One-Shot Image Segmentation with U-Net

Guanyi Zhao1, a, †, He Zhao2, b, †
1Department of Computer Science, George Washington University, Washington DC20052, US
2Faculty of Engineering, University of New South Wales, UNSW Sydney High St Kensington, NSW2052, Australia
agzhao25@gwu.edu, bz5108047@ad.unsw.edu.au
† These authors contributed equally.

Abstract. Image segmentation is an important task in the field of computer vision research. Many applications need accurate and efficient segmentation mechanisms. However, most existing methods need a large support set and have struggled with dealing with new classes in solving computer vision problems. In order to deal with the problem, this paper modified the traditional model to adapt it to the task of few-shot segmentation. Especially, the part of encoders was changed into a Siamese neural network for support branch and query branch. Then, the decoded query image is compared with different levels of features from the encoded support image. Experimental results show that this proposed model achieved the best performance compared with several baseline methods, and the improvement is more than 8%.

1. Introduction
Deep Neural Network has shown its powerful ability in the area of image segmentation. Many popular convolutional networks have been proposed for computer vision tasks, such as U-Net [1], Mask R-CNN [2], FCN [3], SegNet [4], and DeepLab [5]. However, most of these networks require a great amount of data for training, which costs significant time and effort. What’s more, these models also struggle to adapt to new classes in a data-efficient manner. In recent years, one-shot learning has been proposed to deal with the data-sparse problem and shown excellent performance in some tasks. In this paper, we take advantage of the one-shot learning and apply it to the image segmentation task, i.e., one-shot segmentation. One-Shot image segmentation aims to train a network based on a limited labeled dataset and an unknown large dataset.

In One-Seg [6], the task of One-Shot semantic Image Segmentation is proposed. In this paper, we followed the same configuration and task in One-Seg [6]. The task’s goal is to predict a segmentation mask for a semantic class given in the supported image. With only one sample of the target class, the network will be able to segment on the query images. What’s more, we leveraged the Siamese Network to make our model more efficient, for it is a simple and efficient structure that aims to build two branches. In Siamese network, the two branches can share the same parameters. These branches are made of the same network, which uses the same parameters. In actual structure, the two branches can be replaced by only one network. Inducing the Siamese Network into the encoding module makes training and optimization more efficient.

Furthermore, most convolution networks, such as VGG [7] and FCN [3], will lose much information in the convolution process. Since one-shot learning has limited data, losing information will decrease
accuracy significantly. In order to avoid so much information loss, we propose U-net as the backbone of our network. The U-net [1] is a convolutional neural network, which analyzes information from different levels to determine a mask. The U-net’s contracting path follows the typical architecture of a convolutional network. However, its expansive path consists of an up-sampling of the feature map and concatenation with the corresponding cropped feature map from the contracting path. The U-net allows the up-sampling layers to have a large number of features. The network allows propagating context information, which helps module trained by limited dataset to make higher accuracy predictions.

In this work, we proposed a model based on U-net [1] with a Siamese Network as an encoding module. The encoders and the decoders are based on a double convolution module induced from U-net. We use the Siamese Network with encoders to gain the features at different levels from the query image and support the image with its mask. The features are concatenated at a different level and compared to each other. We use Cross Entropy Loss in the model. The experiment shows that our approach has a significant improvement over the baselines.

2. Related Work

2.1 Image Segmentation

Image segmentation is the process of segmenting digital images into several segments. After semantic segmentation, the original image contains several colors, each of which represents a category.

In the computer vision field, recognizing and locating objects are main goals of many modules, and image segmentation can locate the sets of pixels for different objects. Networks used in image segmentation tasks, like FCN, U-net and Mask R-CNN, have their own advantages in different areas, and we should get the suitable one.

2.2 Semantic Segmentation

Recognizing objects is an important goal of many computer vision tasks, so people propose Semantic Segmentation to recognize the specific classes of images. Semantic Segmentation will separate the image into different categories, and objects from the same categories will share the same features. Semantic Segmentation is widely used in various applications, including autonomous driving, image search engines, and human-machine interaction.

2.3 One-shot learning

One-shot learning [8,9] aims to learn information about object categories from one or a few training samples. A large data-set is virtually impossible in some conditions, and this is where one-shot learning comes into play. In contrast, many non-parametric models allow novel examples to be rapidly assimilated, whilst not suffering from catastrophic overfitting[10]. Instead of treating the task as a classification problem, one-shot learning turns it into a different evaluation problem. The learning model takes two images and returns a value that shows the similarity between the two images when adjusted for one-shot learning [11]. There are many types of one-shot learning: metric learning methods, meta-learning [12,13] method, data-augment method, and semantic-based method. The metric learning method learns the mapping images from the same class closed together and away from the different classes. Meta-learning aims to learn new concepts and skills fast with a few training data. Data-augment will generate more useful samples from the few labeled objects. Semantic-based methods aiming to find the relationship between the objective and object properties are introduced in zero-shot learning.

2.4 One-shot Segmentation

One-shot object detection and segmentation aim to develop a model that can localize and segment objects from arbitrary categories when provided with a single visual example from that category. One-shot segmentation takes sparse labeled images as the training set and segments an image from an unseen class in the training process. Few-shot segmentation avoids over-fitting and time-consuming, so it is applied to solve many computer vision problems. Many useful modules have already been proposed.
2.5 Siamese Network

Siamese networks [14] are an effective method for contrast learning that use contrastive loss to embed semantically similar samples closer together and dissimilar images further away [8,15]. Siamese Network works in tandem, taking two images as input and output the difference value. Once the Siamese network is trained, the model will tell them these images are similar enough to be the same category or not if people provide two new images.

3. Method

In this paper, we build a two-branch network based on U-net [1]. The overview of our model is shown in Figure 1. The query branch will take the RGB image as input. In the support branch, the product of support image and support mask is taken as input. The two branches are built based on Siamese network. They share the same weights and bias. By using Siamese network in the structure, we can reduce the load and time for training. The two branches will encode the features of the support and the query into different levels. Then the features will be compared and decoded for predicting segmentation masks. The predicted mask will be used for calculating Cross Entropy Loss with the true mask.

![Figure 1. Overview of the proposed network](image)

A double convolution module is made up of two 3x3 convolution layers, shown in Figure 2. Each of the convolution layers is followed by a rectified linear unit (ReLU) and batch normalization. This module is used as the core part of the whole network. There are five double convolution modules in the support branch, query branch, and decode branch.
The structure and feature sizes of the query branch and support branch are shown in Figure 3. As we are using Siamese network, the two branches share the same structure. The input of the branch will be encoded into different levels for later comparison and decoding.

A decoder module from decode branch is shown in Figure 4. There are five decoder modules in the network. The decoder at bottom is slightly different; the other four are the same. The bottom decoder consists of two double convolution modules. It takes the highest encoded features from query branch and support branch. The features are concatenated and passed through the double convolution modules. The output will be downsampled. The other four decoders consist of only one double convolution module. However, they take three inputs: the output of the previous decoder and a specific level of features from the query branch and support branch.
Finally, the output layer, which is a convolution layer, will calculate the probability of being the object or the background. Then we use the softmax function of finalizing the class of every pixel to get the predicted mask. The predicted mask will be used for calculating the Cross Entropy Loss for the training process and evaluation.

4. Data
The data set PASCAL-5i is constructed for few-shot image segmentation tasks, which are proposed in Shaban et al.’s research [6]. The data set is constructed based on PASCAL VOC 2012 and extra masks from SDS [17]. The data is split into four folds for training and evaluation. The classes contained in the folds are shown in Table 1.

| FOLD | 0 | 1 | 2 | 3 |
|------|---|---|---|---|
| CLASSES | plane, bicycle, bird, boat, bottle | bus, car, cat, chair, cow | dining table, dog, horse, motorbike, person | potted plant, sheep, sofa, train, TV/monitor |

5. Results and Discussion
In the experiment, we set the random seed as 1337, which is popularly used to gain the same result among different tests. The images from the data set are normalized using mean [0.485, 0.456, 0.406] and std [0.229, 0.224, 0.225]. After that, the images are resized to [224, 224]. There are mainly two classes in the task: the target object and the background—the optimizer used in the experiment in Adam with a 0.003 learning rate. The IoU and the loss of every image in the evaluation part will be recorded. We decided to use Cross Entropy Loss as our loss function.

Our network is tested with the four splits of the dataset. The results are evaluated using mean Intersection-over-Union (mean IoU) and shown in Table 2 below.

| Methods | Split 0 | Split 1 | Split 2 | Split 3 | Mean IoU |
|---------|---------|---------|---------|---------|----------|
| 1-NN    | 25.3    | 44.9    | 41.7    | 18.4    | 32.6     |
| Log Reg | 26.9    | 42.9    | 37.1    | 18.4    | 31.4     |
| Finetuning | 24.9    | 38.8    | 36.5    | 30.1    | 32.6     |
| Siamese | 28.1    | 39.9    | 31.8    | 25.8    | 31.4     |
| One-Seg [6] | 33.6    | 55.3    | 40.9    | 33.5    | 40.8     |
| Ours    | 51.2    | 51.5    | 44.8    | 47.9    | 48.8     |
In three folds of the data set, our model can work much better. However, in split 1, the model is worse than the baseline. The reason can be the content in the split. Good predictions are shown below. In some categories, the model has a very good performance. The general shape of the object can be predicted. Only some small error exists in the prediction.

![Figure 5. The optimal predictions](image)

6. Conclusion
In this paper, we proposed a method of one-shot segmentation, which is based on U-net. The structure of U-Net turned out to be good in the setting of one-shot segmentation. It works better than expected. However, there are still some problems. For example, some of the masks are stripped. The pixels of the overall object are not continuous, but the general shape of the object is correct. The problem can be handled by adding additional refinement modules for the masks. For future work, we are planning to design a suitable refinement module for the network. The refinement module aims to rearrange the pixels in the predicted images for improving the performance.

Reference
[1] Ronneberger, O. et al. “U-Net: Convolutional Networks for Biomedical Image Segmentation.” MICCAI (2015).
[2] He, Kaiming et al. “Mask R-CNN.” 2017 IEEE International Conference on Computer Vision (ICCV) (2017): 2980-2988.
[3] Shelhamer, Evan et al. “Fully Convolutional Networks for Semantic Segmentation.” IEEE Transactions on Pattern Analysis and Machine Intelligence 39 (2017): 640-651.
[4] Badrinarayanan, Vijay et al. “SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation.” IEEE Transactions on Pattern Analysis and Machine Intelligence 39 (2017): 2481-2495.
[5] Chen, Liang-Chieh et al. “DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs.” IEEE Transactions on Pattern Analysis and Machine Intelligence 40 (2018): 834-848.
[6] Shaban, Amirreza et al. “One-Shot Learning for Semantic Segmentation.” ArXiv abs/1709.03410 (2017): n. pag.
[7] Simonyan, K., & Zisserman, A. (2015). Very Deep Convolutional Networks for Large-Scale Image Recognition. CoRR, abs/1409.1556.

[8] Fei-Fei, Li, Fergus, Robert, and Perona, Pietro. A bayesian approach to unsupervised one-shot learning of object categories. In Computer Vision, 2003. Proceedings. Ninth IEEE International Conference on, pp. 1134–1141. IEEE, 2003.

[9] Fei-Fei, Li, Fergus, Robert, and Perona, Pietro. One-shot learning of object categories. Pattern Analysis and Machine Intelligence, IEEE Transactions on, 28(4): 594–611, 2006.

[10] Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Koray Kavukcuoglu, Daan Wierstra al. “Matching Networks for One Shot Learning” NeurIPS (2016)

[11] Lake, Brenden M, Lee, Chia-ying, Glass, James R, and Tenenbaum, Joshua B. One-shot learning of generative speech concepts. Cognitive Science Society, 2014.

[12] Shota Horiguchi, Daiki Ikami, and Kiyoharu Aizawa. Significance of Softmax-based Features in Comparison to Distance Metric Learning-based Features. IEEE transactions on pattern analysis and machine intelligence, 2019.

[13] Junlin Hu, Jiwen Lu, and Yap-Peng Tan. Discriminative Deep Metric Learning for Face Verification in the Wild. In CVPR, pages 1875–1882, 2014

[14] BG Kumar, Gustavo Carneiro, Ian Reid, et al. Learning Local Image Descriptors with Deep Siamese and Triplet Convolutional Networks by Minimising Global Loss Functions. In CVPR, pages 5385–5394, 2016.

[15] Haomin Chen et al. “Anatomy-Aware Siamese Network: Exploiting Semantic Asymmetry for Accurate Pelvic Fracture Detection in X-ray Images” ECCV (2020)

[16] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, A. Zisserman, The PAS-CAL Visual Object Classes Challenge 2012 (VOC2012) Results, http://www.pascal-network.org/challenges/VOC/voc2012/workshop/index.html.

[17] Bharath Hariharan, Pablo Arbeláez, Ross Girshick, and Jitendra Malik. Simultaneous detection and segmentation. In European Conference on Computer Vision, pages 297–312. Springer, 2014.