A CNN-BASED SEGMENTATION MODEL FOR SEGMENTING FOREGROUND BY A PROBABILITY MAP

Kunming Luo, Fanman Meng, Qingbo Wu, Wen Shi, Lili Guo

School of Electronic Engineering, University of Electronic Science and Technology of China

ABSTRACT

This paper proposes a CNN-based segmentation model to segment foreground from an image and a prior probability map. Our model is constructed based on the FCN model that we simply replace the original RGB-based three channel input layer by a four channel, i.e., RGB and prior probability map. We then train the model by constructing various image, prior probability maps and the groundtruths from the PASCAL VOC dataset, and finally obtain a CNN-based foreground segmentation model that is suitable for general images. Our proposed method is motivated by the observation that the classical graphcut algorithm using GMM for modeling the priors can not capture the semantic segmentation from the prior probability, and thus leads to low segmentation performance. Furthermore, the efficient FCN segmentation model is for specific objects rather than general objects. We therefore improve the graph-cut like foreground segmentation by extending FCN segmentation model. We verify the proposed model by various prior probability maps such as artifical maps, saliency maps, and discriminative maps. The ICoseg dataset that is different from the PASCAL Voc dataset is used for the verification. Experimental results demonstrates the fact that our method obviously outperforms the graphcut algorithms and FCN models.

Index Terms— Segmentation, Grabcut, Probability Map

1. INTRODUCTION

In computer vision, we usually face the task of segmenting foreground from an image and a prior probability map. For example, in saliency detection[1, 2, 3, 4], we usually first obtain a saliency probability map, and then segment the saliency objects from the saliency map. In discriminative map detection [5, 6, 7, 8], the discriminative regions need to be segmented by the image and the discriminative map. In foreground segmentation task, a foreground probability map is usually first obtained, and is then used to guide the segmentation from the images.

In general, the classical graphcuts [9, 10] method is used to accomplish such task. The graphcut is a minimization algorithm that solves the classical Markov Random Field (MRF) energy, which consists of two terms, unary term (probability map) and pairwise term. The unary term describes the probabilities of a pixel belonging to foreground and background, and pairwise term measures the similarities of neighbour pixels. Given an image, the segmentation is formulated as searching a cut line that has the minimal cut cost, which can be globally obtained by graphcuts algorithm.

In general, the unary term of graphcuts is usually generated by GMM model, where the colors are used for modeling the objects. Since the color feature is a low-level feature that lacks semantic information, the GMM model can not capture global semantic information. When the prior is not good enough such as a part of the objects as commonly appearing in discriminative map, graphcut will lead to the unsuccessful segmentation.

To remedy such shortage, we need to use the global and semantic information. CNN-based segmentation structure[11, 12, 13, 14, 15, 16] is an efficient tool to provide global and semantic segmentation, which discovers the semantic feature by deeply convolving the image. Recently, CNN-based segmentation model such as FCN [17] has been successfully used in semantic segmentation. However, most related methods focus on the specific object segmentation such as “Person” and “Car”, rather than the general objects. A few of methods use CNN for general object segmentation, such as DeepMask [18], SharpMask [19], and FastMask[20]. However, these methods focus on proposal generation, where the foreground maps are not considered.

In this paper, we solve such task by CNN structure that segments foregrounds from an image and a foreground map. Our idea is to capture the semantic information by modifying the input of FCN segmentation model, i.e. modifying the input of FCN as the combination of the image and a prior probability map. Such simple modification makes our model general to object class. To train the model, we consider many types of images and prior probabilities, where the challenging images from the PASCAL VOC dataset are selected, and multiple probability maps of one image are considered. By observing the effectiveness of iteratively implementing our model by updating the prior probability as the output, an iteration process of our model is proposed to improve the performance. We verify our method on kinds of foreground probability map, such as saliency map, discriminative map, and artificial map. ICoseg that is totally different from PASCAL

978-1-5386-2159-2/17/$31.00 ©2017 IEEE

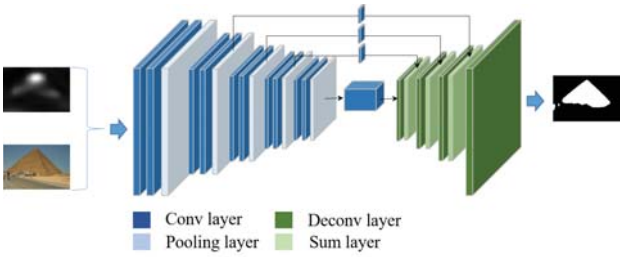

Fig. 1. The structure of the proposed model. The input of the model is an image along with its foreground map and the model can segment out the object related to the prior.

VOC dataset is used for the verification. The experimental result demonstrates that our model can obtain 0.79 average IOU value on ICoseg dataset by considering saliency map that is a totally automatic manner compared with the 0.76 average IOU value obtained by the best existing automatic manner, 0.74 mIoU by the original FCN model.

2. PROPOSED METHOD

2.1. Network Structure

Our task is to segment out the target object based on the image and its prior map. We define this task as foreground-background segmentation. In other words, the task of our model is to identify the object label of each pixel. The output of the model is a mask that assigns foreground and background label to every pixel of the input image. The model of FCN [17] has achieved unprecedented success in the task of semantic segmentation, which benefits from its skip layer structure that fuses the multi-scale features extracted by the convolutional neural network. Our model is designed based on FCN model. In order to achieve better segmentation results, we employ the FCN-8s model as the base model and make improvements on this basis.

The detailed structure of our network is shown in Fig 1. The first half of the network is the Conv-ReLU-Pooling structure of the VGG16 network, then the output feature maps of the pool2 layer, pool3 layer and pool4 layer are respectively connected to the corresponding outputs of the deconvolutional layers by convolutional layers. Finally, the segmentation mask is obtained by a 2 times up-sample deconvolutional layers. Considering that more low-level local information is more better for the network to understand and analyze the correlation between foreground map and target object, we add more skip-layer structures to take advantage of the local detail information. Our experiments have found that the results with more low-level local information is better than with less of that.

The input of the proposed model is an image with its corresponding foreground map. And the size of the input is $W \times H \times 4$, where W and H respectively represent the width and height of the image and 4 channels means three channels of the input image and one channel of the foreground map. The objective of the proposed model is a mask corresponding to the object pointed out by the foreground map. This mask contains two labels: 1 for the foreground pixel and 0 for the background pixel, and the size of the output mask is $W \times H$. The motivation we use such setting is as follows:

- First, we expect that with the same input image, different object can be segmented out according to different foreground map;
- Second, it will make the model more general when the output of the model is a binary map. Because the network do not need to know what the object is, but just segment it out;
- Third, the existing segmentation method can only segment out objects with fixed categories, which is limited by the training data. We expect that the model trained in this way will be able to break the limitation of the training dataset, which means that our model can be used to segment out objects with categories that the training dataset do not have.

2.2. Iteratively Processing

In practice of implementing our model, we find out that the initial foreground prior is not accurate enough.

- The foreground maps sometimes only point out part of the target object, thus the edge information of the target object is very poor;
- There is a lot of noise in the foreground map, which makes the foreground confused with background and thus affects the segmentation quality seriously;
- The input foreground map with more edge information and less noise can make the segmentation result more accurate.

We observe that the output of our model is more accurate than the original foreground map, and it can be considered as a new foreground map. Compared with the initial foreground map, the new foreground map has more global information

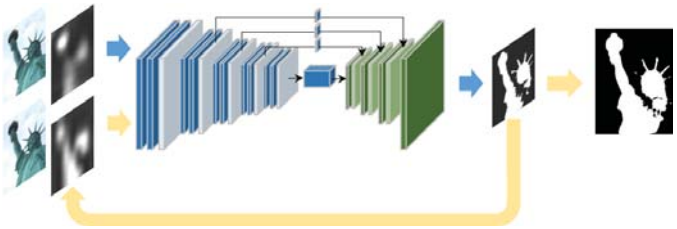

Fig. 2. The iterative segmentation scheme. The first segmentation result based on the initial foreground map is fractured but captures more global information. And we conduct a second segmentation iteration using the new foreground map transformed from the first segmentation result.

and less noise information due to the deep analysis. By using the generated foreground map and the original image for segmentation again, the boundary of the result will be more accurate. Hence, we replace the original foreground map by the new output, and implement our model to obtain new results. Such processes are iteratively implemented until the end. Our iterative segmentation method is shown in Fig 2.

2.3. Construction Of Training Data

We use the Pascal VOC 2012 semantic Segmentation dataset to train the model. When training the model, we randomly select the objects with one of the categories in the groundtruth and treat all the others as background. Then we set the mask produced by this process as the groundtruth corresponding to the image in this training iteration. At the same time, the input data is the original image and the foreground map transformed from the generated groundtruth. In other words, for the same picture, there may be different pairs of input foreground map and corresponding groundtruth label in training process.

In practice, the foreground map generation algorithms are not always able to grasp the full edges of objects, but often point out only some parts of the object, which requires our model to learn the characteristics of the object and to segment out the region with the same characteristics. We randomly remove some foreground area in the groundtruth in order to increase the diversity of the foreground maps used in training process. The advantage of this trick is that the network can only get the localization information of the object from the training foreground map. And driven by the segmentation loss function, the network is forced to learn the edges of the objects.

2.4. Implementation Details

Our model requires foreground maps as a part of input in training and test phases. In practice, we obtain the image foreground map by several type of methods such as saliency detection methods MB [2] and ST [1], and discriminative region generation method CAM [5]. Minimum Barrier (MB) algorithm is an unsupervised algorithm that uses minimum barrier distance (MBD) transform to achieve object salience detection. Salience Tree (ST) is also an unsupervised framework for object salience evaluation. The Class Activation Map (CAM) algorithm is a weakly supervised algorithm, which utilizes image-level labels to train a convolution neural network that includes the global average pooling layer to accomplish image classification and output the discriminative region of each image.

In practical applications, the global average pooling layer often makes the CAM algorithm overestimate the boundary of objects, which seriously affects the segmentation performance. To this end, we propose an improved method (we term VGG-D map) to extract discriminative region, which modestly estimates the object. We use the original VGG16 network structure to achieve image classification, and then transform the parameters of the classification layers into an extractor to extract discriminative region from the feature map which is feed into the classification layer. Since there are three fully connected layers in the VGG16 which contain huge amount of parameters, the extracted discriminative region tends to be appropriate or lower estimated. The following is a brief description of the proposed approach for extracting discriminative map from the VGG-16 net.

Assume the convolution filter parameters of $fc6$, $fc7$ and $fc8$ layers in VGG-16 network are: $fc6f(7 \times 7 \times 512 \times 4096)$, $fc7f(4096 \times 4096)$ and $fc8f(4096 \times c$, where c is the number of categories output by the $fc8$layer). Our first step in building the extractor is formulated as:

$$E^j = \sum_{i=1}^{4096} fc6f^i \cdot \sum_{k=1}^{4096} fc8f_k^j \cdot fc7f_i^k$$

The convolution filter parameter F^n represents the nth filter in F, and the F_m^n represents the mth parameter in F^n. The E^j represents the discriminant region extractor for the jth class. The operation above uses the parameters of the $fc7$ and $fc8$ layers to convert the parameters of the fully-connected layer $fc6$ into the discriminant region extractor. To obtain the discriminative map, we choose a certain class of extractor to compute the Hadamard product with the input feature map of

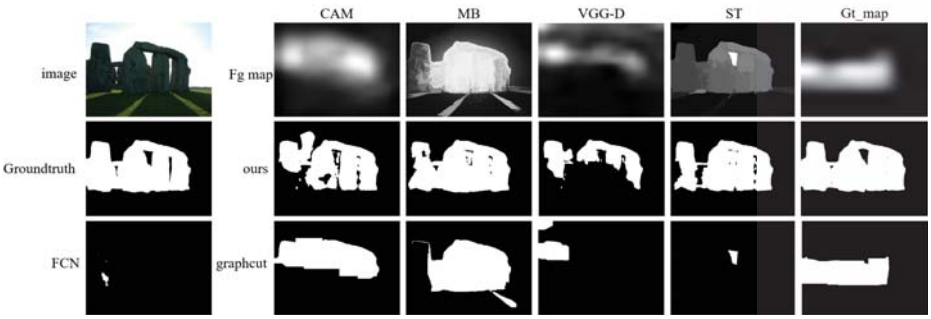

Fig. 3. The example of the segmentation results. This image is the stonehenge in IcoSeg dataset. Our model can segment out the stones as the foreground map indicated, while the FCN[17] model can not segment out anything in this image. Note that the FCN model and our model are trained on the same dataset. Compared with graphcut our model is more powerful in inferring the information of the foreground object

	CAM map[5]	MB map[2]	ST map[1]	VGG-D map
Graphcut[9]	57.21	68.01	68.19	54.22
Graphcut-SF[21]	57.96	67.92	67.52	53.17
Graphcut-HED[22]	58.91	68.52	68.33	54.19
Graphcut-SF-UCM[23]	58.77	68.29	68.11	54.01
Graphcut-HED-UCM	58.81	68.34	68.17	54.01
DFseg-class	74.29	76.91	79.45	76.64
DFseg-class-iter	77.09	79.91	79.24	78.36
DFseg-instance	74.00	76.88	78.50	76.04
DFseg-instance-iter	77.20	78.00	78.80	78.06
DFseg-skip	74.12	79.00	80.10	78.21
DFseg-skip-iter	77.96	80.70	80.20	79.04

Table 1. the comparison of the Segmentation results(mIoU in %). We term our proposed model as DFseg (Deep Foreground segmentation). In this table, DFseg-skip is the model with one more skip layer which is depicted in Fig1 and is trained by the category-level segmentation label. "iter" means the result of the proposed iteratively segmentation scheme. SF[21], HED[22] and UCM[23] are edge detection algorithms we used to produce the second order term in graphcut[9] to improve its segmentation performance.

the $fc6$ layer and compute the sum of the result matric according to its third dimension. In order to use larger scale feature map to obtain the finer discriminant map, we upsample all the extractors and then extract the discriminant map based on the feature map before the pooling layer.

3. EXPERIMENT

We evaluate our approach on the ICoseg dataset. The I-Coseg dataset contains 645 images along with pixel-level binary groundtruth. There are many classes of objects in IcoSeg dataset that do not exist in our model's training dataset Pascal VOC 2012, such as "pyramids", "pandas", "kites", etc. We test our model based on the foreground maps generated by the algorithms above, using the mIoU of the segmentation result to verify the performance of our model. The examples of the segmentation results are shown in Fig3.

For the proposed iterative segmentation scheme, we empirically set different iteration number for different foreground map generation method and test it based on the DFseg and DFseg-skip models, which is shown in Table1 where the segmentation mIoU of the models can be improved at most 3.84%. One example of the iterative segmentation scheme is shown in Fig4. And for skip structure in the network, it can be demonstrated that the model with more skip structure is better than the basis model. For example, for MB maps the mIoU of the basis model DFseg is 76.91%, while it is 79.00% for the model DFseg-skip. We discuss the comparison of our method with the baseline methods: graphcut and FCN in the following two subsections.

3.1. Comparision With Graphcut

We compare our model with graphcut, which is the most widely used segmentation method for image foreground map. We use the foreground maps as the first order term in the

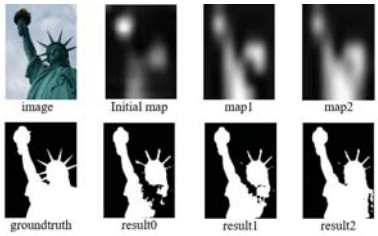

Fig. 4. The example results of the iterative segmentation scheme. The initial foreground map only points out parts of the foreground object. Through the iterative segmentation method, the model gradually learns the information of the foreground object from the image, and finally segment it out successfully.

graphcut, and then we test the following scenarios:

- Using the partial derivative of the horizontal and vertical directions of the image as the second order term;
- Using The boundary graph generated by HED[22] algorithm as the second order term;
- Using The boundary graph generated by SF[21] algorithm as the second order term;
- Using the image's UCM map[23] to enhance the above two boundaries;

We select the optimal parameters for graphcut in experiments. And the performance comparison of Graphcut and our model in Tabel1 has demonstrated that the proposed model is more practical. For foreground maps generated by CAM, MB, ST and VGG-D algorithm, the graphcut-HED algorithm is the best performed in all the graphcut-based approaches, reaching 58.91%, 68.52%, 68.33% and 54.19% respectively. And our model DFseg-skip improves the segmentation performance by at least 10%, reaching 74.12%, 79.00%, 80.10% and 78.21% respectively.

method	mIoU(%)
FCN	74.44
DFseg+CAM map[5]	74.29
DFseg+MB map[2]	76.91
DFseg+ST map[1]	79.45
DFseg+VGG-D map	76.64

Table 2. Performance on IcoSeg dataset. We train the base model of the FCN-8s[17] structure on Pascal VOC 2012 data set. And for our proposed model, we choose the same skip structure compared with the FCN-8s structure and also trained on the same data set. We test the models on the IcoSeg data set for foreground-background segmentation. As is shown in the table, our model is more general due to the ability to segment "new objects".

3.2. Comparison With FCN

We verify our method on the ICoseg dataset, which is different from the training dataset. The original FCN model trained on the same dataset is also considered. The FCN model is designed for the task of semantic segmentation, which is to assign class labels to every pixel in the image. We test the FCN model by ignoring its output class label: all the object label is treated as foreground label. Thus the FCN model may successfully segment out those classes of objects even though its result label is wrong. The skip layer structure of the both model is the same in order to compare their cross-dataset performance in the same condition.

The results of our method is listed in Table2 where the segmentation mIoU of FCN model is 74.44%. It is seen that our method achieves 74.29%, 76.91%, 79.45% and 76.64% under CAM, MB, ST and VGG-D maps, respectively. Our best result is higher than FCN model by 5% on mIoU, which demonstrates the effectiveness of our approach on segmentation of the cross-dataset.

4. CONCLUSION

In this paper, we propose a model for solving the task of foreground-background segmentation based on the image foreground prior. The proposed model can segment out objects indicated by the foreground map, even though they are completely different from those in the training dataset. Experiments demonstrate the superiority of the proposed model to the classical graphcut methods under several different foreground map generation algorithms. And at the same time it outperforms the basic FCN model on cross-dataset performance. In the future work, we will further study the use of our model in the field of multi-instance segmentation.

5. ACKNOWLEDGMENT

This work was supported in part by National Natural Science Foundation of China (No. 61202084, 61601102).

6. REFERENCES

[1] Z. Liu, W. Zou, and O. Le Meur, "Saliency tree: A novel saliency detection framework," *IEEE Transactions on Image Processing*, vol. 23, no. 5, pp. 1937–1952, May 2014.

[2] Jianming Zhang, Stan Sclaroff, Zhe Lin, Xiaohui Shen, Brian Price, and Radomir Mech, "Minimum barrier salient object detection at 80 fps," in *The IEEE International Conference on Computer Vision (ICCV)*, December 2015.

[3] Liangzhi Tang, Hongliang Li, and Tiantang Chen, "Extract salient objects from natural images," in *2010 International Symposium on Intelligent Signal Processing and Communication Systems*, Dec 2010, pp. 1–4.

[4] Fanman Meng, Jianfei Cai, and Hongliang Li, "Cosegmentation of multiple image groups," *Computer Vision and Image Understanding*, vol. 146, no. Supplement C, pp. 67 – 76, 2016.

[5] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba, "Learning deep features for discriminative localization," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[6] Fanman Meng, Jianfei Cai, and Hongliang Li, "Cosegmentation of multiple image groups," *Computer Vision and Image Understanding*, vol. 146, pp. 67–76, May 2016.

[7] Hongyuan Zhu, Jean-Baptiste Weibel, and Shijian Lu, "Discriminative multi-modal feature fusion for rgbd indoor scene recognition," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[8] Hongyuan Zhu, Romain Vial, and Shijian Lu, "Tornado: A spatio-temporal convolutional regression network for video action proposal," in *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[9] Y. Boykov, O. Veksler, and R. Zabih, "Fast approximate energy minimization via graph cuts," Nov 2001.

[10] F. Meng, H. Li, S. Zhu, B. Luo, C. Huang, B. Zeng, and M. Gabbouj, "Constrained directed graph clustering and segmentation propagation for multiple foregrounds cosegmentation," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 25, no. 11, pp. 1735–1748, 2015.

[11] Yunchao Wei, Jiashi Feng, Xiaodan Liang, Ming-Ming Cheng, Yao Zhao, and Shuicheng Yan, "Object region mining with adversarial erasing: A simple classification to semantic segmentation approach," *CoRR*, vol. abs/1703.08448, 2017.

[12] Seong Joon Oh, Rodrigo Benenson, Anna Khoreva, Zeynep Akata, Mario Fritz, and Bernt Schiele, "Exploiting saliency for object segmentation from image level labels," *CoRR*, vol. abs/1701.08261, 2017.

[13] Mohammadreza Mostajabi, Payman Yadollahpour, and Gregory Shakhnarovich, "Feedforward semantic segmentation with zoom-out features," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.

[14] Yunchao Wei, Xiaodan Liang, Yunpeng Chen, Xiaohui Shen, Ming-Ming Cheng, Yao Zhao, and Shuicheng Yan, "Stc: A simple to complex framework for weakly-supervised semantic segmentation," *IEEE TPAMI*, 2016.

[15] Xiaojuan Qi, Zhengzhe Liu, Jianping Shi, Hengshuang Zhao, and Jiaya Jia, "Augmented feedback in semantic segmentation under image level supervision," in *European Conference On Computer Vision*, 2016.

[16] Hongyuan Zhu, Fanman Meng, Jianfei Cai, and Shijian Lu, "Beyond pixels: A comprehensive survey from bottom-up to semantic image segmentation and cosegmentation," *Journal of Visual Communication and Image Representation*, vol. 34, no. Supplement C, pp. 12 – 27, 2016.

[17] Jonathan Long, Evan Shelhamer, and Trevor Darrell, "Fully convolutional networks for semantic segmentation," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.

[18] Pedro O Pinheiro, Ronan Collobert, and Piotr Dollar, "Learning to segment object candidates," in *Advances in Neural Information Processing Systems 28*, C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, Eds., pp. 1990–1998. Curran Associates, Inc., 2015.

[19] Pedro O. Pinheiro, Tsung-Yi Lin, Ronan Collobert, and Piotr Dollár, *Learning to Refine Object Segments*, pp. 75–91, Springer International Publishing, Cham, 2016.

[20] Hexiang Hu, Shiyi Lan, Yuning Jiang, Zhimin Cao, and Fei Sha, "Fastmask: Segment object multi-scale candidates in one shot," *CoRR*, vol. abs/1612.08843, 2016.

[21] Piotr Dollár and C. Lawrence Zitnick, "Fast edge detection using structured forests," *ArXiv*, 2014.

[22] Saining Xie and Zhuowen Tu, "Holistically-nested edge detection," in *The IEEE International Conference on Computer Vision (ICCV)*, December 2015.

[23] P. Arbelaez, M. Maire, C. Fowlkes, and J. Malik, "Contour detection and hierarchical image segmentation," May 2011.