# Object Segmentation

**Lavanya Govindaraju**
7002430
lago00001@teams.uni-saarland.de

## Abstract

Image segmentation is the process of teaching a neural network to generate a pixel-wise mask of an image that can be used to understand the image at a much lower level, namely the pixel level. Image segmentation is used in the pre-processing of images with the aim of dividing the image into components or regions of interest for a more thorough study of one or more of these regions. In task 1, we implement a VGG16 model using the PASCAL VOC dataset. The implementation and testing of the paper "Recurrent Residual Convolutional Neural Network based on U-Net (R2U-Net) for Medical Image Segmentation" is done in task 2. In task3, we implement Attention-UNet which is assumed to perform better than UNet but not R2-UNet.

**Keywords**

Image Segmentation; Vgg16; R2-UNet; U-Net; cityscapes; Attention-UNet.

## 1 Introduction

The Semantic segmentation, or image segmentation, is the task of clustering parts of an image belonging to the same object class together.Since each pixel in an image is labeled according to a category, it is a type of pixel-level prediction.Some of the example benchmarks for image segmentation are Cityscapes, PASCAL VOC and ADE20K.[1] Each annotated pixel in an image is assigned to a single class using image segmentation. It's commonly used to mark images for high-accuracy applications, and it's labor-intensive since pixel-level accuracy is necessary. It may take up to 30 minutes or more to complete a single image. The result is a mask that highlights the shape of the image's object. Although there are many different types of segmentation annotations (such as semantic segmentation, instance segmentation, panoptic segmentation, and so on), image segmentation generally refers to the need to annotate every pixel of an image with a class. [2]The goal of image segmentation is to recognize and comprehend what's in the image down to the pixel level. Unlike object detection, where the bounding boxes of objects can overlap, every pixel in an image belongs to a single class. When a computer vision application needs to be developed with high precision, image segmentation is common for real-world ML models. Autonomous vehicles, medical imaging, retail applications, and other applications use image segmentation.

## 2 Methodology

**Task1:** The main aim is to implement the semantic segmentation over the PASCAL VOC dataset. The Pascal VOC dataset recognizes objects from a number of visual object classes in realistic scenes (i.e. not pre-segmented objects) where the training set of labelled images are provided. For this work, we use the Vgg16 pre-trained model, as Vgg16 acts as a classifier so for changing it to do segmentation we added an extra layer to the model to produce the

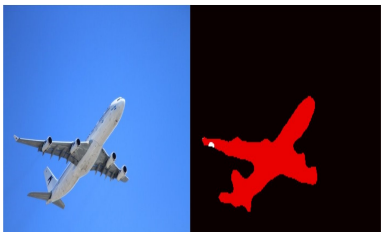

Figure 1: Semantic Segmentation (aws.amazon.com)

output of the same size as input. Long et al. [3] proposed the full convolution network (FCN), which replaces the full connection layer in the classification network framework with the convolution layer, and outputs spatial mapping instead of classification score, achieved good results, and became the cornerstone for the application of deep learning technology in semantic segmentation. The use of spatial mapping instead of classification scores produced good results, and it became the foundation for using deep learning technology in semantic segmentation. [4] The most distinctive feature of VGG16 is that, rather than having a large number of hyper-parameters; it is concentrated on having 3x3 filter convolution layers with a stride 1 and still used the same padding and maxpool layer of 2x2 filter of stride 2. Throughout the design, the convolution and max pool layers are arranged in the same way. It has two fully connected layers at the top, followed by a softmax for output. The 16 in VGG16 refers to the fact that it has 16 layers of different weights. This network is very wide, with approximately 138 million parameters.

Advantages

1. It is a very good architecture for benchmarking on a particular task.
2. Also, pre-trained networks for VGG are available freely on the internet, so it is commonly used out of the box for various applications

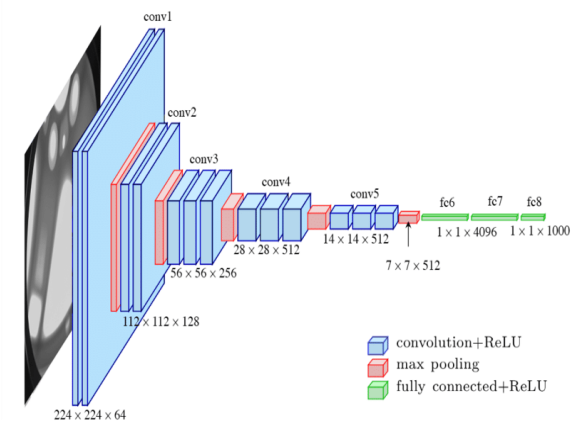


Figure 2: Vgg16 Architecture [5]

**Task2:**
One of the major advantages of R2-UNet is that it adds a residual error and recurrent network on the basis of U-Net. It is used for medical image segmentation. The segmentation in R2-UNet has better effect than U-Net. We are implementing R2-UNet based on "Recurrent Residual Convolutional Neural Network based on U-Net (R2U-Net) for Medical Image Segmentation" paper. The proposed models are tested on three benchmark datasets such as blood vessel segmentation in retina images, skin cancer segmentation, and lung lesion segmentation. In our work, we have implemented R2-UNet architecture on Cityscapes dataset. The Cityscapes Dataset focuses on semantic understanding of urban street scenes that has

30 classes comprising of different diversities such as 50 cities, several months, daytime, Good/medium weather conditions, Manually selected frames (large number of dynamic objects, Varying scene layout, Varying background). **U-Net**: UNET was developed by Olaf Ronneberger et al.[6] for Bio Medical Image Segmentation. The U-Net architecture contains two paths. First path is the encoder, a traditional stack of convolutional and max pooling layers that is used to capture the context in the image. The second path is the decoder which is used to enable precise localization using transposed convolutions. Thus it is an end-to-end fully convolutional network (FCN), i.e. it only contains Convolutional layers and does not contain any Dense layer because of which it can accept image of any size.
**Recurrent residual convolutional Neural network (R2U-Net)** [7] is chosen as the generative network where the network architecture of R2U-Net is like that of a basic U-Net. In this model, the features are accumulated with different time-step, which makes the feature representation better and makes sense of low-level feature extraction replacing all regular forward convolutional layers by recurrent residual blocks makes the R2U-Net model deeper than original U-net.

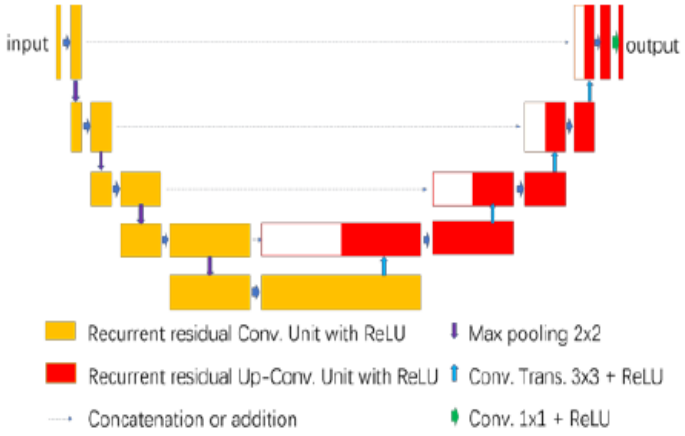


Figure 3: R2U-Net Architecture

**Task 3:**
[8] Attention, in the image segmentation, is a way to highlight only the relevant activations to certain parts of the image during training that results in the reduction of the computational resources wasted on irrelevant activations, providing the network with better generalisation power.
The attention gate takes in two inputs where one of the vector input is taken from the next lowest layer of the network. The vector x have dimensions of 64x64x64 and vector g would be 32x32x32. Vector x slides through a strided convolution such that it's dimensions become 64x32x32 and vector g goes through a 1x1 convolution such that it's dimensions become 64x32x32. The two vectors are summed element-wise which results in aligned weights becoming larger while unaligned weights become relatively smaller. The resultant vector goes through a ReLU activation layer and a 1x1 convolution that collapses the dimensions to 1x32x32 which then goes through a sigmoid layer between the range [0,1], producing the attention coefficients (weights), where coefficients closer to 1 indicate more relevant features. The attention coefficients are multiplied element-wise to the original x vector, scaling the vector according to relevance that is then passed along in the skip connection as normal.
Results obtained by Oktay et al. show that the Attention U-Net has outperformed a plain U-Net in the overall Dice Coefficient Score by a sizeable margin. While the Attention U-Net has more parameters, it is not significantly more and the inference time is only marginally longer. In conclusion, attention gates are a simple way to improve the U-Net consistently in a large variety of datasets without a significant overhead in terms of computational cost.

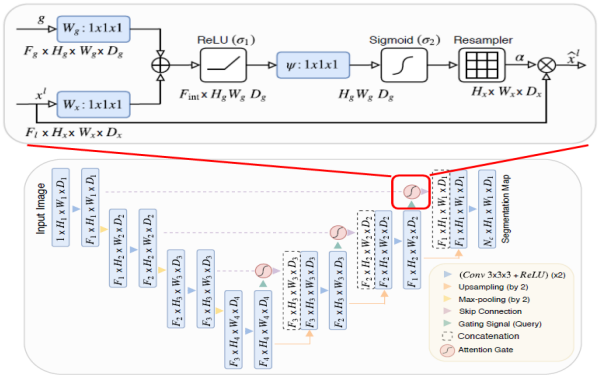


Figure 4: Attention-UNet Architecture

## 3 Experimental Evaluation and Results

For task1, we evaluate the model on 2 metrics namely, F1-score and dice-coefficient.

| Metrics | Values |
|---|---|
| F1 score | 0.8560 |
| Dice Co-efficient | 0.7482 |

Table 1: Performance metrics of vgg16(task1)

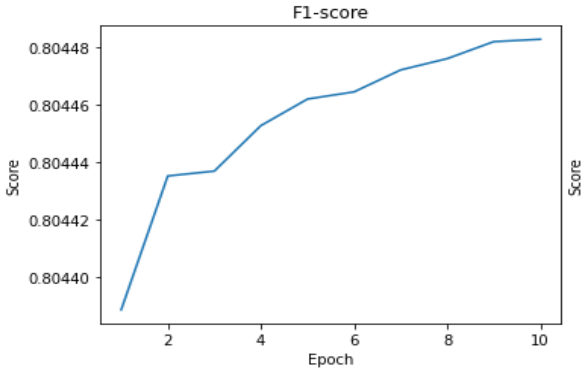


Figure 5: F1-Score plot in vgg16 (task1)

For task2, we evaluate the model on the 5 metrics namely, accuracy, specificity, sensitivity, F1-score, and jaccard score.
For the calculation of metrics, we require attributes like true positive(TP), true negative (TN), false positive (FP) and false negative(FN) respectively.

Accuracy(AC): Accuracy is one metric for evaluating classification models which is calculated as the fraction of predictions our model got right.

$$AC = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

Sensitivity(SE): Sensitivity is the metric that evaluates a model's ability to predict true positives of each available category.

$$SE = \frac{TP}{TP + FN} \tag{2}$$

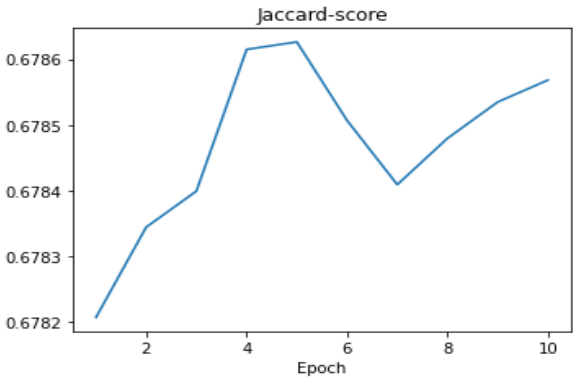


Figure 6: Jaccard score plot in vgg16

F1-Score: F1-score is a measure of a model's accuracy on a dataset.

$$DC = 2\frac{|GT \cap SR|}{|GT| + |SR|} \tag{3}$$

Jaccard Score(JS): Jaccard score is a metric that is used for gauging the similarity and diversity of sample sets.

$$JS = \frac{|GT \cap SR|}{|GT \cup SR|} \tag{4}$$

| Metrics | Values |
|---|---|
| Accuracy | 0.9840 |
| Specificity | 0.9876 |
| F1-Score | 0.8821 |
| Jaccard Score | 0.7570 |

Table 2: Performance metrics of R2-UNet

The evaluation metrics of Attention-UNet are as below:

| Metrics | Values |
|---|---|
| Accuracy | 0.9130 |
| Specificity | 0.9086 |
| F1-Score | 0.8489 |
| Jaccard Score | 0.7340 |

Table 3: Performance metrics of Attention-UNet

## 4 Conclusion

Working on this project we have successfully completed the tasks given to us. Task 1 results can be improved by running it for more epochs. In task 2, the R2-UNet provides expected results. For task 3, we tried to improve the results for task 2 by using Attention-UNet but it doesnot perform as good as R2-UNet.

## References


[1] "A 2021 Guide to Semantic Segmentation." AI  Machine Learning Blog, 7 Mar. 2021, https://nanonets.com/blog/semantic-image-segmentation-2020/.

[2] Liu, Li, et al. "Deep Learning for Generic Object Detection: A Survey." International Journal of Computer Vision, vol. 128, no. 2, Feb. 2020, pp. 261–318. Springer Link, doi:10.1007/s11263-019-01247-4

[3] Long J, Shelhamer E, Darrell T. Fully Convolutional Networks for Semantic Segmentation. IEEE Transactions on Pattern Analysis  Machine Intelligence. 2014; 39(4):640–651

[4] Brownlee, Jason. "A Gentle Introduction to Pooling Layers for Convolutional Neural Networks." Machine Learning Mastery, 21 Apr. 2019, https://machinelearningmastery.com/pooling-layers-for-convolutional-neural-networks/.

[5] Automatic localization of casting defects with convolutional neural networks – Scientific Figure on ResearchGate. Available from: https://www.researchgate.net/figure/Fig-A1-The-standard-VGG-16-network-architecture-as-proposed-in-32-Note-that-only$_fig3_322512435$.

[6] Ronneberger, Olaf, et al. "U-Net: Convolutional Networks for Biomedical Image Segmentation." Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015, edited by Nassir Navab et al., vol. 9351, Springer International Publishing, 2015, pp. 234–41. DOI.org (Crossref), doi:10.1007/978-3-319-24574-4$_28$

[7] Le, Kening, et al. "Auto Whole Heart Segmentation from CT Images Using an Improved Unet-GAN." Journal of Physics: Conference Series, vol. 1769, no. 1, Jan. 2021, p. 012016. DOI.org (Crossref), doi:10.1088/1742-6596/1769/1/012016.

[8] Oktay, Ozan, et al. "Attention U-Net:  Learning Where to Look for the Pancreas." ArXiv:1804.03999 [Cs], May 2018. arXiv.org, http://arxiv.org/abs/1804.03999.