NNTI Project: Object Segmentation

Student 1:
Akash Kumar
7009735
akku00001@teams.uni-saarland.de

Student 2:
Harisree Kallakuri
7009317
haka00001@teams.uni-saarland.de

Abstract

In this project, we implemented the semantic segmentation on two different datasets. Semantic segmentation is recognizing, understanding what's the image in pixel level.The goal of semantic image segmentation is to label each pixel of an image with a corresponding class of what is being represented. We used VGG16 (Task 1), R2-Unet(Task 2) and FocusNet(Task 3) for performing the segmentation. Results show FocusNet performs slightly better than the R2-Unet.

Keywords:

Object segmentation, deep learning, recurrent convolutional neural network, R2-Unet, vgg16, , FocusNet

1 Introduction

The extensive use of mobile camera and digital imagery tools increases popularity and demands for foreground. The increasing number of pixels like in animal furs, human hair, insect antennas, floral stamps, cavity of jewelry within the ships, handrail of boats, etc., revealed further details in the resolution of an image.

From 1970s-90, sequential application of pixel-based treatment (filter editors, regionally grow) and mathematical (fitting lines, circles and ellipse) were initially performed to develop compound rule-based systems to solve particular tasks. In addition, medical image analysis was performed in the 1990s. There's an analogy with expert systems with many if-then-else declarations, popular during the same time in artificial intelligence.

According to (1), Segmentation is a process in which an image is subdivided by a particular feature, to capture a region of interest, into several sub-regions. In the field of medicine, segmentation has huge applications. There have been many efforts in the field of research and development to solve the problems facing the segmentation process, but more efficient and effective work is needed(3). Segmentation of the medical image, which identifies pixels of organs or injuries from background medical images like CT or MRI, is one of the most demanding tasks of analyzing medical images to provide critical information on the shapes and volumes of these organs. By applying available technologies, several researchers have proposed various automated segmentation systems. Previous systems are based on traditional methods like filters and mathematical methods for detecting edges. Then, machine

learning approaches that extract handmade features have for a long time become a dominant technique (2)

The design, extraction and implementation of these features were always the main concern for the development of such a system and their complexity was considered to be considerable limit. In the 2000s, deep learning approaches were developed due to improvements in hardware and their significant capabilities in image processing were demonstrated. The promising ability to learn deeply has made them primary to segment images and, in particular, to segment medical images(4).

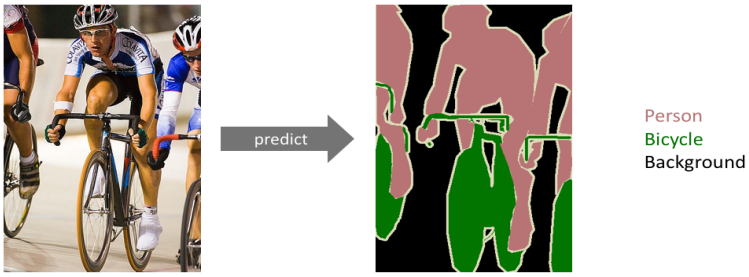

Figure 1: Segmented image of man on cycle

2 Methodology

The project illustrates object segmentation via three tasks. Given below explains the entire procedure, methods to carry out the investigation. It also explains how results were obtained and analyzed.

2.1 Task 1:

The major objective of this task is to identify objects in realistic scenes from a variety of visual object classes (i.e. not pre-segmented objects). In the sense that a training set of labelled images is provided, it is fundamentally a supervised learning learning problem(11). The task is trained on PASCAL VOC dataset.The twenty object classes that have been selected are:
Person: person
Animal: bird, cat, cow, dog, horse, sheep
Vehicle: aeroplane, bicycle, boat, bus, car, motorbike, train
Indoor: bottle, chair, dining table, potted plant, sofa, tv/monitor

2.1.1 Network Architecture

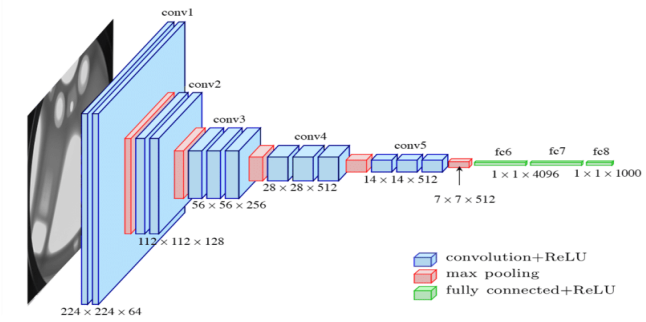

Figure 2: Architecture of vgg16 model

The implementation uses VGG16 model, the overall structure includes 5 sets of convolutional layers, followed by a MaxPool as shown in figure 2. But the difference is that as the depth increases that is as we move from VGG11 to VGG19 more and more cascaded convolutional layers are added in the five sets of convolutional layers.
As VGG16 acts as a classifier and for doing the segmentation task that is to predict the each pixel of the image, we need to connect coarse outputs back to the pixel (1). For doing so we attached an extra convolution layer at the end of the VGG16 architecture. The model uses Adam and cross-entropy loss as optimization algorithm and loss function respectively.

2.2 Task 2:

The task to is performed in reference to the paper (2). In our task, we use the cityscape dataset. The dataset focuses on the semantic understanding of street scenes. The classes used on this dataset are:
flat: road, sidewalk, parking, rail track
human: person, rider
vehicle: car,truck, bus, on rails, motorcycle, bicycle, caravan, trailer
construction: building, wall, fence, guard rail, bridge, tunnel
object: pole, pole group, traffic sign, traffic light
nature: vegetation, terrain
sky: sky
void: ground, dynamic, static
The evaluation is done by calculating the metrics. The metrics are given by the following following formulae:
TN = True Negative
TP = True Positive
FN = False Negative
FP = False Positive
Accuracy is the proportion of the total number of predictions that were correct. Accuracy 'AC' is given by:

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

Sensitivity is a measure of the proportion of actual positive cases that got predicted as positive (or true positive).Sensitivity 'SE' is given by:

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

Specificity is defined as the proportion of actual negatives, which got predicted as the negative (or true negative).Specificity 'SP' is given by:

$$SP = \frac{TN}{TN + FP} \tag{3}$$

Dice Coefficient is defined as two times the area of the intersection of A and B, divided by the sum of the areas of A and B. Dice Coefficient 'DC' is given by:

$$DC = 2\frac{|GT \cap SR|}{|GT| + |SR|} \tag{4}$$

Jaccard Similarity is defined as the area of the intersection divided by the area of the union:Jaccard Similarity 'JS' is given by:

$$JS = \frac{|GT \cap SR|}{|GT \cup SR|} \tag{5}$$

In this work according to the, we have evaluated the proposed approaches on both patch-based and entire image-based approaches. To switch from the patch-based approach to the pixel-based approach that works with the entire image, we must be aware of the class imbalance problem. In the case of semantic segmentation, the image backgrounds are

assigned a label and the foreground regions are assigned a target class.
Cross-entropy loss and dice similarity are introduced for efficient training of classification and segmentation tasks. Each pixel is assigned a class label with a desired boundary that is related to the contour of the target lesion in identification tasks. To define these target lesion boundaries, we must emphasize the related pixels. However, semantic segmentation approaches that utilize Deep Learning have become very popular in recent years in the field of medical image segmentation, lesion detection, and localization.

The architecture of segmentation tasks requires both convolutional encoding and decoding units. The encoding unit is used to encode input images into a larger number of maps with lower dimensionality. The decoding unit is used to perform up-convolution (de-convolution) operations to produce segmentation maps with the same dimensionality as the original input image.
This task performs segmentation using the model R2-Unet. R2U-Net architectures are designed to have the same number of network parameters when compared to U-Net and ResU-Net, and RU-Net and R2U-Net show better performance on segmentation tasks. These two approaches utilize the strengths of all three recently developed deep learning models. RCNN and its variants have already shown superior performance on object recognition tasks using different benchmarks.

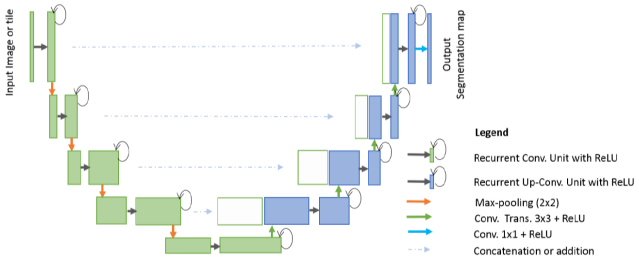

Fig. 3. RU-Net architecture with convolutional encoding and decoding units using recurrent convolutional layers (RCL) based U-Net architecture. The residual units are used with RCL for R2U-Net architecture.

Figure 3: Architecture of R2U-net model

2.3 Task 3:

2.3.1 Network Architecture

For this task, we have to improve the existing results of task 2. So in this work, we used an attention gate, these are widely used for image captioning, machine translation, and classification tasks in natural image analysis, knowledge graphs, and language processing (NLP)(14). As proposed in (15), for this work we combined the R2-UNet model with an attention gate,known as FocusNet and previously it worked better than the start-of-the-art architecture (16). Rest all parameters and components are the same as the model implemented in task 2.

3 Results and Discussion

3.1 Task 1:

In this, we trained our model for 30 epochs and got the following metrics :

Metrics	Values
F1 score	0.8560
Dice Co-efficient	0.7482

Table 1: Values of respective metrics and values using vgg16

3.2 Task 2:

In this, we trained the model for 40 epochs with Adam as an optimizer and cross entropy as a loss. We got the following metrics on **training** and **testing** datasets respectively.

Metrics	Values
Accuracy (AC)	0.9904
Sensitivity (SE)	0.8713
Specificity (SP)	0.9950
F1 score (F1)	0.8713
Jaccard similarity (JQ)	0.7720

Table 2: Values of respective metrics and values using R2U-net on training data

Metrics	Values
Accuracy (AC)	0.9843
Sensitivity (SE)	0.8042
Specificity (SP)	0.9918
F1 score (F1)	0.8042
Jaccard similarity (JQ)	0.6725

Table 3: Values of respective metrics and values using R2U-net on testing data

3.3 Task 3:

In this we implemented the same model as in task2 with an added attention gate. We have the following results for training and testing dataset respectively.

Metrics	Values
Accuracy (AC)	0.9982
Sensitivity (SE)	0.8854
Specificity (SP)	0.9932
F1 score (F1)	0.8632
Jaccard similarity (JQ)	0.8056

Table 4: Values of respective metrics and values using FocusNet on training data

Metrics	Values
Accuracy (AC)	0.9862
Sensitivity (SE)	0.8323
Specificity (SP)	0.9543
F1 score (F1)	0.8413
Jaccard similarity (JQ)	0.7523

Table 5: Values of respective metrics and values using FocusNet on testing data

4 Conclusion

In this project, we completed all the assigned tasks for which we used three different models to perform image segmentation. In Task 1, we got 0.8560 as F1 score but the predicted mask was not much accurate. So, to improve the results we can use more complicated architecture than VGG16. In Task 2, R2-Unet performs very well on both training and testing datasets. In Task 3, we improved the results of task 2 by adding an attention unit to the R2-Unet, known as FocusNet.The FocusNet performs slightly better than R2-Unet and we would get better result by tuning the parameters and training it for more epochs.

References

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

[2] Alom, Md Zahangir, et al. 'Recurrent Residual Convolutional Neural Network Based on U-Net (R2U-Net) for Medical Image Segmentation'. ArXiv:1802.06955 [Cs], 5, May 2018. arXiv.org, http://arxiv.org/abs/1802.06955.

[3] Maier, Andreas. 'Segmentation and Object Detection — Part 1'. Medium, 20 Aug. 2020, https://towardsdatascience.com/segmentation-and-object-detection-part-1-b8ef6f101547.

[4] Yang, Chenglin, et al. Meticulous Object Segmentation. Dec. 2020. arxiv.org, https://arxiv.org/abs/2012.07181v1.

[5] Gandhi, Rohith. 'R-CNN, Fast R-CNN, Faster R-CNN, YOLO — Object Detection Algorithms'. Medium, 9 July 2018, https://towardsdatascience.com/r-cnn-fast-r-cnn-faster-r-cnn-yolo-object-detection-algorithms-36d53571365e.

[6] Sharma, Anshika Singh, Pradeep Khurana, Palak. (2016). Analytical review on object segmentation and recognition. 524-530. 10.1109/CONFLUENCE.2016.7508176.

[7] Hesamian, M.H., Jia, W., He, X. et al. Deep Learning Techniques for Medical Image Segmentation: Achievements and Challenges. J Digit Imaging 32, 582–596 (2019). https://doi.org/10.1007/s10278-019-00227-x

[8] Litjens, Geert, et al. 'A Survey on Deep Learning in Medical Image Analysis'. Medical Image Analysis, 2, vol. 42, Dec. 2017, pp. 60–88. arXiv.org, doi:10.1016/j.media.2017.07.005.

[9] Masood, Saleha Sharif, Muhammad Masood, Afifa Mussarat, Yasmin Raza, Mudassar. (2015). A Survey on Medical Image Segmentation. Current Medical Imaging Reviews. 11. 3-14. 10.2174/1573405611011504231034 41.

[10] Image Analysis (Medical Imaging) - an Overview | ScienceDirect Topics. https://www.sciencedirect.com/topics/medicine-and-dentistry/image-analysis-medical-imaging. Accessed 31 Mar. 2021.

[11] The PASCAL Visual Object Classes Homepage. http://host.robots.ox.ac.uk/pascal/VOC/. Accessed 31 Mar. 2021.

[12] Social Network for Programmers and Developers. https://morioh.com/p/43eab5c453b9. Accessed 31 Mar. 2021.

[13] 'An Overview of Semantic Image Segmentation.' Jeremy Jordan, 22 May 2018, https://www.jeremyjordan.me/semantic-segmentation/.

[14] 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.

[15] leejunhyun. LeeJunHyun/Image Segmentation. 2018. 2021. GitHub, https://github.com/LeeJunHyun/Image Segmentation.

[16] Kaul, Chaitanya, et al. "Focusnet: An Attention-Based Fully Convolutional Network for Medical Image Segmentation." 2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019), IEEE, 2019, pp. 455–58. DOI.org (Crossref), doi:10.1109/ISBI.2019.8759477.