# Skin Lesion Segmentation and Classification – UnB entry at the ISIC Challenge

Frederico Guth*, Teófilo E. de Campos †
Departamento de Ciência da Computação, Universidade de Brasília (UNB)
*fredguth@fredguth.com, †t.decampos@st-annes.oxon.org

***Abstract*—This paper describes our approach for the ISIC Challenge 2018 - Skin Lesion Analysis Towards Melanoma Detection. For Part 1 - Lesion Segmentation, we developed a U-Net based convolutional neural network pretrained with the ImageNet dataset and applied several data augmentation and hyperparameters optimization strategies, obtaining threshold Jaccard index of 77.5%. For Part 3 - Lesion Classification, we developed an ensemble strategy that achieved an online score of 84.6%.**

## I. Introduction

According to the World Health Organization, between 2 and 3 million non-melanoma skin cancers and 132,000 melanoma skin cancers occur globally each year [1]. Despite representing less than 6.5% of all skin cancers, melanomas are the most dangerous type, accounting for aproximately 75% of all skin cancer related deaths [1], [2]. Early detection is critical to increase survival expectancy and visual inspection still is the most common diagnostic technique.

Deep convolutional neural networks (CNNs) already exceed human performance in visual classification [3]. In some areas of oncology, such as histological image analysis, CNNs have also proven to match the performance of experts, e.g. [4]. In an attempt to improve the scalability of diagnostic expertise, CNNs have been developed to locate and classify skin cancers in images with dermatologist-level accuracy [2].

Dermoscopy is a technique for examination of skin lesions that, with proper training, increase dianostic accuracy from 60% (unaided expert visual inspection) to 75%-84% [5]. The International Skin Imaging Collaboration (ISIC) has a large-scale publicly acessible dataset of more than 20,000 dermoscopy images and host an annual benchmark challenge on dermoscopic image analysis since 2016. The challenge comprise 3 tasks of lesion analysis: Part 1 - Segmentation, Part 2 - Dermoscopic feature extraction, Part 3 - Classification.

In this paper, we describe our approach for the ISIC Challenge 2018:

- **Section II** describes our methodology for **lesion segmentation**;
- **Section III** describes our methodology for **lesion classification**;

## II. Lesion Segmentation

### A. Segmentation Model: Unet34

Introduced in 2015, U-Net is an encoder-decoder architecture designed for biomedical image segmentation [6], with great results in image segmentation problems [7]. In an U-Net, the output is an image with the same dimension of the input, but with one channel. The encoder path is a typical CNN, where each downsampling step doubles the number of feature channels. But what makes this architecture unique is the decoder path, where each upsampling step input is a concatenation of the output of the previous step with the output of the corresponding (same height) downsampling step. This strategy enables precise localization with a very simple network.

Resnet is a very successful architecture in several computer vision tasks [8]. It mitigates the degradation problem that happens when very deep networks starts converging. Instead of learning a direct mapping $H(x) = y$, it learns the residual function $F(x) = H(x) - x$, which can be reframed into $H(x) = F(x) + x = y$, where $F(x)$ is a stack of non-linear layers and $x$ is the identity function(input=output). The formulation of $F(x) + x$ can be implemented by feedforward neural networks with "shortcut connections". Resnet34, specifically, is composed of an initial convolutional layer, 16 blocks of 2 layers and a final fully connected layer.

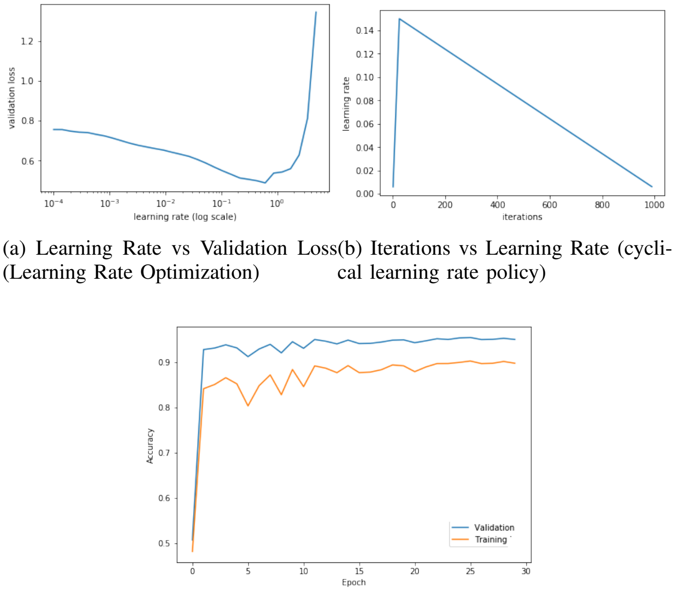
Unet34 is the idea of using a pretrained Resnet34 model as an U-Net encoder path [9]. First, every step from the adaptive pooling onwards is removed, keeping only Resnet backbone. Then we save the output of results of the initial layer, 3rd , 8th, and 14th blocks (of 16 in total). During the upsampling we concatenate the output of those with the ouputs of upsampling steps. We used Adam optimizer and Binary Cross Entropy with Logits as the loss function.

### B. Data and Augmentation

We used "ISIC 2018: Skin Lesion Analysis Towards Melanoma Detection" grand challenge datasets [10], [11] and no aditional external data. All images were first resized to 128x128 pixels, 256x256 and 512x512; and preprocessed to adjust color balance. Random transformations on input images to agument the dataset were made: dihedral transformation, rotation (up to 44 degrees), zooming (up to 1.05), fliping and random lightining changes. The official training dataset was then splited in 3-folds of training and validation datasets.

### C. Segmentation Experiments and Training

Our training strategy was to first train the model with 128x128 images and transfer this learning to train the same model with images with 256 x 256 images. We would suggest



(a) Learning Rate vs Validation Loss (Learning Rate Optimization)

(b) Iterations vs Learning Rate (cyclical learning rate policy)

(c) Epoch vs Accuracy (superconvergence)

Figure 1: Learning rate optimization strategy

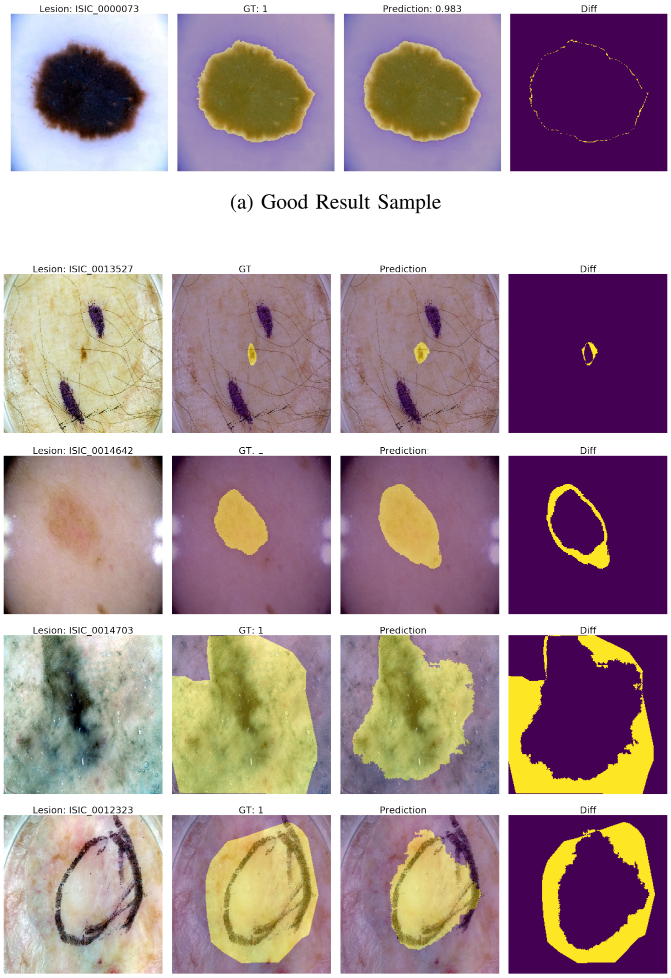
using the same strategy to go from 256x256 to 512 x 512 images, but due to GPU memory constraints, we did not manage to acccomplish this last step.

The training procedure was the same on 128x128 and 256x256:

1) Freeze the first layer group.
2) Define the optimal learning rate with the method proposed by [12] and implemented by [9], where one batch is trained with different learning rates, starting at very low and linearly increasing it at every iteration and ploting a chart of the learning rate versus loss (see figure 1a).
3) We use the 1 cyclical learning rate policy (figure 1b), also proposed by [12], to obtain training convergence in only 30 epochs (superconvergence).
4) Unfreeze the model, keeping only the batch normalization layers frozen, and repeat steps 2 and 3.

It would be advisable to use a loss function more similar to the avaliation criteria. As jaccard is not differentiable, one could use a soft jaccard variation [7]. However, in our priliminar trials the implemented soft jaccard did not improve over the Binary Cross Entropy with Logits loss function and we decided to keep the later.

We developed two strategies for Task 1: *BestDice*: This strategy just predicts the input with the model that presented the best dice index on the validation of our splited training dataset. *Ensemble*: We used the 3-folds of our training dataset and trained with BestDice model and ensemble than to give a prediction.



(a) Good Result Sample

(b) Bad Result Samples

Figure 2: Qualitative assessment of segmentation result

### D. Segmentation Results

The best result on our validation set was obtained using the Ensemble strategy. It achieves a 85.39% Jaccard index and 78.43% Threshold Jaccard index (with cut at 65%). Surprisingly, it did not score so well with the online score and given official validation set, scoring 71.4%. The small size of the official validation set and the threshold at 65% may be the reason for that (one change in prediction from 64% or 66% Jaccard makes a huge difference in the average threshold jaccard index). The BestDice strategy achieved an online score of 75.5% with the official validation set. Anyway, the top ranked participant in 2017 achieved an average Jaccard Index of 76.5%, which should be compared with our 85.39% score.

Visually the best segmentations are almost indentical to the ground truth (see Figure 2a). But we can learn even more from our mistakes. Analysing the worse segmentations there are cases where as non specialists is hard to say if the algorithm was wrong or the ground truth was; there are cases where our algorithm got confused by the pen marker or the glass used by the doctor; and it is clear that in general it doesn't do a good job when the lesion is small relative to the overall image.

## III. Lesion Classification

### A. Classification Model

We used Resnet50, an architecture which has been very succesful in image classification problems. We tried using Resnet101 and Resnext101, but due to the lack of computer resources they didn't suit the fast iteration process our research needed. We used the fastai implementation of Resnet50, which has some improvements over the original model proposed by [13]. We customized the Resnet50 model pretrained with the Imagenet dataset [14], by first, removing all the steps from the adaptive pooling onwards keeping only Resnet backbone as a feature extractor. Then we plugged a very simple network consisting of:

- BatchNorm1d (1024, momentum=0.1)
- Dropout (p=0.25)
- Linear (input=1024, output=512)
- ReLU
- BatchNorm1d (1024, eps=1e-5, momentum=0.1)
- Dropout (p=0.5)
- Linear (input=512, output=7)
- LogSoftmax

We used Adam optimizer and the negative log likelihood loss function.

### B. Data and Augmentation

We used "ISIC 2018: Skin Lesion Analysis Towards Melanoma Detection" grand challenge datasets [10], [11]. When we used a pretrained Task1 model, we considered that external data was used. All images were first resized to 224x224 and a color balance copy of the dataset was made. Random transformations on input images to agument the dataset were made: dihedral transformation, rotation (up to 35 degrees), zooming (up to 1.05), fliping and random lightining changes. The official training dataset was then splited in 3-folds of training and validation datasets.

### C. Classification: Experiments and Training

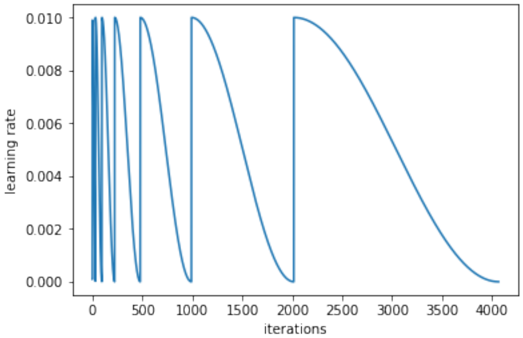


Figure 3: Cosine annealing learning rate strategy.

Priliminar results showed that the color balanced dataset was performing worse than the original and we kept the later.

Table I: Classification Results.

| | Accuracy (%) | Online Score |
|---|---|---|
| SnapEsem | 89.71 | 84.6 |
| CropBest | 93.26 | 76.4 |
| KfoldCrop | 92.31 | 75.4 |

The training procedure was similar to Task 1 (II-C), but we used a cosine annealing learning rate strategy (see figure 3) instead of the one cycle policy (both proposed by [12]). We then tried 3 strategies:

*SnapEsem*: We saved snapshots of our model during training and then made an ensemble of those snapshots, as proposed by [15].

*CropBest*: We applied our segmentation algorithm from Task1 to the dataset of Task3, croping the images by the lesion mininum bounding rectangle. We used a pretrained model from SnapEsem and retrained it using the croped images.

*KFoldCrop*: We repeated CropBest strategy to 3 different folds and ensemble them.

On all strategies we applied Test Time Augmentation (TTA) with 5 augmentations per input to improve results.

### D. Classification Results

Although the best result within our validation set was obtained using the CropBest strategy, we believe that the KfoldCrop can be more general. Interesting to point that all strategies had better accuracy than dermatologists which have an accuracy of 75% to 84% [5].

Better than dermatologist specialists (diagnostic accuracy of 75%-84%).

## IV. Computational Resources

We used Paperspace GPU Cloud service for running all our experiments. The instance server had: 8 cores CPU with 30GB of RAM, Quadro P4000 8 core GPU with 8GB of RAM. The code was developed in PyTorch and Fast.ai [9] on Jupyter Notebooks.

## V. Conclusion and Future Work

This paper descibes our approach for the ISIC Challenge 2018. For the Lesion Segmentation Task, our model Unet34 achieves 77.5% in the competition threshold jaccard index, which is better than previous results in the competition. For the Lesion Classification, we developed an ensemble strategy that achieves an online score of 84.6%, which is on the upper bound of dermatologists accuracy range.

Future work could improve in many ways. For lesion segmentation, we could repeat the experiment using other architectures as feature extractors. Test time augmentation could be applied, it made difference in our classification algorithm and was not implemented due to the difficulties of inversing the augmentations to later average. We suggest to implement a preprocessing pipeline for removing the pen marks in images and maybe a 2 step process can be developed to roughly finding where the lesion is and croping a part of the input image to prevent the bad results due to lesion size. For lesion classification, one could different resnet variations as backbone (resnet18, resnet101, resnext101, etc) and/or using inception. We didn't have time to use the snapshots of the croped images and just used the final model in the kfold, and this may show better results.

## References

[1] B. Stewart, *World cancer report 2014*. Lyon, France Geneva, Switzerland: International Agency for Research on Cancer,WHO Press, World Health Organization, 2014.

[2] A. Esteva, B. Kuprel, R. A. Novoa, J. Ko, S. M. Swetter, H. M. Blau, and S. Thrun, "Dermatologist-level classification of skin cancer with deep neural networks," *Nature*, vol. 542, no. 7639, pp. 115–118, jan 2017. [Online]. Available: https://doi.org/10.1038/nature21056

[3] J. D. L. Fei-Fei, "Where have we been? where are we going?" http://image-net.org/challenges/talks_2017/imagenet_ilsvrc2017_v1.0.pdf, 2017, [Online; accessada 28 de Junho de 2018].

[4] M. Veta, P. J. van Diest, S. M. Willems, H. Wang, A. Madabhushi, A. Cruz-Roa, F. Gonzalez, A. B. Larsen, J. S. Vestergaard, A. B. Dahl, D. C. Ciresan, J. Schmidhuber, A. Giusti, L. M. Gambardella, F. B. Tek, T. Walter, C.-W. Wang, S. Kondo, B. J. Matuszewski, F. Precioso, V. Snell, J. Kittler, T. E. de Campos, A. M. Khan, N. M. Rajpoot, E. Arkoumani, M. M. Lacle, M. A. Viergever, and J. P. Pluim, "Assessment of algorithms for mitosis detection in breast cancer histopathology images," *Medical Image Analysis*, vol. 20, no. 1, pp. 237 – 248, 2015. [Online]. Available: http://dx.doi.org/10.1016/j.media.2014.11.010

[5] N. C. F. Codella, D. Gutman, M. E. Celebi, B. Helba, M. A. Marchetti, S. W. Dusza, A. Kalloo, K. Liopyris, N. K. 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)," *CoRR*, vol. abs/1710.05006, 2017. [Online]. Available: http://arxiv.org/abs/1710.05006

[6] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," *CoRR*, vol. abs/1505.04597, 2015. [Online]. Available: http://arxiv.org/abs/1505.04597

[7] V. Iglovikov, S. Mushinskiy, and V. Osin, "Satellite imagery feature detection using deep convolutional neural network: A kaggle competition," *CoRR*, vol. abs/1706.06169, 2017. [Online]. Available: http://arxiv.org/abs/1706.06169

[8] S. Xie, R. B. Girshick, P. Dollár, Z. Tu, and K. He, "Aggregated residual transformations for deep neural networks," in *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, 2017, pp. 5987–5995. [Online]. Available: https://doi.org/10.1109/CVPR.2017.634

[9] J. Howard *et al.*, "fastai," https://github.com/fastai/fastai, 2018.

[10] N. C. F. Codella, D. Gutman, M. E. Celebi, B. Helba, M. A. Marchetti, S. W. Dusza, A. Kalloo, K. Liopyris, N. K. 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)," *CoRR*, vol. abs/1710.05006, 2017. [Online]. Available: http://arxiv.org/abs/1710.05006

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

[12] L. N. Smith, "A disciplined approach to neural network hyper-parameters: Part 1 - learning rate, batch size, momentum, and weight decay," *CoRR*, vol. abs/1803.09820, 2018. [Online]. Available: http://arxiv.org/abs/1803.09820

[13] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *CoRR*, vol. abs/1512.03385, 2015. [Online]. Available: http://arxiv.org/abs/1512.03385

[14] M. Guillaumin and V. Ferrari, "Large-scale knowledge transfer for object localization in ImageNet," in *2012 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, jun 2012. [Online]. Available: https://doi.org/10.1109/cvpr.2012.6248055

[15] G. Huang, Y. Li, G. Pleiss, Z. Liu, J. E. Hopcroft, and K. Q. Weinberger, "Snapshot ensembles: Train 1, get M for free," *CoRR*, vol. abs/1704.00109, 2017. [Online]. Available: http://arxiv.org/abs/1704.00109