Pixel-level semantic segmentation based on gradient features

Xiaoshuo Jia*, Zhihui Li, Kangshun Li
School of Computer Science, Guangdong University of Science and Technology, Dongguan 523079, Guangdong, China

ABSTRACT

Due to the mechanism of pooling and convolutional layers, many important features and the correlation between the features are lost in the forward propagation process in the pixel-level semantic segmentation tasks. Therefore, here we analyze the edge features of the image by means of second-order difference, propose gradient features and design the corresponding gradient convolution layer. Based on the gradient convolution layer, we use the residual structure to achieve the fusion of high-resolution gradient features and low-resolution gradient features. Finally, we designed the GraDNet. In the tests on the Cityscapes and ADE20K datasets, GraDNet achieves the best results in both accuracy and speed compared to some SOTA algorithms.

Keywords: Pixel-level semantic segmentation, second-order difference, residual structure, convolution layer

1. INTRODUCTION

Semantic segmentation is a hot topic in computer vision. Likewise, pixel-level semantic segmentation is an important and complex task, and feature extraction is a difficult task. In the traditional feature extraction algorithm, the Roberts and Prewitt operators extract the edge features of the image through the first-order difference, and the Sobel operator obtains the edge features of the image through the second-order difference. In the effect of image segmentation, Sobel operator is superior to Roberts and Prewitt operator. On the basis the Sobel operator, the Robinson operator adds 8 convolution kernels in different directions to ensure that the extracted information is more accurate. However, the parameters of the traditional algorithm are fixed, so the generalization ability of this kind algorithm is relatively weak.

The CNN algorithm has achieved excellent results in image classification, image segmentation, target tracking and other directions in Kaggle[1] and AI Challenger[2] competitions by using the characteristics of multi-parameters. FCN[3] uses the deconvolution to upsampling, which make the extracted features more detailed. U-Net[4] utilizes the network symmetric structure to fuse high-dimensional features and low-dimensional features, which can weight edge features. In CPFNet[5], the dilated convolution is proposed which can expand the field of the convolutional layer to extract more features, and then combine the inception module to achieve context-based feature fusion, which achieve superior results in medical datasets. STDC[6] performs multiple scale fusion processing on images based on FPN[7], which increases the diversity of image features, so the accuracy rate is superior to the CPFnet algorithm. BiseNetV2[8] adopts a bilateral segmentation structure on the STDC, namely Detail Branch and Semantic Branch. Detail Branch can obtain more low-level features by expanding the channel of convolution layer. Semantic Branch expands the receptive field of convolution layer through a lightweight convolution layer, which can obtain more high-level features. At the same time, the Semantic Branch also solves the problem of structural redundancy. Although the CNN-based algorithm has high accuracy, due to the problem of the pooling layer mechanism, it is easy to cause the extracted features to lose a lot of spatial information. In the end, there are some problems such as redundant network structure, large amount of computation, and segmentation errors in semantic segmentation.

Here, we utilize second-order difference to design gradient convolutional layers (Gra layer). The Gra layer extracts the gradient features in the image through second-order difference, which can retain the spatial location information of the features. Compared with traditional convolution layer, Gra layer can effectively preserve the spatial information of features. Finally, on Gra layer, GraDNet is designed on the Resnet's residual structure[9], which not only achieves contextual information fusion, improves accuracy, but also makes the model lighter. Under the datasets of Cityscapes[10] and ADE20K[11], GraDNet compared with some SOTA algorithms. The experimental results show that GraDNet has certain advantages in terms of accuracy and speed. This paper makes three contributions:

* gxnujiashuo@163.com

Third International Conference on Computer Science and Communication Technology (ICCSCT 2022)
edited by Yingfa Lu, Changbo Cheng, Proc. of SPIE Vol. 12506, 125066V
 · 0277-786X · doi: 10.1117/12.2661794

- This paper uses the second-order difference method to design the Gra layer.
- On the Gra layer, this paper uses the residual structure to realize the fusion of contextual information, and designs GraDNet.
- GraDNet compared with some SOTA algorithms under two deferent datasets. The effectiveness of GraDNet comprehensively reflected from four indicators.

2. METHOD

2.1 Gra layer

In the second-order difference formula and the Pierce correlation coefficient theory, we know that there is no correlation between two independent discrete data, and the gradient features of the image edge position are more obvious. So we here stipulate that the step size l of the sliding window should be less than size of the window, ie (l<h,l<w), The schematic diagram of the Gra layer algorithm is shown in Figure 1.

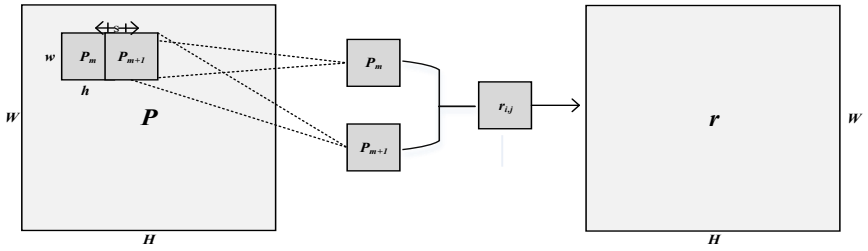

Figure 1. Schematic diagram of the Gra layer.

Algorithm 1. Image second-order difference processing

Input: P_m is the feature map obtained by the feature map P through the sliding window, K*C is the size of P_m;

Output: g_m is the gradient feature of P_m, same size as P_m;

λ represents a hyperparameter used to determine whether there is an edge feature in P_m.

if delta(P_m) <λ

then

for k=0 to K-1 do

for c=0 to C-1 do

$$g_{m|(k,c)} = P_{m|(k+1,c)} + P_{m|(k,c+1)} - 2* P_{m|(k,c)}$$

return g_m

else

return g_m=0

The feature P_m, m∈{1,2,3,…,M} intercepted by the sliding window from the feature map P is calculated by Algorithm 1 to obtain the corresponding gradient feature g_m, m∈{1,2,3,…,M}.

Algorithm 2. Gradient feature extraction

Input: P_m and P_{m+1} are two continuous feature maps，where the K*C is the size of

P_m ∘

Output: r_m is the corresponding gradient features of P_m

for k=*l* to K do

for c=*l* to C do

if $P_{m|(k,c)} == P_{m+1|(k-l,c-l)}$

then $r_{m|(k,c)} = P_{m|(k,c)}/2$

else $r_{m|(k,c)} = 0$

return r_m

Two consecutive feature maps P_m and P_{m+1} obtain the final gradient feature r_m through Algorithm 2, as shown in Figure 1.

2.2 GraDNet

In terms of features, the Gra layer extracts the gradient features of the image, which can ensure the spatial position invariance of edge features during the forward propagation process. In terms of structure, this paper uses the residual structure to realize the information fusion of high-dimensional resolution and low-dimensional resolution, and further realizes the precise positioning of edge features. The structure diagram of GraDNet is show in Figure 2. The network structure parameters are show in the following Table 1.

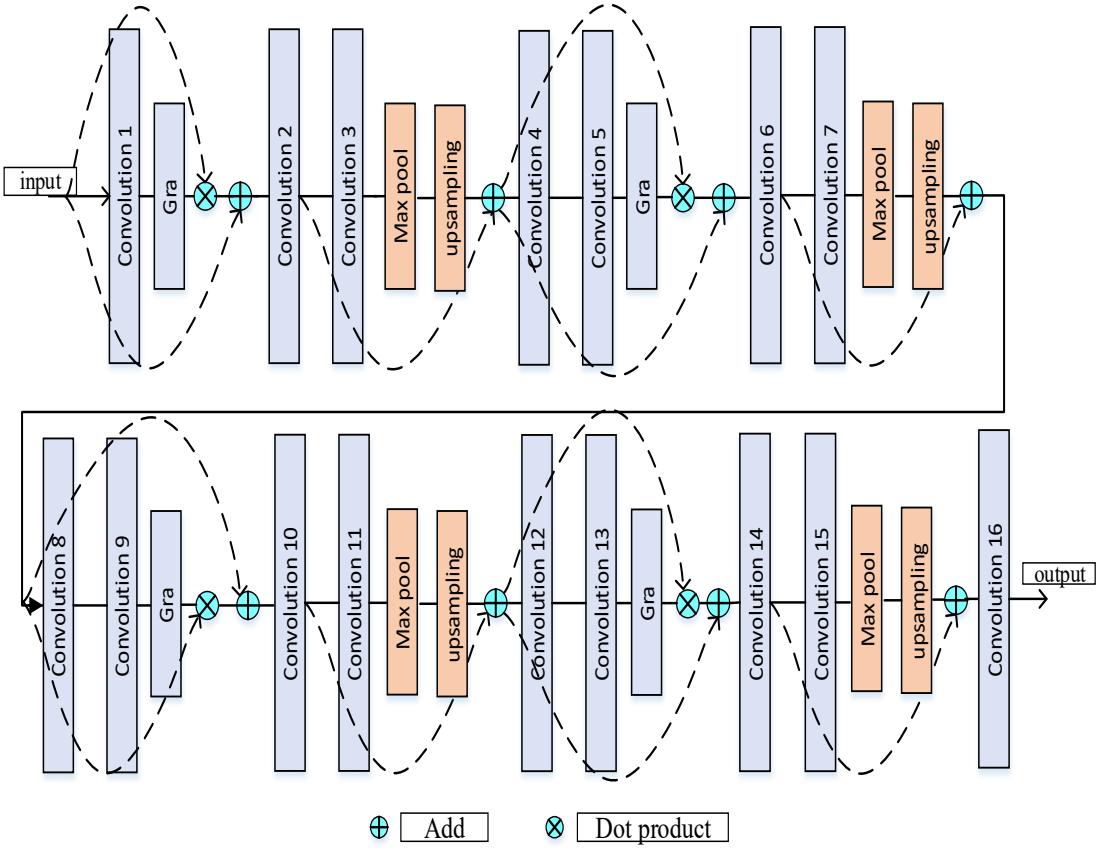

Figure 2. The structure of GraDNet.

Table 1. Network parameters of GraDNet.

Layer	Kernel/stride
Conv1	128*11*11/4
Gra	3*3/1
Conv2/6/10/14	16*1*3/2
Conv3/7//11/15	16*3*1/2
Conv4/8/12	8*1*5/2
Conv5/9/13	8*5*1/2
Max pool	3*3/1
Conv16	1*1*1/1

First, the input data pass to the 11*11 convolutional layer，which can expand the range of visual field extraction. Then, that data pass to the Gra layer to obtain the gradient features. The gradient features are dot-multiplied with the input data, and finally added to the input data. In this way, the edge features can be strengthened on the one hand, and the correlation between features can be preserved on the other hand. Therefore, the point multiplication is to calibrate the spatial position of the edge features, and the purpose of addition is to strengthen the information of the edge features and preserve the correlation between the features. The GraDnet structure refers to the residual effect of Resnet, which effectively extracts edge features on the one hand, and makes the model more lightweight on the other hand.

3. EXPERIMENT

3.1 Datasets

The Cityscapes dataset contains 5000 fine images, of which 2975 are training images, 500 validation images and 1525 testing images. In addition, the dataset contains 20k roughly annotated images. However, the performance of the algorithm is evaluated on the average precision metric of the dataset's 8 semantic classes.

The ADE20K dataset has 25k images, of which the training set is 20k, the validation set is 2k, and the test set is 3k. This dataset covers various annotations for scenes, objects. There are a total of 150 different scenes and objects, with an average of 19.5 instances and 10.2 object classes per image class.

3.2 Comparison with the some SOTA algorithms

We first crop the images of the ADE20K and Cistyscapes datasets to a size of 500*500, and set the initial learning rate to 1*e-6 and epoch to 24000. The training platform is Ryzen 7 3800X and RTX 2070. The optimize function is the Adam optimizer. The loss function calculates the error between the true value and the predicted value using Equation (1), and evaluates the test results using mIoU. yp and yt represent the predicted and actual values, respectively.

$$\mathrm{dice}(p,t) = 2^{*} | y_p \cap y_t | / (| y_p | + | y_t |) \tag{1}$$

Here we compare GraDNet with some SOTA algorithms, such as Deep snake[12], Unet, PANet[13], FCIS[14], ESE[15], etc. The results for the two datasets are shown in Tables 2 and 3 below respectively.

Table 2. Comparison results of ADE20K datasets.

Network	GraDNet	Deep (2021)	UNet (2015)	PANet (2018)	FCIS (2017)	ESE (2019)	STS
AUC (%)	73.8	67.4	68.6	70.2	66.7	65.3	69.0
fps	20.1	12.3	16.4	13.6	13.2	15.1	14.7
Model (Mb)	22.1	30.6	28.7	32.6	26.1	28.8	32.1

Table 3. Comparison results of Cityscapes datasets.

Network	GraDNet	Deep (2021)	UNet (2015)	PANet (2018)	FCIS (2017)	ESE (2019)	STS
AUC (%)	88.6	82.4	78.4	86.5	79.4	68.6	63.2
fps	28.6	14.6	18.6	17.5	16.2	18.7	16.5
Model (Mb)	22.1	30.6	28.7	32.6	26.1	28.8	32.1

Judging from the accuracy results of the two datasets, GraDNet can achieve a good result. UNet and PANet can enhance the features of corresponding locations by concatenating data dimensions. Through feature splicing, the high-resolution features are enhanced by the low-resolution features, so the edge features can be accurately segmented by locating the enhanced features. However, due to the problem of feature loss in the pooling layer, the segmentation position is inaccurate. FCIS and ESE will be trained on the fully convolutional network by means of encode-decode. FCIS can effectively avoid the problem of inaccurate information caused by the loss of features in the pooling layer, but it will also reduce the calculation speed due to the excessive number of convolutional layers. Here we extract relevant feature about the edge features of the image through the Gra layer. GraDNet uses these features to enhance the edge feature and achieve feature positioning, and then achieve the effect of image segmentation, as shown in Figure 3. Compared with Deep and Unet, GraDNet's Gra layer can obtain more gradient features, so that fewer convolutional layers used to achieve the effect of segmentation.

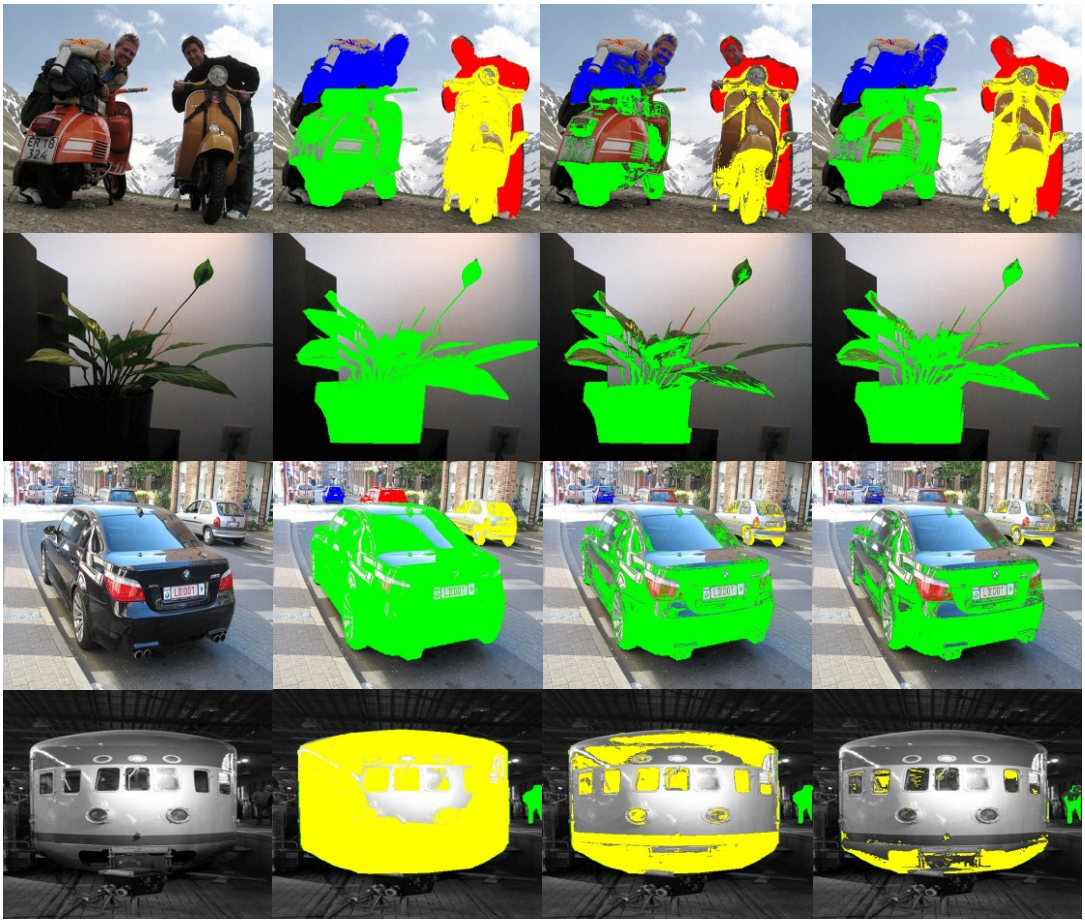

Figure 3. Image is the original image. And GraDNet, Deepnet, Unet correspond to the segmentation effect image of these algorithms respectively.

From the comparison results in Figure 3, it can be directly seen that GraDNet can accurately segment the target edge. Unet and Deep snake cannot accurately locate the fine boundary contour, and have the problem of inaccurate segmentation for small volume targets. It can be seen from the comparison results of segmentation renderings and accuracy that GraDNet has certain advantages.

4. CONCLUSION

In this paper, the gradient feature is proposed through the second-order difference, the Gra layer is designed, and the GraDNet is designed with the residual structure. In terms of feature, the features extracted by the Gra layer are not only more expressive, but also reduce the loss of features and improve the fault tolerance rate. Structurally, the extracted features are fused with context information under the residual structure, so as to achieve the effect of semantic segmentation. Compared with Deep Snake, GraDNet has a 9% increase in accuracy and an 8fps increase in speed. Under the ADE20K and Cityscapes datasets, a comprehensive comparison with some SOTA algorithms is carried out, which proves that GraDNet has good performance in terms of accuracy, model size and speed.

ACKNOWLEDGMENTS

This work is supported by The Natural Science Foundation of Guangdong Province under the Grant No.2020A1515010784, The National Natural Science Foundation of China under the Grant No.61976063 and Natural Science Program of Guangdong University of Science and Technology under the Grant No. GKY-2021KYQNK-2.

REFERENCES

[1] Iglovikov, V., Mushinskiy, S. and Osin, V., "Satellite imagery feature detection using deep convolutional neural network: A kaggle competition," arXiv preprint arXiv:1706.06169, (2017).
[2] Wu, J., Zheng, H., Zhao, B., et al., "Large-scale datasets for going deeper in image understanding," 2019 IEEE International Conference on Multimedia and Expo (ICME), IEEE, 1480-1485(2019).
[3] Long, J., Shelhamer, E., Darrell, T., "Fully convolutional networks for semantic segmentation," Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 3431-3440(2015).
[4] Ronneberger, O., Fischer., P., Brox, T., "U-net: Convolutional networks for biomedical image segmentation," International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, Cham, 234-241(2015).
[5] Feng, S., Zhao, H., Shi, F., et al., "CPFNet: Context pyramid fusion network for medical image segmentation," IEEE Transactions on Medical Imaging, 39(10), 3008-3018(2020).
[6] Fan, M., Lai, S., Huang, J., et al., "Rethinking BiSeNet for real-time semantic segmentation," Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 9716-9725(2021).
[7] Lin, T. Y., Dollár, P., Girshick, R., et al., "Feature pyramid networks for object detection," Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2117-2125(2017).
[8] Yu, C., Gao, C., Wang, J., et al., "Bisenet v2: Bilateral network with guided aggregation for real-time semantic segmentation," International Journal of Computer Vision, 129(11), 3051-3068(2021).
[9] He, K., Zhang, X., Ren, S., et al., "Deep residual learning for image recognition," Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 770-778(2016).
[10] Cordts, M., Omran, M., Ramos, S., et al., "The cityscapes dataset for semantic urban scene understanding," Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 3213-3223(2016).
[11] Everingham, M., Eslami, S. M., Van Gool, L., et al., "Assessing the significance of performance differences on the pascal voc challenges via bootstrapping," Technical Note, 1-4 (2013).
[12] Peng, S., Jiang, W., Pi, H., et al., "Deep snake for real-time instance segmentation," Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 8533-8542(2020).
[13] Armato, S. G., Roberts, R. Y., Mcnitt-Gray M. F., et al., "The Lung Image Database Consortium (LIDC) and Image Database Resource Initiative (IDRI): A completed reference database of lung nodules on CT scans," Academic Radiology, 14(12), 1455-1463(2007).
[14] Xu, W., Wang, H., Qi, F., et al. "Explicit shape encoding for real-time instance segmentation," Proceedings of the IEEE/CVF International Conference on Computer Vision, 5168-5177(2019).
[15] Zhou, X., Wang, D. and Krahenbuhl P., "Objects as points," Arxiv preprint arxiv:1904.07850, (2019).