A Brief Analysis of U-Net and Mask R-CNN for Skin Lesion Segmentation

Erick Alfaro
Computing Engineering School
Instituto Tecnológico de Costa Rica
Cartago, Costa Rica
erick.ah06@gmail.com

Ximena Bolaños Fonseca.
Computing Engineering School
Instituto Tecnológico de Costa Rica
Cartago, Costa Rica
xbolanosfonseca@gmail.com

Enrique M. Albornoz
Instituto sinc(i)
FICH-UNL / CONICET
Santa Fe, Argentina
emalbornoz@sinc.unl.edu.ar

César E. Martínez
Instituto sinc(i)
FICH-UNL / CONICET
Santa Fe, Argentina
cmartinez@sinc.unl.edu.ar

Saúl Calderón Ramrez
Computing Engineering School
Instituto Tecnológico de Costa Rica
Cartago, Costa Rica
sacalderon@itcr.ac.cr

sinc(*i*) Research Institute for Signals, Systems and Computational Intelligence (fich.unl.edu.ar/sinc)
E. Alfaro, X. Bolaños Fonseca, E. M. Albornoz, C. E. Martínez & S. Calderón Ramrez; "A Brief Analysis of U-Net and Mask R-CNN for Skin Lesion Segmentation"
IEEE International Work Conference on Bioinspired Intelligence (IWOBI 2019), 2019.

***Abstract*—A brief analysis on the use of two deep neural architectures, the U-Net and Mask R-CNN for the segmentation of skin lesions in dermoscopic images is presented. The two systems were adapted to use the dataset provided by the International Skin Imaging Collaboration (ISIC) for its 2017 challenge and different experiments were carried out. Results showed that the Mask-R-CNN obtained better performance than U-Net, also with lower computation times, being a feasible architecture to further analysis and application also to skin lesion classification.**

***Index Terms*—segmentation, melanoma, skin lesions, dermoscopic images, deep learning**

I. Introduction

Malignant skin lesions have become a very common disease in the last years due to higher levels of ultra violet sun ray exposure, as the ozone layer becomes more fragile. In many cases, the disease remains undetected until its last stages. People often underestimate the impact of the sun radiation, where a 10% of its whole energy is in the from of ultraviolet light (UVA, UVB) the main culprit of skin lesions in the world [1]. The lesions include skin cancer, the most common cancer in the US (with more than 5 million cases annually) and its most lethal form, the melanoma, causes around 9000 deaths per year [2].

The ability to isolate the lesion and diagnose it with a high enough precision becomes a major problem in medicine because there are cases where doctors in the area skip certain processes due to its complexity or simply they trust more in their professional expertise, which can lead to wrong diagnostics [2]. Furthermore, often patients have to wait long times before an appointment with skin lesion expert in public health services, or have to invest significant time moving to the main hospitals and clinics [3]. Therefore, an automatic skin lesion assessment for a computer assisted diagnosis would enable more accurate, faster, and even remote diagnosis of patients. It also would enable dermatologists to focus in patients with the most dangerous skin lesions and treat them sooner, an usual problem specially in the case of public health services with long wait lines.

In this work we propose a system based on deep neural networks for the segmentation of skin lesion in dermoscopic images. Skin lesion segmentation is the first step in its analysis in most of computer based skin lesion analysis systems [4]. We compare the performance of two fully convolutional architectures, U-net and the R-Mask convolutional network. Up to our knowledge, the R-Mask convolutional network have not been used for this task.

II. State of the Art

The segmentation of dermoscopic images is not a new topic, and actually the ISIC has made challenges for three consecutive years since 2016 and so a wide range of papers have been released on the matter. One of the first deep learning approaches reported consists on an improved Convolutional Neural Network (CNN) architecture, with the original image pixels as input in a hierarchical way to learn a set of nonlinear transformations that represent the contents of the image. The method achieved a high accuracy score, but the melanoma test results lagged due to the lack of image intensity, uniformity, and presence of imaging artifacts, which resulted in a few false positive results [5].

In [6], a multitask network is proposed based on the GoogleNet network; the outputs are then passed as input to three main components where one is devoted to the segmentation (4 deconvolutional layers) and the other two output the probability of the lesion being a melanoma or a seborroic keratosis (two fully connected layers each). In [7] authors proposed a multi-stage Fully Convolutional Network (FCN) which iteratively learns the lesion boundaries while a parallel integration join all the segmentation results during the process to approach a higher segmentation level.

On these works it has been noted a common coarse in the segmentation generated by FCN and SegNet since they

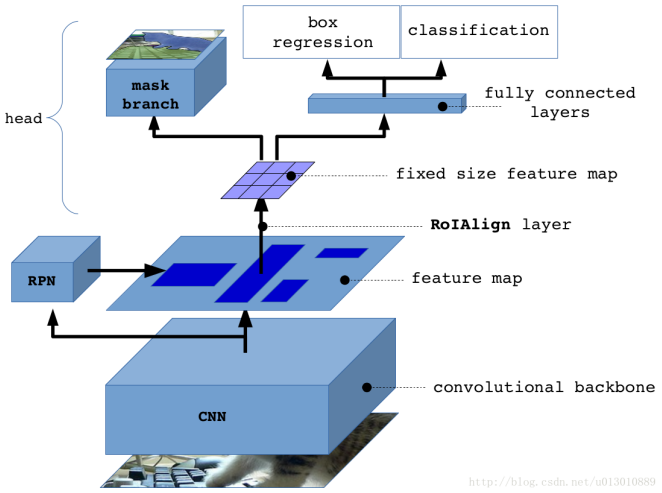


Fig. 1: Mask R-CNN architecture.

do not obtain anatomically plausible masks. To overcome this a Recurrent Neural Network (RNN) was trained to learn contextual relationships between pixels, preserving a local and global perspective with the use of a Long Short Term Memory (LSTM), obtaining a Jaccard score of 0.93 on the ISBI 2016 challenge [8].

There are papers submitted on the ISBI 2017 challenge where the participants accomplished great results, beginning with a Resnet network which adds various convolutional-deconvolutional layers that upsample the Resnet feature maps to output the score mask, also adding up to 8,000 images to the original 2,000; the training phase used the pre-trained Imagenet weights and went on fine tuning the network to reach the third place on the challenge [9]. The second place used a modified U-Net architecture with predictions were equal sized as its inputs, a Relu activation function was used with for all non-linear layers, and a significant amount of transformations was used to increase the training dataset from 2,000 to 20,000 [4]. As for the segmentation accuracy, a Jaccard index of 0.71 was achievec. Finally, the first place focused on creating a robust CDNN framework capable of handling images under various conditions instead of dealing with complex pre and post processing algorithms; as a pre-processing step they added three more channels to the images and in the post-processing they used a high threshold on the output map to determine the lesion center, then a lower one to look around it and finally taking as segmentation the area which embraces this center [10]. The authors yielded a Jaccard index of 0.784.

III. Materials and Methods

The Mask R-CNN architecture is described in [11], which consists of three main sections: the backbone, the region proposal network (RPN) and the head as can be seen in Figure 1. The backbone for our purposes was the Resnet-101 with its five stages.

The input image is first processed by the backbone which result is later fed into the RPN to search for potential regions of interest (ROI). Here, the purpose is to focus the labor of the following layers only on these regions instead of the whole image; the head layers consists of three branches (essentially composed of fully connected layers) each processing in parallel and predicting a different aspect of the ROIs, having as final outputs the regions mask, class and bounding box.

However, as we are only focusing on, binary segmentation our primary interest will be the mask output which is used on a post-processing stage to generate the actual image segmentation as asked by the ISIC. This is because Mask R-CNN can predict multiple masks when a big lesion comes in; even though this can be seen as over-segmentation, when they are joined together to actually make a better boundary detection. This point will be explained in next sections.

A. Implementation

The network was trained and evaluated using the released ISIC 2017 challenge dataset [8], [12] which has 2000 training images, 150 validation images and 600 test images.

The method was implemented with Python 3.5 using pytorch 0.3. The experiments were conducted on a system with a 12 Gb Geforce Titan GPU and a Intel(R) Core(TM) i7-8700 CPU @ 3.20GHz.

B. Pre-processing

The network is designed to work with the COCO dataset [13], and expects an annotations file. So, the training sub-set had to be processed to generate this files in order to initiate the network training. No other pre-processing was made.

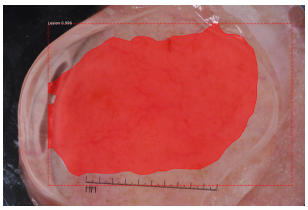
C. Training

The total number of epochs was set to 360 divided as follows: 40 to train only the network head with a learning rate of 0.01; 120 to train only the Resnet backbone from the 4 stages and up with a learning rate of 0.01, and finally 200 to fine-tune the whole network with a learning rate of 0.001. A 5-fold cross validation scheme was applied. All these stages used the Adam optimizer and, as a common practice, a validation was executed after each epoch.

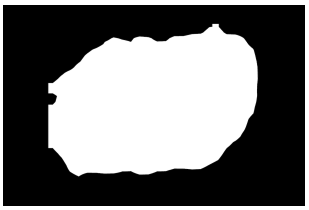
The network itself makes a resize of every image which was set to 1024×1024 pixels, but the output image has the same dimension as the input image. Although, the network's output is not actually an image as we can observe in the post-processing section.

D. Post-processing

For prediction purposes, the results of the network must be post-processed because its actual output is an array of masks (usually one per each ROI). These masks were overlapped over the original image with different colors and an alpha factor for transparency, as can be seen in Figure 2a. We modified this format to make all the masks black-and-white binary images, which let us join all the possible different masks into a final binary mask, as shown in Figure 2b.



(a) *Original Mask R-CNN output format*



(b) *Out modified output format*

Fig. 2: Comparison between our outputs and Mask R-CNN original outputs.

TABLE I: Evaluation over validation subset with U-Net

Number of epochs	Dice	Jaccard
160	0.606299	0.515993
180	0.616017	**0.539274**
200	0.630315	0.521581

IV. Experiment and Results

All the predictions made by the network were evaluated using the well-known similarity indexes for images segmentation, Dice and Jaccard, being the last one the one used by the ISIC to officially rank the participants' predictions. To ensure comprehension these indexes are calculated as follows:

$$Jaccard(A,B) = \frac{|A \bigcap B|}{|A \bigcup B|}$$

$$Dice(A,B) = \frac{2|A \cdot B|}{|A| + |B|}$$

with A being the ground truth mask and B the prediction of the network.

A. Experiments with U-Net

The results obtained using the U-Net as a stand-alone classifier are shown in Tables I and II.

It can be observed that for 180 epochs the network reach its best performance in terms of the Jaccard index. The number of epochs reflects the final portion of the training, but approximately at 75 epochs the training gets almost its final performance.

B. Experiments with Mask R-CNN

We chose to take various checkpoints from the training processes as the potential best model, so we tested the models from the epochs 160, 180 and 200 against the test subsets to check whether or not more epochs conducted to better predictions. It turns out not to be the case as shown in Tables III and IV where it can be seen that the model at 180 epochs gives better average results than the 200 epochs model. It also seems to be the case of the differences in the predictions between models trained with different number of epochs. As an example, Figure 3 shows that the validation loss tends to jump in certain range, not varying much from the mean

TABLE II: Evaluation over test subset with U-Net.

Number of epochs	Dice	Jaccard
160	0.675687	0.575864
180	0.676534	**0.581769**
200	0.681308	0.562630

TABLE III: Evaluation over validation subset with R-CNN.

Number of epochs	Dice	Jaccard
160	0.785716	0.780492
180	0.790384	**0.785926**
200	0.776682	0.773677

value, therefore the changes made from epoch to epoch on later training do not really make a significant improvement on the predictions. An example of a good prediction can be seen in Figure 4 where both, ground truth and prediction, are very similar.

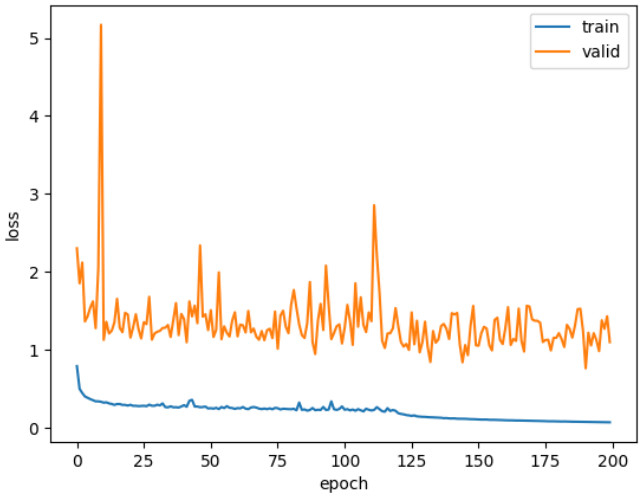
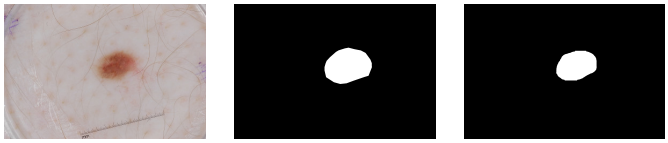


Fig. 3: Network loss during training.

It is important to notice two issues about the network in its current state. First, that the network gives very good predictions when the lesion is small and centered; however it turns out problematic when the lesion is bigger and/or if borders are touching a wide area of the image frame, as can be seen in Figure 5. Second, when the prediction mask is too small or the lesion is not identified, the network cannot handle this situation and crashes.



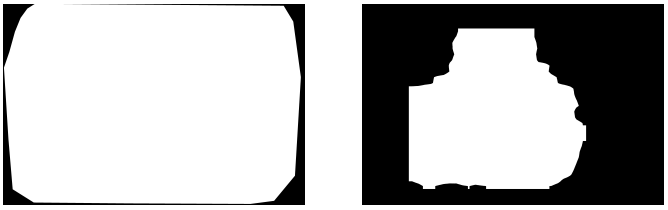
(a) *Input image* (b) *Ground truth* (c) *Prediction*

Fig. 4: Good segmentation example.

Despite these problems and considering the very few changes applied to the network, it performed relatively well. The highest Jaccard score obtained was 0.743 on the test set,

TABLE IV: Evaluation over test subset with R-CNN.

Number of epochs	Dice	Jaccard
160	0.776743	0.739569
180	0.781558	**0.743864**
200	0.781415	0.740744



(a) *Ground truth* (b) *Prediction*

Fig. 5: Bad segmentation example.

close to the best score reported in the ISIC 2017 results, a Jaccard index of 0.76 [14].

V. Conclusions

In this work, an analysis of two deep neural architectures to the segmentation task of skin lesions is presented. Particularly, the U-Net and the Mask-R-CNN were adapted and tested using data of the ISIC challenge.

Results showed that the Mask-R-CNN reaches significant better performance than U-Net when comparing both, the predictions binary masks obtained and the required computation time for training (much lower for Mask-R-CNN).

Future works will be focused on improving the performance coefficients with two ideas: the use of data augmentation techniques and the addition of extra layers to the input image with texture or some other information about the lesion that could help the networks, for instance morphological information [15] and enhancement. These neural architectures could also carry out the classification of the lesion in different type of skin diseases, which is very important in order to generate a complete identification and diagnosis of the lesions. Clearly, this topic will be addressed in future works.

VI. Acknowledgments

The authors gratefully acknowledge NVIDIA Corporation with the donation of the Titan Xp GPU used for this research, and the support of UNL (CAI+D 50020150100055LI, CAID-PIC-50420150100098LI) and ANPCyT (PICT 2016-0651).

References

[1] T. L. Diepgen and V. Mahler, "The epidemiology of skin cancer," *British Journal of Dermatology*, vol. 146, pp. 1–6, 2002.

[2] N. C. F. Codella, D. Gutman, M. E. Celebi, B. Helba, M. A. Marchetti, S. W. Dusza, A. Kalloo, K. Liopyris, N. Mishra, H. Kittler, and A. Halpern, "Skin lesion analysis toward melanoma detection: A challenge at the 2017 international symposium on biomedical imaging (ISBI), hosted by the international skin imaging collaboration (ISIC),"

[3] N. Homedes and A. Ugalde, "Privatización de los servicios de salud: las experiencias de Chile y Costa Rica," *Gaceta Sanitaria*, vol. 16, no. 1, pp. 54–62, 2002.

[4] M. Berseth, "ISIC 2017 skin lesion analysis towards melanoma detection," p. arXiv preprint arXiv:1703.00523, 2017.

[5] X. Zhang, "Melanoma segmentation based on deep learning," *Computer Assisted Surgery*, vol. 22, no. sup1, pp. 267–277, 2017.

[6] X. Yang, Z. Zeng, S. Y. Yeo, C. Tan, H. L. Tey, and Y. Su, "A novel multi-task deep learning model for skin lesion segmentation and classification," *arXiv preprint arXiv:1703.01025*, 2017.

[7] L. Bi, J. Kim, E. Ahn, A. Kumar, M. Fulham, and D. Feng, "Dermoscopic image segmentation via multistage fully convolutional networks," *IEEE Transactions on Biomedical Engineering*, vol. 64, no. 9, pp. 2065–2074, 2017.

[8] M. Attia, M. Hossny, S. Nahavandi, and A. Yazdabadi, "Skin melanoma segmentation using recurrent and convolutional neural networks," in *2017 IEEE 14th International Symposium on Biomedical Imaging (ISBI 2017)*, pp. 292–296, IEEE.

[9] L. Bi, J. Kim, E. Ahn, and D. Feng, "Automatic skin lesion analysis using large-scale dermoscopy images and deep residual networks," *arXiv preprint arXiv:1703.04197*, 2017.

[10] Y. Yuan, "Automatic skin lesion segmentation with fully convolutional-deconvolutional networks," *arXiv preprint arXiv:1703.05165*, 2017.

[11] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," in *Proceedings of the IEEE international conference on computer vision*, pp. 2961–2969, 2017.

[12] P. Tschandl, C. Rosendahl, and H. Kittler, "The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions," *Scientific data*, vol. 5, p. 180161, 2018.

[13] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *European conference on computer vision*, pp. 740–755, Springer, 2014.

[14] N. C. Codella, D. Gutman, M. E. Celebi, B. Helba, M. A. Marchetti, S. W. Dusza, A. Kalloo, K. Liopyris, N. Mishra, H. Kittler, *et al.*, "Skin lesion analysis toward melanoma detection: A challenge at the 2017 international symposium on biomedical imaging (isbi), hosted by the international skin imaging collaboration (isic)," in *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*, pp. 168–172, IEEE, 2018.

[15] E. Decencière, S. Velasco-Forero, F. Min, J. Chen, H. Burdin, G. Gauthier, B. Lay, T. Bornschloegl, and T. Baldeweck, "Dealing with topological information within a fully convolutional neural network," in *International Conference on Advanced Concepts for Intelligent Vision Systems*, pp. 462–471, Springer, 2018.