Conference on Image Processing (ICIP). IEEE, 2019: 2434-2438.

[6] Chen L C, Papandreou G, Kokkinos I, et al. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs[J]. IEEE transactions on pattern analysis and machine intelligence, 2017, 40(4): 834-848.

[7] Ronneberger O, Fischer P, Brox T. U-net: Convolutional networks for biomedical image segmentation[C]// International Conference on Medical image computing and computer-assisted intervention. Springer, Cham, 2015: 234-241.

[8] Badrinarayanan V, Handa A, Cipolla R. Segnet: A deep convolutional encoder-decoder architecture for robust semantic pixel-wise labelling [J]. arXiv preprint arXiv:1505.07293, 2015.

[9] Yu F, Koltun V. Multi-scale context aggregation by dilated convolutions[J]. arXiv preprint arXiv:1511.07122, 2015.

[10] He K, Zhang X, Ren S, et al. Deep residual learning for image recognition[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 770-778.

[11] Pan X, Gao L, Zhang B, et al. High-resolution aerial imagery semantic labeling with dense pyramid network[J]. Sensors, 2018, 18(11): 3774.

[12] Hu J, Shen L, Sun G. Squeeze-and-excitation networks[C]// Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 7132-7141.

[13] Chen L C, Papandreou G, Kokkinos I, et al. Semantic image segmentation with deep convolutional nets and fully connected crfs[J]. arXiv preprint arXiv:1412.7062, 2014.

[14] Chen L C, Papandreou G, Schroff F, et al. Rethinking atrous convolution for semantic image segmentation[J]. arXiv preprint arXiv:1706.05587, 2017.

[15] Hu J, Shen L, Sun G. Squeeze-and-excitation networks[C]// Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 7132-7141.