OBJECT SEGMENTATION

Anush Onkarappa
Saarland University
7010620
anon00001@stud.uni-saarland.de

Hitesh Kotte
Saarland University
7010571
hiko00001@stud.uni-saarland.de

March 31, 2021

Abstract

In the history of Computer Vision one of the most challenging task is the object segmentation. Object segmentation has now evolved and many models are arising for the process of object segmentation. It is more important than the object recognition nowadays as a human being can recognize without even knowing what the is the object (for instance, in satellite imagery or medical X-ray scans, there may be several objects which are unknown, but they can still be segmented within the image typically for further investigation. This has helped science grow and now the improvising task is to find out which model when used to particular data gives the best accuracy. Herein this work, we have applied variety of models to depict the image segmentation for two types of data i.e, the VOC dataset and cityscapes dataset.

1 Introduction

Image Segmentation is an essential part of many visual comprehension systems. It entails breaking down images (or video frames) into separate segments or objects. Medical image processing (e.g., tumor boundary extraction and tissue volume measurement), autonomous vehicles (e.g., navigable surface and pedestrian detection), video monitoring, and virtual reality are only a few of the technologies that use segmentation. entire image. From the earliest methods, such as thresholding, histogram-based bundling, region-growing, k-means clustering, and watersheds, to more sophisticated algorithms, such as active contours, graph cuts, conditional and Markov random fields, and sparsity-based methods, numerous image segmentation algorithms have been established in the literature. Deep learning (DL) networks, on the other hand, have produced a new generation of image segmentation models with remarkable performance improvements—often achieving the highest accuracy rates on popular benchmarks—resulting in what many consider a paradigm shift in the field. Image segmentation can be expressed as a problem of pixel labeling with semantic labels (semanticsegmentation) or partitioning of individual objects (partitioning) (instancesegmentation). Semantic segmentation performs pixellevel labeling for all image pixels using a series of object categories (e.g., person, vehicle, tree, sky), making it a more difficult task than image classification, which predicts a single label for the the entire image. By detecting and delineating each object of interest in the image, instance segmentation broadens the reach of semantic segmentation. Our project now is divided in 3 tasks which we work on image segmentation for 2 datasets i.e., the VOC dataset and the cityscapes dataset. The image below shows the output of object segmentation forpretrained model. [1]

2 Methodology

We are performing 2 tasks in this project. Task 1 we need to train a model to get the object segmentation output for the VOC Dataset. We have used the VGG16 model for this task. The VGG16 model architecture was introduced by Simonyan in their research paper released in 2014. This network is characterized by its simplicity, using only 3x3 convolutional layer stacked on top of each other in increasing depth. The volume

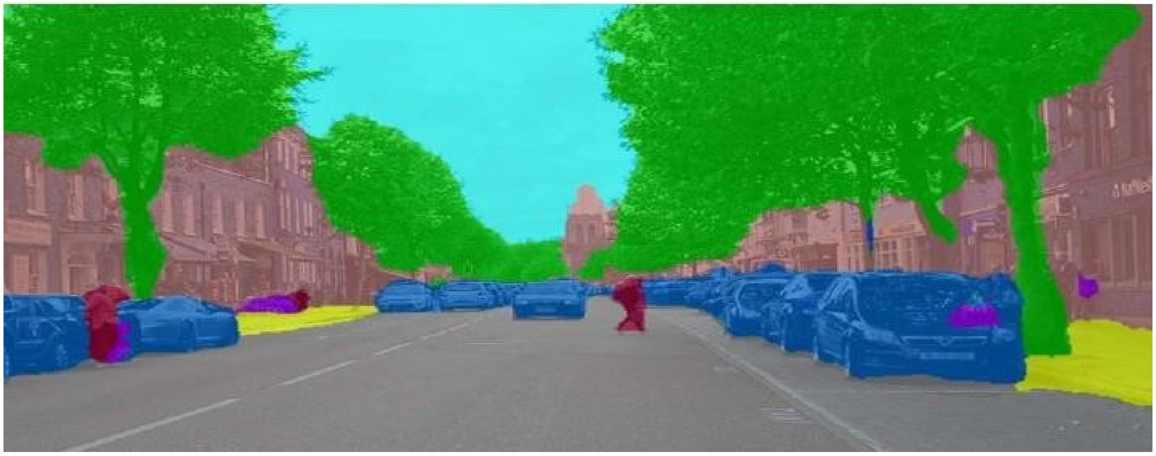

Figure 1: Example of object segmentation of a city(City Scapes Dataset)

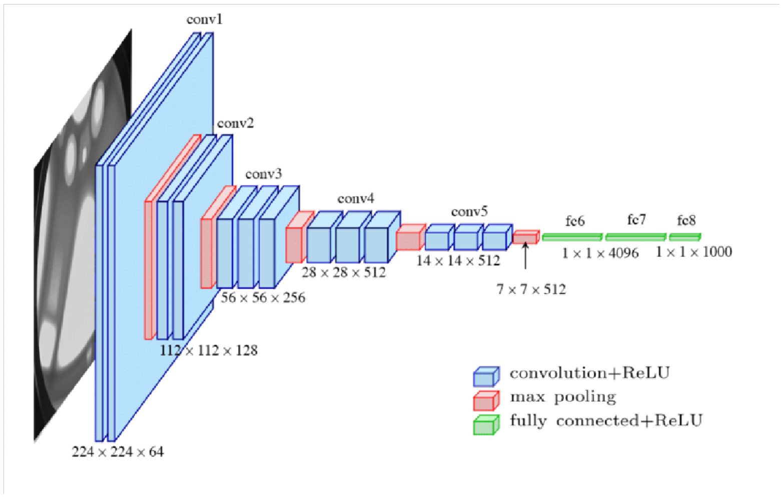

Figure 2: Image of the layers in VGG16

size is controlled by max pooling. Two fully-connected layers, each with 4096 nodes are then followed by a softmax classifier.

Task 1. We have been given a dataset of VOC12 dataset and we are asked to develop a model. VGG16 model has 2 layers that have 64 channels of 3*3 filter size and padding. After this it follows max Pooling with a stride of 2. Now we have two layers that have 256 convolutional layers and filter size (3,3). Now, max pooling is applied followed bys stride (2,2) which is same as the previous layer. Then there are 2 convolution layers of filter size (3,3) and 256 layers. VGG16 consists of 4 main classes and 21 sub classes and the main classes of the dataset are as ‘Person’, ’Vehicle’, ’Animal’, ‘Indoor’. So, in this task we are supposed to train the VOC12 dataset. We are training the model with VGG16 model. We used VGG16 model as it focuses on convolutional layers of 3x3 filter with stride 1 instead of having large number of hyperparameters and always uses same padding and maxpool layer of 2x2 filter of stride 2, Also, the main reason of using VGG16 model is that it follows the same arrangement of convolution and max pool layers throughout the whole architecture. [2]

Task 2. We are given cityscapes dataset and we have a task to implement a Recurrent Residual Neural Network based on the U-Net architecture for semantic segmentation. Cityscapes Dataset is a huge collection of preceding and trailing video frames. Each annotated image is the 20th image from a 30 frame video snippet (1.8s). It consists of 5000 annotated images with fine annotations and 2000 annotated images with coarse annotations. The classes are classified as Flat, Human, Vehicle, Construction, Object, Nature, Sky, Void. For medical fields semantic segmentation has become a major help and these techniques have been widely used and are generating good results. The most famous network the U-Net is widely used and many sub techniques have evolved from this technique. RU-Net and R2U-Net utilize the powers of U-Net and generates better results by considering more layers.

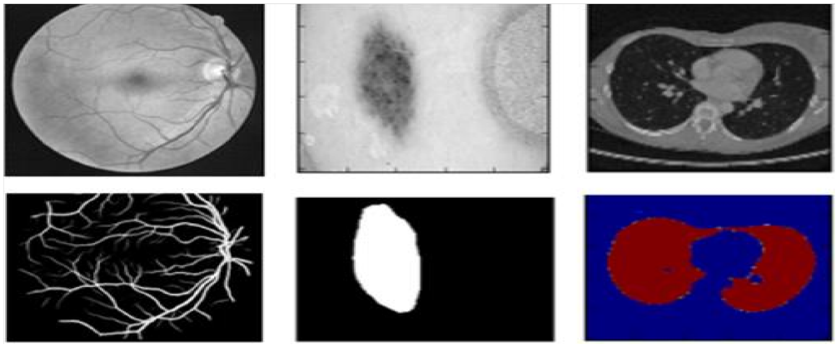

Figure 3: Object Segmentation of Medical field of skin cancer.

U-Net U-Net is a semantic segmentation architecture. It has two paths: one that contracts and one that extends. The convolutional network's contracting direction follows the traditional architecture. It consists of two 3x3 convolutions (unpadded convolutions) that are applied repeatedly, each accompanied by a rectified linear unit (ReLU) and a 2x2 max pooling operation with stride 2 for down sampling. We double the number of feature channels with each down sampling step. An up sampling of the feature map is followed by a 2x2 convolution ("up-convolution") that halves the number of feature channels, a concatenation with the correspondingly cropped feature map from the contracting direction, and two 3x3 convolutions, each followed by a ReLU in the expansive path. Due to the loss of border pixels in any convolution, cropping is required. A 1x1 convolution is used at the final layer to map each 64component function vector to the desired number of groups. The network has a total of 23 convolutional [3] layers.

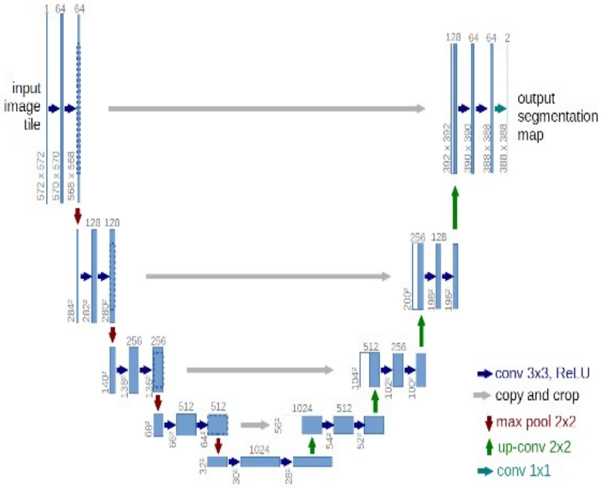

Figure 4: U-Net.

R2U-Net From the deep neural networks there are models that have been developed namely the RU-Net and the R2U-Net. These models utilize the powers of the previous existing models. The RCNN and its variants have already shown superior performance using different benchmarks. The recurrent residual convolutional operation can be expressed mathematically:

Let xi be the input sample in the l th layer of the RRCNN block and a center pixel of a patch located at (i,j) in an input sample on the k th feature ma in the RCL.

Eq. (1)

$$O^{l}_{ijk}(t) = (w^{f}_{k})^{T} * x^{f(i,j)}_{l}(t) + (w^{r}_{k})^{T} * x^{r(i,j)}_{l}(t-1) + b_{k}.$$

Now, the outputs of the RC L are fed to the standard ReLU activation function f are expressed as follows:

Eq. (2)

$$\mathscr{F}(x_l, w_l) = f[O^{l}_{ijk}(t)] = \max[0, O^{l}_{ijk}(t)],$$

The output of the previous variable is used for down sampling and up sampling layers in the convolutionl encoding and decoding units of the RU-Net model, respectively. The final outputs of RCNN units are passed through the residual unit as shown in the figure below:

Eq. (3)

$$x_{l+1} = x_l + \mathscr{F}(x_l, w_l).$$

3 Results and Discussion

Task 1 Metrics and Values

Metrics	Values
F1 score	79.41
Jaccard Score	65.15

Table 1: Values of respective metrics and values

Task 2 Metrics and Values

Metrics	Values
Accuracy	96.37
F1 score	96.32
Sensitivity	96.36
Jaccard Score	93.1

Table 2: Values of respective metrics and values

Task 3 Metrics and Values The epochs ran were limited compared to the Task2 but the results yielded proves that it may not yeild better results even on more epochs

Metrics	Values
Accuracy	96.22
F1 score	96.21
Sensitivity	96.22
Jaccard Score	92.82

Table 3: Values of respective metrics and values

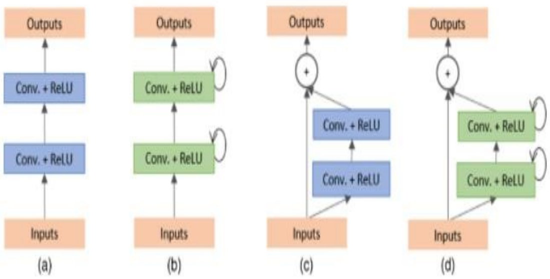

Figure 5: Different variants of the convolutional and recurrent convolutional units (RCUs) including (a) the forward convolutional unit, (b) the recurrent convolutional block, (c) the residual convolutional unit, and (d) the recurrent residual convolutional unit.

4 Conclusion

In this paper we have worked on Object segmentation and have performed the same with various models to conclude the best model. Attention R2U-Net model chosen in task3 performed nearly with the same accuracy for epochs chosen in this project, but if trained for higher epochs Attention R2U-Net might perform slightly better than the R2Unet for the object segmentation tasks

5 References

References

[1] (PDF) Image Segmentation Using Deep Learning: A Survey (researchgate.net)

[2] Image Classification using pre-trained VGG-16 model - KGP Talkie

[3] Recurrent Residual U-Net : University of Dayton, Ohio (udayton.edu)

[4] Recurrent residual U-Net for medical image segmentation (spiedigitallibrary.org)

[5] [1505.04597v1] U-Net: Convolutional Networks for Biomedical Image Segmentation (arxiv.org)