Comparing U-Net and Mask R-CNN for medical cell segmentation

Bjørn Kurt Hansen
Technical University of Denmark
Lyngby, Denmark
s193035@student.dtu.dk

***Abstract*—Finding quick and efficient ways to perform image segmentation has been of importance in many fields ever since AI has become popular. Image segmentation for medical fields is of particular importance as it can help ease the time constraint put on medical professionals as well as minimise human error in diagnosing disease. Over time, deep learning based image segmentation has become more reliable and in recent years have become robust enough to be considered for clinical use. A popular medical image segmentation task is to segment out areas of an image which contain cells from areas which do contain cells. In this work the task of segmenting glands from colon histology images is done. Two different methods are tested on the same dataset namely, the U-net architecture and the masked-RCNN algorithm. It was shown that both methods perform well on the segmentation task, the U-net achieving an IoU score of 0.33 and F1 score of 0.64 and mask-RCNN achieving an IoU score of 0.42 and F1 score of 0.59.**

***Index Terms*—Cell segmentation, U-Net, Mask R-CNN, Deep Learning**

I. Introduction

The use of deep learning in connection to performing image segmentation within medical fields is of particular importance as it can help ease the time constraint put on medical professionals as well as minimise human error in diagnosing disease. Many diseases use screening techniques to evaluate risk and give localization information for a given patient. These screenings are obviously of great importance for the health of the patient as quick and accurate detection of diseases lead to the patient receiving adequate care in time. Colo rectal cancer is an example of a common disease which relies on accurate and timely detection for the patient to receive necessary care. The first step to detection is the segmentation of the area of interest from a medical image. In clinical settings, however, the segmentation and detection of abnormal cancer areas is challenging; the task is in great need of a highly accurate and automatic segmentation method. The segmentation of cancerous colon tissue in images is difficult mainly because the cancerous tissues vary in size and because the edges of the areas of interest are often blurred. [1]

Many deep learning architectures have been developed and applied to the task of medical image segmentation in the recent past; a popular architecture for the task of segmentation is the U-net which was originally introduced in 2015 [2]. The U-net has been shown to be successful for segmentation tasks within various areas; because of this success, various architectures similar to U-net, such as U-Net++ and ResUNet++ have been developed and shown improved segmentation performance [3] [4].

The U-net architecture looks as Figure 1 shows, and is a fully connected convolutional network architecture. The U-net consists of an encoder side and a decoder side; the encoder which encodes images into a feature space of small dimension by applying kernels, and a decoder which maps this information into spatial categorisation to perform segmentation. The characteristic of the U-net model is that the input of each encoder layer is also concatenated to the output of its corresponding decoder. By performing multiple operations in the encoder, some spatial information is lost, and the link between the encoder and decoder is able to recover this lost information which produces the segmentation characteristic.

Work done on segmentation tasks spans outside of medical imagery; frameworks such as the mask RCNN have been developed for multi use object instance segmentation and detection tasks [5]. The mask RCNN is an extension of the Faster RCNN framework [6] and has two fundamental stages. The first stage generates proposals about the regions where there might be an object based on the input image. The second stage predicts the class of the defined objects, defines bounding box coordinates, and generates a pixel mask for the object. The described mask RCNN framework can be seen in Figure 2.

This first stage consists of a region proposal network which through *scanning* the feature map proposes regions which may contain an object. Next, a set of predefined so called anchors bind features to the original image location. The anchors area a set of boxes with predefined locations and scales relative to the given input image. Binary classes (object and background) and bounding boxes are assigned to individual anchors according to a intersection over union (IoU) threshhold value. The concept of the first stage in the mask RCNN is relatively similar to the concept of extracting information through convolving, downsampling, and upsampling.

The second stage of the mask RCNN is a neural network which is similar to the first stage, however, here the it takes the proposed regions of interest (ROI) and with a method called ROIAlign, the areas of interest are located to the the relevant areas of the feature map. From here there are various neural network branches which perform assignments independently for each object on a pixel level, and therefore have their own loss functions. There is a classifier, bounding box, and mask generating network.

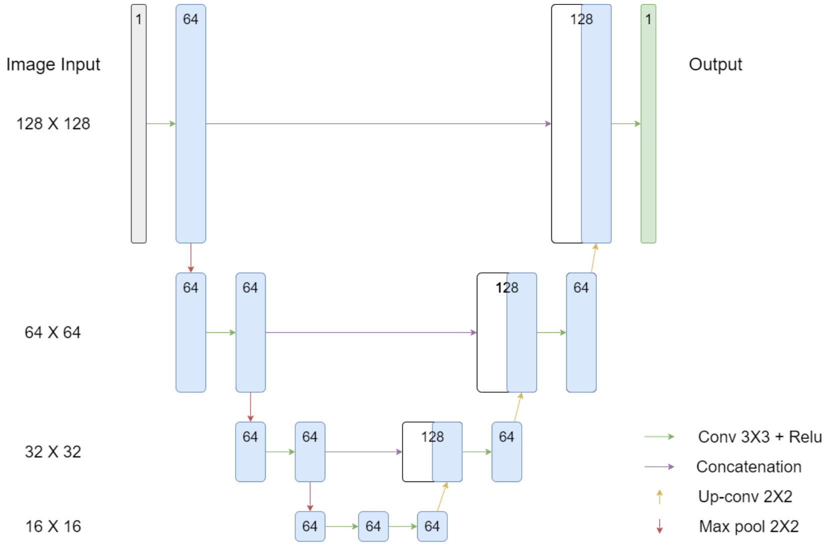


Fig. 1: U-Net Framework

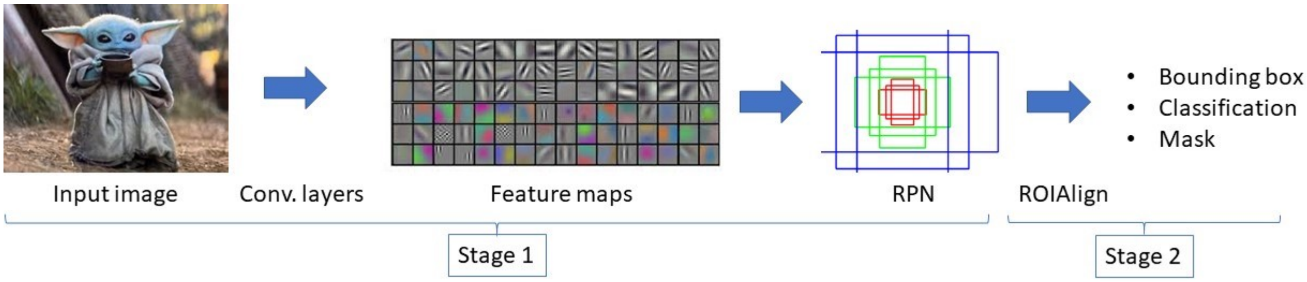


Fig. 2: Mask RCNN Framework

II. Methods

The code used for this work is publicly available under MIT license at https://github.com/bh1995/02456-Deep-learning/tree/master/Project. The data used in this work is publicly available, and was originally used in a research contest called the GlaS Challenge Contest [7] [8].

A. Data description

The data are stained colon histology images, the contest where these images were originally used, divided the images into two parts; part A images and part B images. Both parts are similar but Part B images are slightly more difficult to segment. In this project the part labeling was disregarded and all test and training images were used equally. A total of 105 images were used, the training set consisted of 85 images, and the test set of 20. The majority of the images have the pixel dimensions 775x522, however a few images have different dimensions.

B. Deep Learning Models

The two models which are explored in this work are the U-Net and the Mask RCNN. The U-Net architecture as seen in Figure 1 was implemented using the deep learning framework, PyTorch. A unique data loader function was created for the U-Net; the data loader is implemented such that each input image is converted into a normalized pixel array and re-sized to dimensions of 128x128. Better performance on the test set was seen when the images were resized, as well as the model training considerably faster. Various loss functions were tried such as focal and dice loss, but the binary cross entropy was found to give the best results and was therefore used in the training of the U-Net.

The Mask RCNN framework implementation is different compared to the U-Net implementation. The mask RCNN takes as input, the raw input image, and a dictionary which holds information about the target image. The dictionary contains the mask of the input image, the labels (class) of each object, and the coordinates of the bounding boxes for each object in the image. For this, firstly a data loader is constructed which, via the different colors of the pixels in the mask, is able to find the number of objects the image holds, the coordinates of these objects and their class (which is binary, either cell or background).

Since the mask RCNN produces various outputs (bounding box, mask, classifier) by having a branch-like framework, there are various loss functions at play simultaneously. The fully connected layer portion of the network uses per-pixel softmax and a multinominal loss. This means that the mask prediction task (the boundaries of the object) and the class prediction task (what is the object being masked) are coupled. Mask-RCNN decouples these tasks such that the existing bounding-box prediction head predicts the class while the mask branch generates a mask for each class. The loss being used is per-pixel sigmoid with binary cross entropy. To summarize, the mask RCNN uses Sigmoid with binary cross entropy, and the fully connected portion uses soft-max.

Since the data set size is small, transfer learning was utilized with the help of a pre-trained model from the torchvision modelzoo [9]. The model used was a mask RCNN model with a 50 layer ResNet backbone, which was pre-trained on the coco data set. The final layer of this model was fine tuned via training on the colon histology images.

C. *Interpretation*

In order to measure the performance of the models, the F-score and IoU score were calculated on the test images and reported.

III. Results

Both the U-net framework as well as the mask RCNN frameworks were trained and tested using the Google Colab environment. An example of segmentation masks which each framework produced can be seen in Figure 3 and Figure 4. The U-net was trained for a total of 200 epochs, whereas the pre-trained mask RCNN was trained for nine epochs. As can be seen in Figure 5, the mask RCNN was trained for far fewer epochs compared to the U-net but in practice took about five times longer to train. The U-net plateaued around 200 epochs, whereas the pre-trained mask RCNN plateaued around eight epochs.

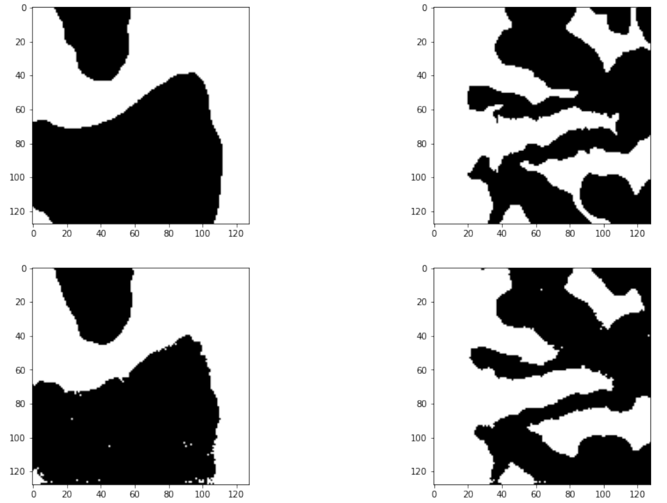


Fig. 3: True segmentation (top), U-Net predicted segmentation (bottom)

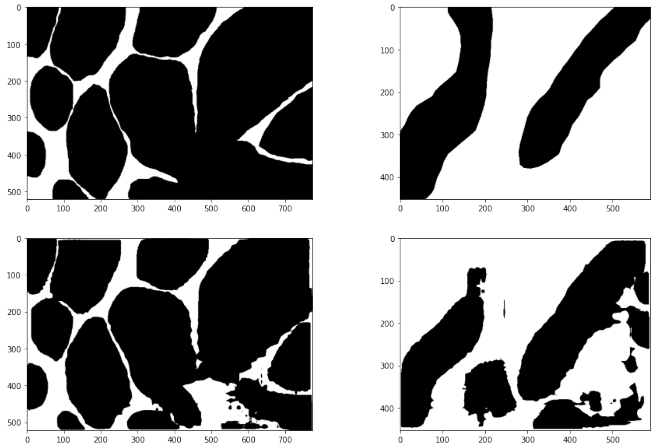
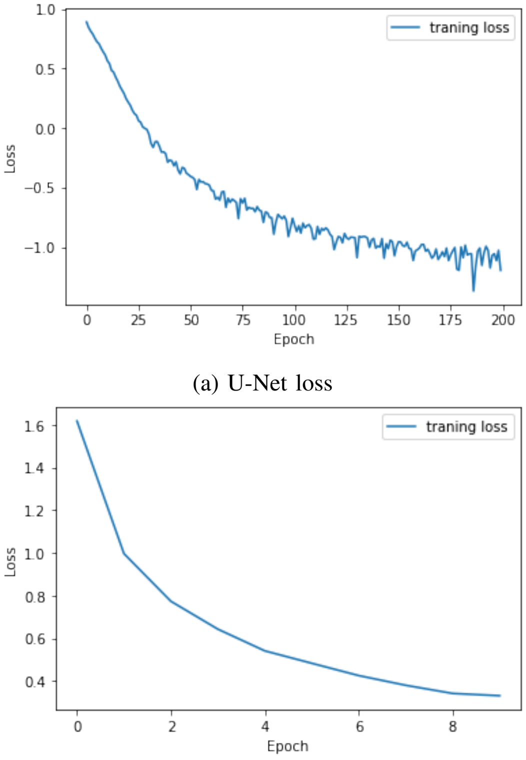


Fig. 4: True segmentation (top), Mask RCNN predicted segmentation (bottom)



(a) U-Net loss

(b) Mask RCNN loss

Fig. 5: Learning curves during training for both frameworks.

A. *Performance*

As seen in Table I, the IoU score acheived by the mask RCNN is slightly higher than the IoU score of the U-net, whereas for the F1 score, the U-net scores higher than the mask RCNN. In general, there is no standard benchmark values for IoU or F1 which applies to all datasets. So, to give a sense of the quality of the results in this work, they can be compared to the results of the top teams from the original GlaS contest where the data was originally used [8]. As previously mentioned, the GlaS contest divided the images up into two parts whereas in this work all images were used equally. The top five teams in the GlaS competition achieved F1 scores in the range of 0.6 and 0.8.

TABLE I: Performance of the different models on the test data set.

Model	IoU-score	F1-score
U-Net	0.332	0.643
Mask RCNN	0.419	0.592
Bench mark from GlaS contest	-*	0.6-0.8

* IoU was not a considerd metric in the GlaS contest

IV. DISCUSSION

In this work, the U-net framework was compared to the mask RCNN framework for the task of medical image segmentation. The first thing to point out is that both frameworks were successful in segmentation of the images and that both achieved similar segmentation performance. This performance is mid range compared to the other teams in the original GlaS contest [8]. It should be noted that the mask RCNN was pre-trained on the coco data set which is a large standard benchmark data set comprised of 80 *everyday* object classes such as dogs and cars. The objects in the coco data set are vastly different in nature compared to the colon histology images; it is likely that if the mask RCNN model were fully trained on a larger amount of colon histology images greater performance would be seen.

The mask RCNN model took a considerably longer time to train compared to the U-net, but produces more information than the U-net (such as the mask scores and bounding boxes). The extra information in this work was not much use, as the goal was simply to compare two architectures on the task of segmentation, however, in some practical circumstances such as live colonoscopy inspection, the bounding boxes could be of great use. The mask RCNN is an interesting model which could by applied to many computer vision tasks with high success where detection, classification, and segmentation of typical everyday objects is required. For the sole task of segmentation, however, the U-net is arguably preferable do to its simplicity which not only gives performance advantages but also makes the architecture easier to build upon and adapt, such as the UNet++ architecture has done. The mask RCNN on the other hand, due to its shear complexity, would be far more difficult to alter and develop.

REFERENCES

[1] X. Jia, X. Xing, Y. Yuan, L. Xing, and X. Meng, "Wireless capsule endoscopy: A new tool for cancer screening in the colon with deep-learning-based polyp recognition," *Proceedings of the IEEE*, vol. 1, pp. 178–197, 2019.

[2] O. Ronneberger and T. Fischer, Brox, "U-net: Convolutional networks for biomedical image segmentation," *MICCAI*, pp. 234–241, 2015. [Online]. Available: https://arxiv.org/abs/1505.04597

[3] Z. Zhou, M. Siddiquee, N. Tajbakhsh, and J. Liang, "Unet++: A nested u-net architecture for medical image segmentation," *DLMIA*, vol. 4, pp. 3–11, 07 2018. [Online]. Available: https://arxiv.org/abs/1807.10165

[4] D. Jha, P. Smedsrud, M. Riegler, D. Johansen, T. De Lange, P. Halvorsen, and H. Johansen, "Resunet++: An advanced architecture for medical image segmentation," *IEEE ISM*, vol. 4, p. 225–2255, 11 2019. [Online]. Available: https://arxiv.org/abs/1911.07067

[5] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," *Computer Vision and Pattern Recognition*, 03 2017. [Online]. Available: https://arxiv.org/abs/1703.06870

[6] S. Ren, K. He, J. Sun, and R. Girshick, "Faster r-cnn," *Computer Vision and Pattern Recognition*, 06 2015. [Online]. Available: https://arxiv.org/abs/1506.01497

[7] D. of Computer Science University of Warwick. (2020, 12) Glas contest download. [Online]. Available: https://warwick.ac.uk/fac/sci/dcs/research/tia/glascontest/download/

[8] K. Sirinukunwattana and et.al, "Gland segmentation in colon histology images: The glas challenge contest," *Computer Vision and Pattern Recognition*, 03 2016. [Online]. Available: http://arxiv.org/abs/1603.00275

[9] pytorch.org. (2020) Torchvision.models. [Online]. Available: https://pytorch.org/docs/stable/torchvision/models.html