Improving Semi-automatic Segmentation with U-Net

Pengze Liu
Department of Computer Science
City University of Hong Kong
pengzeliu2-c@my.cityu.edu.hk

Shuhan Li
Department of Electronic Engineering
City University of Hong Kong
shuhanli3-c@my.cityu.edu.hk

Abstract

In this project, we replicate two traditional computer vision algorithms, i.e., GrabCut and Binary Segmentation for semi-automatic segmentation. We also introduce a Deep Learning method to demonstrate the effectiveness of Convolutional Neural Networks to extract semantic information from visual information. We evaluate the results on the dataset previously published by Oxford VGG. Intersection over Union (IoU) is selected as the judging metric for segmentation. The experiment suggests that U-Net, the selected CNN architecture surpassed both traditional algorithms in this benchmark. The U-Net is robust to all categories of objects in the dataset, indicating its good generalization ability. The code, models, and results are available at https://github.com/lpzjerry/GrabCut-FCN.

1. Introduction

Segmentation or pixel-wise classification is a fundamental and critical problem in Computer Vision. With the rapid development of Deep Learning in recent years, many Neural Networks [9, 4, 8, 7] are proposed to address the problem of Semantic Segmentation. Due to the increasing availability of annotated data [2, 6, 1], the performances of these Deep Learning methods are reasonably good in many scenarios. However, it is costly and time-consuming to manually annotate a large amount of data by pixel, semi-automatic segmentation that provides pixel-wise labels with limited human input become a demanding task. To address this problem, many previous works [11] use the graph-based model for clustering pixels, such that with raw annotations in lines or bounding boxes, the models are capable of performing binary segmentation for a single object in a small region.

The task of Semi-automatic Segmentation can be formulated into a pixel-wise binary segmentation problem. It is a simplified version of multi-label semantic segmentation since we only focus on the foreground without considering the category of the objects. Inspired by the recent development of Deep Learning, we train a U-Net [10] to perform binary segmentation on multiple categories of objects, expecting that the network is capable of generalizing to other unseen objects. As a comparison, we also replicated two traditional methods for Semi-automatic Segmentation, known as GrabCut and Binary Segmentation. We evaluate the effectiveness of all methods on Geodesic Star Convexity Dataset [3] and compare the performance of the methods by calculating the IoU of the output masks. The result shows that U-Net [10] is the best model in this dataset with good generalization ability. We also visualize the segmentation results to perform a qualitative evaluation to obtain a better understanding of the performances.

2. Methodology

2.1. GrabCut

Grabcut [11] was first presented in 2004 by researchers Carsten Rother, Vladimir Kolmogorov and Andrew Blake in their paper, "GrabCut": interactive foreground extraction using iterated graph cuts. This algorithm is used to segment foreground and background, and it would provide better performance when there is a bounding box of the foreground or an annotation mask with foreground and background information.

The principle of the theory is based on Gaussian Mixture Model (GMM) which can model the foreground and background. As mentioned before, if we give the bounding box of the foreground or the annotation information, GMM will learn and create new pixel distribution according to them. The unknown pixels are labeled either probable foreground or probable background depending on its relation with the other hard-labeled pixels in terms of color statistics. Then, a graph is built based on pixel distribution. Source node and sink node are added to the graph. Connect all possible foreground to the source node and all possible background to the sink node. The weights of edges connecting pixels to the source node/end node are defined by the probability of a pixel being foreground/background. The weights between the pixels are defined by the edge information or pixel sim-

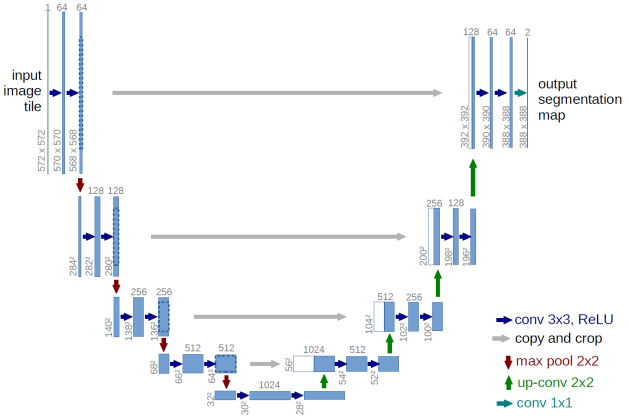

Figure 1. U-Net Architecture.

ilarity. Then a min-cut algorithm is used to segment the graph. It cuts the graph into two separating source node and sinks node with a minimum cost function. The cost function is the sum of all weights of the edges that are cut. After the cut, all the pixels connected to the source node become foreground and those connected to the sink node become background. Do the process iteratively until the classification converges.

2.2. Binary Segmentation

Given sparse markings of foreground and background by the user, it calculates SLIC superpixels, and runs a graph-cut algorithm. Color histograms are calculated for all superpixels and foreground-background. This algorithm takes into account superpixel-superpixel and superpixel-Foreground/Background interaction to obtain a final binary image segmentation.

2.3. U-Net and CNN-based Semantic Segmentation

Convolutional Neural Networks [5], typically Fully Convolutional Networks (FCNs) [12, 9] achieved remarkable result in Semantic Segmentation. These FCN networks typically comprise a downsampling network with a stack of convolution layers and an upsampling network with a stack of transposed convolution layers. The downsampling and upsampling layers are symmetric such that the final output is of the same spatial size as the input image. Pixel-wise classification is performed as the gradient is calculated as the classification loss of each pixel. The network is optimized end-to-end, and the weights in the network are updated by the gradient backpropagation.

In this project, we choose U-Net as our network architecture, as shown in Figure 1. The U-Net is an optimized version of FCN. It is symmetric, and a skip connection between the downsampling and upsampling branch is introduced to concatenate local information with global information in the upsampling stage. As the Semi-automatic Segmentation task can be formulated as a binary segmentation task, we may use a Binary Sigmoid Cross Entropy (BCE) loss (for ultimate pixel accuracy) or dice coefficient (for best IoU) to optimize the network. In the test phase, all pixels in the output mask over Sigmoid with a value of greater than 0.5 is mapped to positive and vice versa. The network is trained using Stochastic Gradient Descent (SGD) algorithm with a learning rate of 0.1 and a weight decay of 0.0005. The U-Net is trained by five epochs over the dataset, and we select the best model on the validation set as the model in the test phase.

3. Experiment

To evaluate our result, we use the dataset introduced in Geodesic Star Convexity for Interactive Image Segmentation [3]. This dataset is a combination of *PASCAL VOC 2007* dataset [2], *GrabCut* dataset [11] and *Alpha matting* dataset, consisting of 151 images with both binary mask annotations and raw foreground/background annotations. To provide extra supervision, we further manually annotate the bounding boxes for each object of interests in this dataset. Since the U-Net requires training, we split the dataset into a training and validation set with 121 images and use the rest 30 images as the test set that is invisible to the network except the evaluation stage.

For the evaluation metric, we use Intersection over Union (IoU):

$$IoU = \frac{target \cap prediction}{target \cup prediction}$$

We test all three methods using different annotations as input. For GrabCut and Binary Segmentation, we tested the methods with or without manual annotations. For the U-Net, we trained both on the original images and the cropped images from our annotated bounding boxes. Due to the limited time and computational resources, we did not fully explore the hyperparameters for U-Net.

4. Result Analysis

4.1. Quantitative Evaluation

The result of the experiments is shown in Table 1. Result suggests that U-Net with bounding box is significantly better than the two traditional methods. For both GrabCut and U-Net, the bounding box is critical to the result. It could be explained that semantic information is hard to be directly explored since the superpixels are identical to each other such that it is hard for the models to predict the foreground without any prior knowledge. The models without parameters are not as good as the model with parameters, while GrabCut and Binary Segmentation is more sufficient since the training stage is omitted. The U-Net also shows

Method	Input	IoU
GrabCut	Image	0.1950
	Image+Box	0.4096
	Image+Box+Annotations	0.4217
Binary Segmentation	Image+Annotations	0.4408
U-Net	Image+Box	**0.5527**

Table 1. Result of the replicated methods. U-Net with bounding box achieves the best result in this dataset.

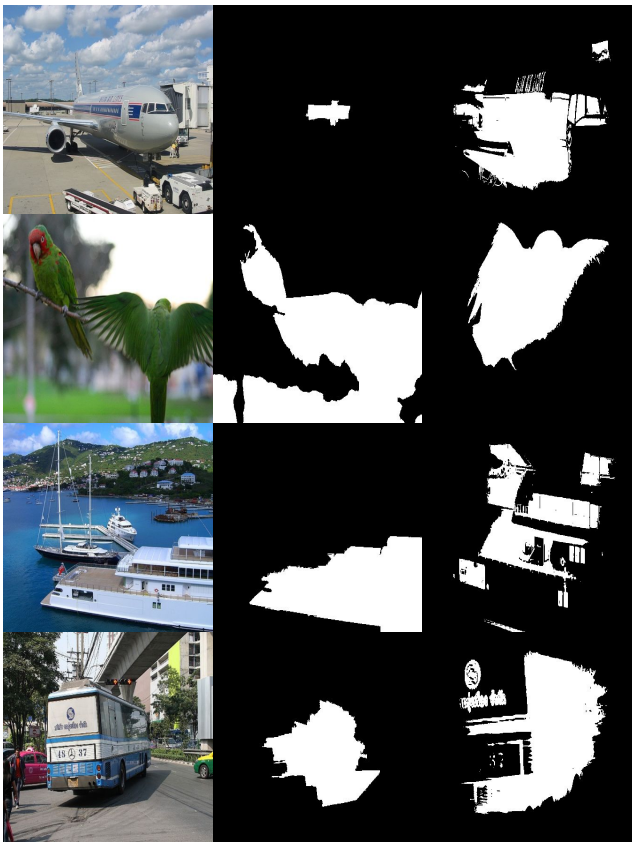

Figure 2. Qualitative Evaluation. From left to right: original image, result by GrabCut, result by U-Net.

good generalization ability, as the dataset contains multiple classes of object. It suggests that deep learning is capable of exploring semantic information from the extracted features given a sufficient amount of training data.

4.2. Qualitative Evaluation

Figure 2 are images randomly selected from the test set. Intuitively, the result of U-Net is better than the GrabCut results, especially for the bus image. It suggests that U-Net has good generalization ability and can perform semi-automatic segmentation relatively well over many categories. While the U-Net tends to output masks with coarse boarders, as the predictions are made by each pixel instead of a cluster of pixels together as GrabCut or Binary Segmentation, this problem could probably be addressed by future experiments to ensemble these methods together.

5. Conclusion

In this project, we replicated three major Semi-automatic Segmentation Algorithms on the VGG dataset, and we evaluated these methods both from the quantitative aspect and qualitative aspect. Our result proves that U-Net or deep learning models has good generalization ability to extract semantic information from features and are robust to the geometrical shape or size of objects. Our experiment suggests that using low-level vision algorithms to avoid coarse boarders could probably improve the performance of FCNs. Future experiments are expected to evaluate our hypothesis.

References

[1] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3213–3223, 2016.

[2] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88(2):303–338, 2010.

[3] V. Gulshan, C. Rother, A. Criminisi, A. Blake, and A. Zisserman. Geodesic star convexity for interactive image segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2010.

[4] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017.

[5] Yann LeCun, Bernhard Boser, John S Denker, Donnie Henderson, Richard E Howard, Wayne Hubbard, and Lawrence D Jackel. Backpropagation applied to handwritten zip code recognition. *Neural computation*, 1(4):541–551, 1989.

[6] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.

[7] Chenxi Liu, Liang-Chieh Chen, Florian Schroff, Hartwig Adam, Wei Hua, Alan Yuille, and Li Fei-Fei. Auto-deeplab: Hierarchical neural architecture search for semantic image segmentation. In *CVPR*, 2019.

[8] Shu Liu, Lu Qi, Haifang Qin, Jianping Shi, and Jiaya Jia. Path aggregation network for instance segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8759–8768, 2018.

[9] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015.

[10] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

[11] Carsten Rother, Vladimir Kolmogorov, and Andrew Blake. Grabcut: Interactive foreground extraction using iterated graph cuts. In *ACM transactions on graphics (TOG)*, volume 23, pages 309–314. ACM, 2004.

[12] Pierre Sermanet, David Eigen, Xiang Zhang, Michaël Mathieu, Rob Fergus, and Yann LeCun. Overfeat: Integrated recognition, localization and detection using convolutional networks. *arXiv preprint arXiv:1312.6229*, 2013.