Journal of Artificial Intelligence Practice (2023)
Clausius Scientific Press, Canada

DOI: 10.23977/jaip.2023.060104
ISSN 2371-8412 Vol. 6 Num. 1

Research on semantic segmentation of unmanned aerial vehicle visual image based on deep learning—take the outdoor environment of Anhui University of Finance & Economics as an example

Lei Zhang[1], Shiyu Fang[1], Daijin Li[1], Xiangrong Xue[1], Xinlei Wu[1], Hao Wu[1,*]

[1]School of Management Science and Engineering, Anhui University of Finance & Economics, Bengbu, Anhui, 233030, China
**Corresponding author*

Keywords: Deep learning, Image semantic segmentation, Convolution neural network, UAV vision

Abstract: In recent years, with the rapid development of unmanned aerial vehicle (UAV) technology, UAV processing visual information, especially image semantic segmentation technology, has developed rapidly. This paper proposes a semantic segmentation model, which has achieved high accuracy on CityScapes data set, and has been verified on the newly collected data set, and the verification results are in line with the actual situation.

1. Introduction

With the continuous development of UAV technology, the demand of UAV visual information processing is more and more extensive, especially the demand of UAV autonomous image analysis and target recognition is more and more intense. This requires the UAV itself to have certain artificial intelligence. At present, the mainstream scheme is to carry the artificial intelligence image semantic segmentation model using deep learning training for UAV. Image semantic segmentation is to use depth learning to mark different kinds of objects in the image at pixel level [1], so as to assist the machine to locate the target object, so as to carry out the next operation. It is of great significance in UAV navigation, UAV recognition of obstacles and flying areas. The principle of image semantic segmentation is to obtain the features of the input image through the depth convolution neural network, and then sample the feature map to output the category of each pixel point, and the color of the pixels of the same category is consistent [2]. In practical applications, an original captured image is input into the semantic segmentation neural network, and different object categories are marked with different colors in the output image. This paper will first introduce the research history and basic theory of semantic segmentation methods, then propose a semantic segmentation model based on the characteristics of UAV visual images, and apply the model to the newly collected data sets, and finally prospect the future development of image semantic segmentation.

2. Related research

2.1. Traditional semantic segmentation method

The traditional semantic segmentation method probably appeared before 2010. Due to the limited computing power of the computer, this method mainly extracts the low-level features of the image (color, texture, shape, spatial relationship and other features), and then disjoint the different objects in the image. The traditional semantic segmentation method is generally unsupervised learning, and the segmentation result has no semantic annotation [3]. However, because traditional semantic segmentation methods cannot learn image features independently, they usually need manual feature processing, model selection and optimal parameter search, which not only consumes time and energy, but also has poor results and accuracy of model segmentation. Among the traditional semantic segmentation methods, the most commonly used method is based on graph theory. With the increase of the diversity and complexity of image scenes and the increase of the size of data sets, the traditional semantic segmentation methods are no longer applicable.

2.2. Semantic segmentation method combining deep learning and traditional methods

The semantic segmentation method combining deep learning and traditional methods probably appeared between 2010 and 2015. With the improvement of computer computing ability, this method combines the traditional semantic segmentation method based on low-level features of the image with deep learning to automatically annotate different object categories in the image, which is convenient for image segmentation in the later stage [4]. Among the semantic segmentation methods combining deep learning and traditional methods, the most commonly used is the Laplace pyramid algorithm. The algorithm trains the input image by building a convolution neural network, and classifies the features of each pixel of the input image. However, the algorithm still uses a part of traditional semantic segmentation, so the segmentation accuracy is low.

2.3. Image semantic segmentation method based on deep learning

With the further improvement of computer data processing ability, the algorithm model of using deep learning to solve image semantic segmentation has been constantly proposed, and the structure of deep learning semantic segmentation network has become more and more complex, including FCN [5], SegNet [6], DeepLab [7-10], RefineNet [11] and PSPNet [12].

The image semantic segmentation method based on deep learning can be divided into three categories according to the different labeling conditions of the dataset: supervised semantic segmentation model, unsupervised semantic segmentation model and semi-supervised semantic segmentation model, which correspond to supervised learning, unsupervised learning and semi-supervised learning respectively [13]. Supervised semantic segmentation refers to the training of annotated data sets. Sometimes, in order to improve the accuracy of model training, complex data preprocessing and data enhancement are also required, and the labor cost is high, but the corresponding accuracy of this model is higher than the other two models; Unsupervised semantic segmentation refers to the training of unlabeled data sets, which can better learn the statistical rules or potential structures in the data, and has good adaptability to complex and diverse environmental images, with low labor costs, but the accuracy of semantic segmentation is relatively poor; Semi-supervised semantic segmentation refers to training some labeled and some unlabeled data sets, usually with a small amount of labeled data and a large amount of unlabeled data. This segmentation method uses the information in the unlabeled data to assist the labeled data for supervised learning. It requires a lot of time for model training, and its segmentation accuracy is

lower than that of supervised semantic segmentation.

3. Basic theory of image semantic segmentation

3.1. The basic structure of convolution neural network

The basic structure of convolution neural network includes convolution layer, activation function, pooling layer, full connection layer and loss function.

(1) Convolution layer

The convolution layer is used to extract the information in the image, that is, to extract the feature data used to input the depth learning model. Each convolution layer uses multiple convolution cores to extract information, and the parameters such as the size and step size of convolution cores are set manually. The features that the convolution kernel may extract include texture features, color features, etc.

(2) Activation function

Convolution is a linear calculation. Even if the convolution layer is continuously superimposed and the depth of the model is increased, the model is still a linear structure and cannot deal with nonlinear problems. Most of the problems that need to be dealt with are nonlinear problems and cannot be well simulated with linear structures. Therefore, it is necessary to introduce nonlinear activation functions, so that the depth learning model can simulate any nonlinear problem, that is, the universal approximation principle. The result of convolution operation is input into the activation function, that is, the result of linear operation is given nonlinear characteristics, and the result is used as the output of the next layer. Common activation functions include sigmoid, ReLU, Tanh, etc.

(3) Pooling layer

After many convolution operations, the image will obtain a large number of image features. The pooling layer is used to reduce the dimension of the image feature data obtained from the convolution layer, while preserving the key features, so as to reduce the amount of computation required, speed up the model training rate, prevent overfitting, and improve the model generalization ability. Commonly used pooling layers include maximum pooling layer and average pooling layer [14].

(4) Full connection layer

The fully connected layer is generally used as the output layer of the convolutional neural network, and its main function is to combine features and act as a classifier. Because the convolution layer can only extract the features of the image, but many objects may have the same type of features, so only local features can not determine the specific category. The full connection layer can comprehensively judge the category of objects according to the image features extracted from the network through the convolution layer and the pooling layer, and play the role of "classifier".

(5) Loss function

The loss function is used to measure the gap between the output predicted value and the real value [15], and then use the gap value to fine-tune the network structure parameters, and continuously reduce the loss value to obtain the optimal parameters. The process of adjusting the parameters is completed by the optimizer, and the cycle training is carried out to reduce the loss value.

3.2. Data enhancement

Data enhancement refers to the expansion of the existing training data set. The main methods

include geometric transformation, color space transformation and random erasure of the original data set. The model can extract more information from the training enhanced data set, thus narrowing the gap between the training set and the verification set, reducing over-fitting, and improving the generalization ability of the model.

4. Construction of image semantic segmentation model

4.1. Main research contents

In this paper, an image semantic segmentation algorithm model for UAV outdoor scene is proposed. The algorithm model will be tested on the unmanned vehicle intelligent driving dataset, and the model will be verified on the newly collected dataset. The main research contents are as follows:

(1) Build image semantic segmentation model. The CityScapes data set is used for model training, the model structure and parameters are adjusted through the model accuracy, and the experimental results are evaluated.

(2) Validate the model on the newly collected data set, and evaluate the effectiveness and generalization ability of the model through the prediction results.

4.2. Overall model design

The general idea of image semantic segmentation algorithm is as follows:

(1) Input image and extract image features using depth convolution neural network

(2) Up-sampling the feature map and outputting the category of each pixel

(3) Use the loss function to optimize the model and optimize the classification result of each pixel to the nearest real value

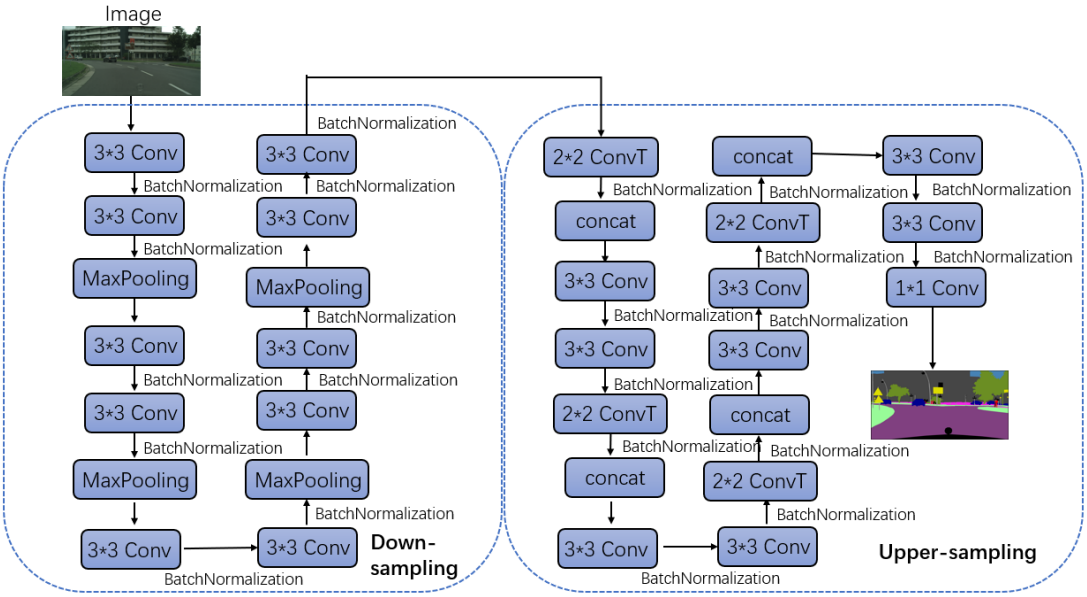

Figure 1: Overall model structure.

5. Image semantic segmentation experiment and result analysis

5.1. Experimental environment

The parameters of the experimental platform used in the experiment are shown in Table 1.

Table 1: Experimental platform parameters.

Experimental platform indicators	main parameter
operating system	Linux
GPU	NVIDIA GTX 3090
Video storage	24G
CUDA	11.0
development tool	Pycharm
Python	Python3.8
Tensorflow version	2.4.0

5.2. Experimental data set

At present, there is no reliable data set for model training of UAV image semantic segmentation algorithm, while the outdoor environment of UAV and UAV is similar, and the driving/flying area has certain similarity. For deep learning, a better model can also obtain better results by retraining parameters with a small amount of data on data sets of other similar scenes. Therefore, this paper selects the unmanned vehicle intelligent driving dataset CityScapes for model training and structural verification.

5.3. Data preprocessing and training strategy

5.3.1. Data enhancement and preprocessing

In order to extract more information from the model and reduce over-fitting, this experiment carried out data enhancement operations on all training data, including random flipping, random scaling, random clipping and other methods. The scale of random scaling is 0.75, 1.0, 1.5, 2.0. In order to facilitate batch processing during training, the data set is cut into a unified scale in this experiment. The original size of CityScapes dataset is 2048 × 1024 size, this experiment will cut it and then scale it to 256 × 256 size. At the same time, for all the input images, this experiment carries out the normalization operation, that is, subtract the average value of the whole image pixel from each pixel of the image.

5.3.2. Training strategy

The parameter update method of this experiment adopts Adam method. Adam method combines Momentum method and AdaGrad method.

The update formula of Momentum method is shown in Formula (1) and (2).

$$v^{(n+1)} = \alpha v^{(n)} - \eta \frac{\partial L}{\partial \omega_i^{(n)}} \tag{1}$$

$$\omega_i^{(n+1)} = \omega_i^{(n)} + v^{(n+1)} \tag{2}$$

Where, $v^{(n)}$ represents the momentum value at the iteration n, α represents the momentum coefficient. The momentum coefficient used in this experiment is 0.9, η indicates the learning rate

at iteration n, $\omega_i^{(n)}$ represents the weight value of the i.th weight in the n iteration.

The update formula of AdaGrad method is shown in equations (3) and (4).

$$h^{(n+1)} = h^{(n)} + \frac{\partial L}{\partial \omega_i^{(n)}} \odot \frac{\partial L}{\partial \omega_i^{(n)}} \tag{3}$$

$$\omega_i^{(n+1)} = \omega_i^{(n)} - \eta \frac{1}{\sqrt{h^{(n+1)}}} \frac{\partial L}{\partial \omega_i^{(n)}} \tag{4}$$

Where, η Indicates the learning rate at iteration n, $\omega_i^{(n)}$ represents the weight value of the i.th weight in the n iteration, the variable $\boldsymbol{h}$ saves the sum of squares of all previous gradient values. When updating parameters, the learning scale can be adjusted by multiplying by $\frac{1}{\sqrt{\boldsymbol{h}}}$.

5.4. Model validation

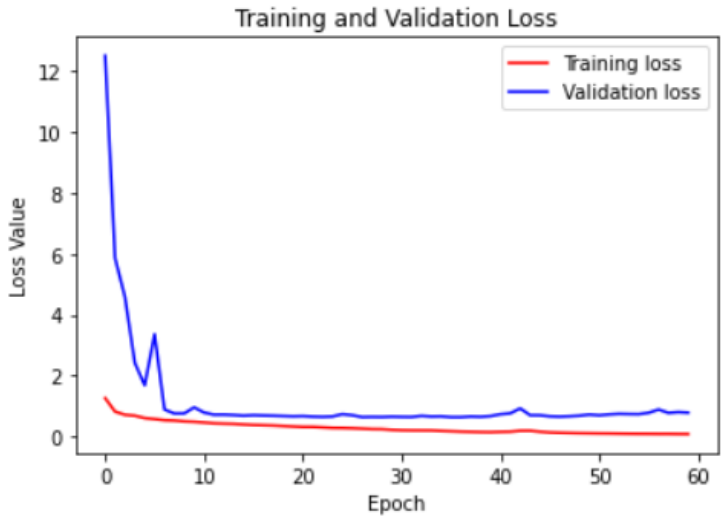

Figure 2: Visualization of training process loss function.

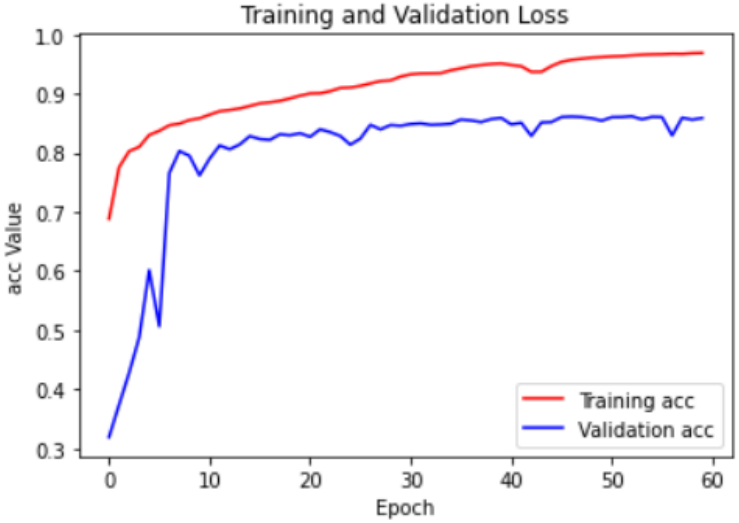

Figure 3: Visualization of training process accuracy.

The model is trained on the CityScapes dataset. The loss rate and accuracy rate of the training process are shown in Figure 2 and Figure 3.

The newly collected data set is the outdoor environment of Anhui University of Finance and Economics. Put the new data set into the model for prediction and observe the output results. Figure 4 is the original input image, and Figure 5 is the prediction result diagram.

Figure 4: Original input image.

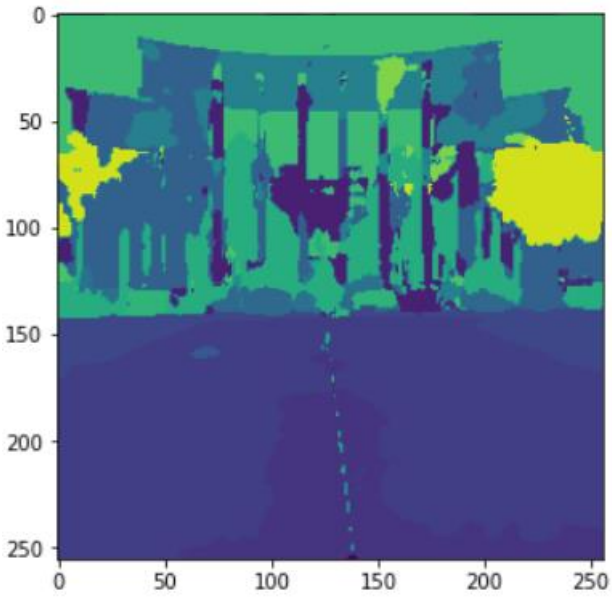

Figure 5: Forecast result image.

The result of semantic segmentation of the new dataset is in good agreement with the actual situation, so the model can better complete the task of image semantic segmentation.

6. Conclusions and prospect

This paper proposes an image semantic segmentation algorithm model for the outdoor scene of UAV. The algorithm model has achieved a high accuracy rate on the unmanned vehicle intelligent driving data set. At the same time, the model is verified on the newly collected data set, and the verification results are in line with the actual situation.

At present, image semantic segmentation still has some limitations. First, when the image is input to the network at the initial stage, because the convolution kernel of CNN is small, the model can only use local information to understand the input image, which may affect the distinguishability of the final extracted features of the encoder. Secondly, the task of image semantic segmentation is the classification of pixel level. When a 512 * 512 image is labeled for

segmentation tasks, the number of labeling required is theoretically 512 * 512 times that of image classification tasks. Because of this, the input acquisition of segmentation requires a lot of resources. Thirdly, the model generalization ability of image semantic segmentation is poor.

Acknowledgements

This paper is funded by the Innovation and Entrepreneurship Training Program for College Students of Anhui University of Finance & Economics, with the project number of 202110378068. The ownership of the project research results belongs to Anhui University of Finance & Economics.

References

[1] Xueliang Jing. Design of Semantic Vision SLAM System Based on Deep Learning [D]. Beijing University of Posts and Telecommunications, 2020.
[2] Qing Xu. Research on vegetation extraction from high-resolution remote sensing images based on attention mechanism [D]. Wuhan University, 2020.
[3] Na Zhao. Research on Semantic Segmentation Network for Rivet Surface Defect Detection Based on U-Net++ [D]. Donghua University, 2022.
[4] Qing Cheng, Man Fan, YanDong Li, Yuan Zhao, Chenglong Li. A Survey of Semantic Segmentation of UAV Aerial Images [J]. Computer Engineering and Applications, 2021.
[5] SHELHAMER E, LONG J, DARRELL T. Fully convolutional networks for semantic segmentation [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39(4):640-651.
[6] BADRINARAYANAN V, KENDALL A, CIPOLLA R. SegNet:A deep convolutional encoder-decoder architecture for image segmentation[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39(12):2481-2495.
[7] CHEN L C, PAPANDREOU G, KOKKINOS I, et al. Semantic image segmentation with deep convolutional nets and fully connected CRFs [J]. arXiv:1412.7062, 2014.
[8] CHEN L, 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, 2018, 40(4):834-848.
[9] CHEN L C, PAPANDREOU G, SCHROFF F, et al. Rethinking Atrous Convolution for Semantic Image Segmentation [J]. arXiv:1706.05587,2017.
[10] CHEN L C, ZHU Y, PAPANDREOU G, et al. Encoderdecoder with atrous separable convolution for semantic image segmentation[C]//2018 European Conference on Computer Vision(ECCV), 2018:833-851.
[11] LIN G, MILAN A, SHEN C, et al. RefineNet:Multi-path refinement networks for high-resolution semantic segmentation[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition(CVPR), 2017.
[12] ZHAO H, SHI J, QI X, et al. Pyramid scene parsing network[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition(CVPR), 2017.
[13] Zhen Wang. Research and application of image semantic segmentation algorithm based on depth neural network [D]. Jiangxi Agricultural University, 2022.
[14] Ziwen Chen. Research on quantitative methods of urban visual attribute perception [D]. Xiangtan University, 2020.
[15] Shuangshuang Lei. Analysis and improvement of the generating countermeasure image repair network based on multi-column convolution [D]. Southwest Jiaotong University, 2021.