Automating Melanoma Segmentation

Ning (Louis) Li [*]
Department of Computer Science
Stanford University
louisli@stanford.edu

Grace Wang [*]
Department of Computer Science
Stanford University
gracenol@stanford.edu

Makena Low [*]
Department of Electrical Engineering
Stanford University
makenal@stanford.edu

Abstract

Accurate segmentation of skin lesions (i.e. melanoma, vitiligo, etc) is important to maximizing dermatological care. Currently, manual segmentation is required for patient diagnosis, which is labor-intensive and often variable between healthcare professionals. Thus, there exists a need for a quick and accurate automated segmentation tool. Our paper introduces two novel U-Net based models to segment the ISIC 2018 melanoma data set. We also explore transfer learning by using pre-trained VGG16 and ResNet50 weights for the U-Net-based models as encoders. Our best score, a Jaccard Index of 0.826 and a Threshold Jaccard Index of 0.785, suggests that using the ResNet50 weights for the U-Net architecture delivers a strong segmentation result. Based on the 2018 competition results, this score could have placed within the top 10 models in the ISIC 2018 Skin Lesion Analysis Towards Melanoma Detection competition. An Attention U-Net built on a BiggerLeakyU-Net, also achieved similar results.

1 Introduction

Skin cancer is the most prevalent form of cancer in humans, and the deadliest skin cancer is melanoma [13]. Manual segmentation of melanoma lesions is required to diagnose patients. Unfortunately, this segmentation process is time consuming, labor intensive, and gives variable results between healthcare providers. A robust computer-aided diagnostic system to recognize melanoma will help speed up accurate detection and diagnosis tremendously.

In this project, our input was a colored dermoscopic image in JPEG format. We then used U-Net - a popular architecture in image segmentation - and its extension Attention U-Net to output a binary mask prediction image (1 is the existence of a lesion at that pixel, 0 is the absence). Transfer learning techniques were leveraged by using pre-trained weights from VGG16 and Resnet50 as encoders to the U-Net models.

2 Related work

- Hand-built features. Automated skin lesion segmentation research goes as early as the late 1980s. In the early days, research was mostly focused on hand-built features using statistics or math algorithms, such as transforming RGB colors into spherical color space using coordinates,

[*] Equal Contribution

CS230: Deep Learning, Winter 2018, Stanford University, CA. (LateX template borrowed from NIPS 2017.)

transforming principle-components in a user selected color space, or by finding first the average color of a small area of a lesion and the average color of a small area of the background interactively. Though hand-build features made some progress in lesion segmentation, this process was time-intensive and not efficient for larger datasets. [1][2][3]

- Deep Learning. Until recently, with the emergence of deep learning algorithms, increasing state of the art research has been done in lesion segmentation with larger datasets. Yet skin lesion segmentation remained a challenging task. Issues such as low contrast between the lesion area and regular skin, the irregular and fuzzy borders, including hair in the area obstructing the view, and other factors make accurate segmentation much harder. The ISIC competition started in 2016 and is attracting more models that leverage a variety of robust algorithms like LinkNet152 with HAM10000, SLSDeep Ensemble with Dual Attention Network, U-Net, Attention U-Net, and Mask RCNN. Based on this research, our team focused on U-Net-based models and transfer learning, which seemed to give good results in the literature [4][5].

3 Dataset and Features

Our data was extracted from the "ISIC 2018: Skin Lesion Analysis Towards Melanoma Detection" grand challenge datasets, with 2594 samples in total [9][10]. Each sample consisted of a dermoscopic lesion RGB image with a labeled ground truth image (binary black and white masks). Original image sizes ranged from 600 x 600 to 5000 x 4000 pixels.

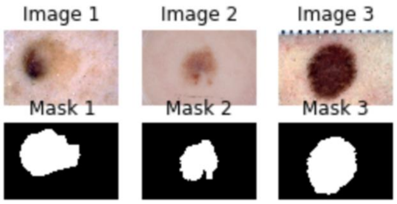


Figure 1: Data Samples

Due to the small dataset size, the data was split into 60/20/20 (1555/519/520) for training, validation and testing, respectively. All images were reshaped to 256 x 256 resolution. The RGB images were then converted to grayscale. Then, the images were normalized by subtracting the mean and dividing by the standard deviation of all channels for each pixel. Keras data augmentation was used to flip, rotate, zoom, shear and shift the original images, as augmentation is an essential component of U-Net. Training the model without augmentation led to overfitting wit the U-Net.

4 Methods

We started with U-Net, due to its viability with our problem statement. U-Net is a network and training strategy that relies on the strong use of data augmentation to use the available annotated samples more efficiently. The architecture consists of a contracting path to capture context and a symmetric expanding path that enables precise localization. This network can be trained end-to-end from very few images. Therefore it is perfect to be used in medical projects, which commonly have limited datasets [4].

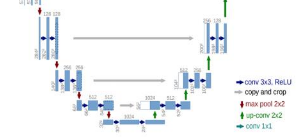


Figure 2: Architecture of U-Net.

We then explored the Attention U-Net. The attention model traditionally has been used heavily in RNN, in which it helps neural networks make choices about which features are more useful for a given task [5]. It was an innovative idea to apply the attention model in a CNN architecture by adding the Attention Gate architecture to the existing U-Net. In fact, this model turned out to be one of the top performing models for image segmentation.

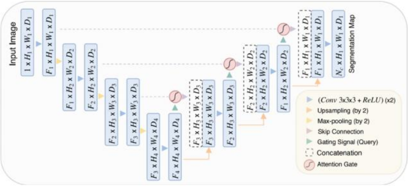

Input Image
$1 \times H_1 \times W_1 \times D_1$
$F_1 \times H_1 \times W_1 \times D_1$
$F_1 \times H_2 \times W_2 \times D_2$
$F_2 \times H_2 \times W_2 \times D_2$
$F_2 \times H_3 \times W_3 \times D_3$
$F_3 \times H_3 \times W_3 \times D_3$
$F_3 \times H_4 \times W_4 \times D_4$
$F_4 \times H_4 \times W_4 \times D_4$
$F_3 \times H_3 \times W_3 \times D_3$
$F_3 \times H_3 \times W_3 \times D_3$
$F_2 \times H_3 \times W_3 \times D_3$
$F_2 \times H_2 \times W_2 \times D_2$
$F_2 \times H_2 \times W_2 \times D_2$
$F_1 \times H_2 \times W_2 \times D_2$
$F_1 \times H_1 \times W_1 \times D_1$
$F_1 \times H_1 \times W_1 \times D_1$
$N_c \times H_1 \times W_1 \times D_1$
Segmentation Map
(Conv 3x3x3 + ReLU) (x2)
Upsampling (by 2)
Max-pooling (by 2)
Skip Connection
Gating Signal (Query)
Concatenation
Attention Gate

Figure 3: Attention U-Net architecture

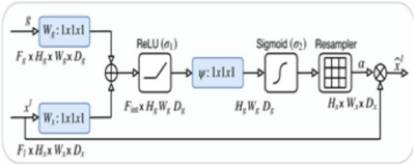

g
$W_g: 1x1x1$
$F_g \times H_g \times W_g \times D_g$
x^l
$W_x: 1x1x1$
$F_l \times H_x \times W_x \times D_x$
ReLU (σ_1)
$F_{int} \times H_g W_g D_g$
$\psi: 1x1x1$
$H_g W_g D_g$
Sigmoid (σ_2)
Resampler
$H_x \times W_x \times D_x$
α
$\hat{x}^l$

Figure 4: Attention gate architecture

After getting the baseline with original U-Net architecture, a number of fine tunings have been performed to achieve better performance. The team also implemented U-Net architecture combined with pre-trained weight from existing VGG16 and ResNet50 models.
Performance is measured by the pixel-wise Jaccard Index:

$$J(A,B) = \frac{A \cap B}{A \cup B} = \frac{A \cap B}{|A| + |B| - |A \cap B|}$$

The Jaccard Index was chosen because it was the most heavily weighted metric in the ISIC competition. We chose not to use the Threshold Jaccard Index as our main metric for easier error analysis.

5 Experiments/Results/Discussion

5.1 Hyperparameter Tuning

- Learning rate (LR): We started with a learning rate of 1e-4. The loss stopped improving after 21 epochs. We hypothesized that perhaps the LR caused the model to bounce around the minima without converging, so we lowered our LR to 1e-5. With this change, training continued to show improvement for 200 epochs. Two slightly higher LRs were also tested with the basic U-Net model (3e-5 and 5e-5), but there was not much improvement in performance or training time.

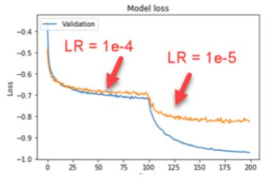


Figure 5: Resnet loss curve with different LR

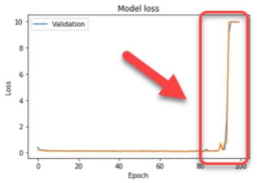


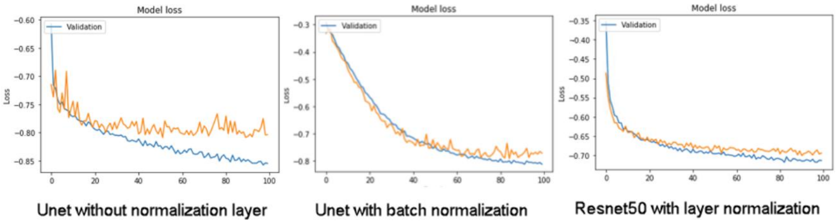
Figure 6: Stability issue with Focal Loss function

- Activation functions. ReLU is used in the original implementation of U-Net. We tried to replace it with LeakyReLU, as it is used in other segmentation papers [11][12]. This change improved validation metrics by 3 percent in our early tests.
- Loss functions: Three loss functions common to segmentation were experimented with: Binary Cross Entropy(BCE), Negative Jaccard Index (NJI) and Focal Loss [14]. The negative Jaccard Index is simply the negative value of the Jaccard Index. The focal loss is a modification of BCE that focuses more on low confidence or incorrect predictions, with the formula

$$FL(p_t) = -(1 - p_t)^\gamma log(p_t)$$

  Focal loss function was not used in any of the final networks as we were unable to stabilize it - loss could be a very large number during the training as shown above in Figure 6.

- Layer Normalization: We tested without layer normalization, batch normalization, and layer normalization. Our team implemented layer normalization from scratch, as it does not appear to be built in to Keras. Batch norm appeared to slow down the training in U-Net network as shown in Figure 7. However, it significantly improved training performance in Resnet + U-Net network.



Unet without normalization layer

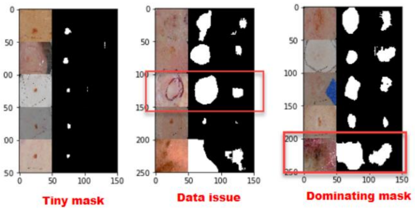
Unet with batch normalization

Resnet50 with layer normalization

Figure 7: Normalization

5.2 Input sizes, Resolution, and Error Analysis

After the first few round of training, we visualized the lowest scored predictions and found the low scores landed in three categories (Tiny Mask, Mislabeled Ground Truth, and Dominating Ground Truth Mask) :



Tiny mask

Data issue

Dominating mask

Figure 8: Error Analysis

The 10 lowest scores were due to tiny masks. To improve, we explored a higher resolution of 512 x 512 images as well as RGB (3 color channels instead of one). The result was greatly improved - the tiny mask issue disappeared after using RGB images. A higher resolution of images did not significantly improve performance.

5.3 Models/Results

We implemented several more networks besides the base models. We chose to use VGG16 and ResNet due to their success as pretrained encoders with semantic segmentation [6][7]. All code was written by the team.

- BigLeakyU-Net - Original U-Net + LeakyReLU + one more layer on the both encoder side and the decoder side.
- BiggerLeakyU-Net - BigLeakyU-Net + one more layer on each side.
- BiggerLeakyU-Net with Attention U-Net - The original Attention U-Net model had overfitting issues even with aggressive dropout. However, it performed very well using BiggerLeakyU-Net.
- Transfer learning with VGG16 as the encoder - Using VGG16 network and pre-trained weights as encoder for U-Net, adding up-sampling layers for masking generating.
- Transfer learning with ResNet50 as the encoder - Unlike VGG16, Resnet doesn't have a conv layer at top level. We added a trainable conv layer (yellow box in Fig 9) to be able to concatenate with up-sampling layer. In the final model, we replaced our trainable top conv blocks with pretrained VGG16 top conv blocks.

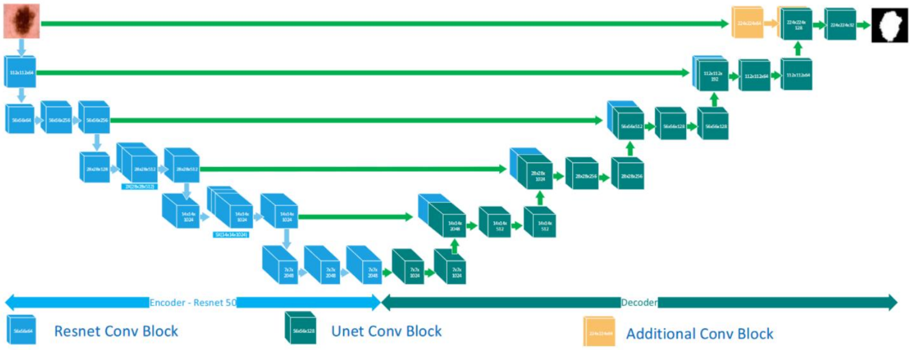
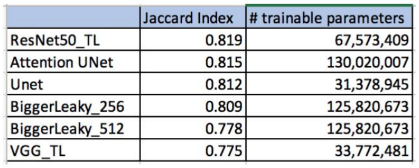


Figure 9: U-Net with ResNet as encoder

Our best result on the test dataset achieved 0.826 in Jaccard Index and 0.785 in Threshold Jaccard Index (threshold = 0.65), using the U-Net based model and 256x256 resolution of RGB images. This result put us in Top 10 in ISIC 2018 leaderboard. Attention U-Net, built on top of the BiggerLeakyU-Net, also achieved comparable results. The hyperparameters used in these models were

- U-Net/Attention U-Net: LR =1e-5, batch size=2, input size=256 x 256
- VGG16/Resnet transfer learning: LR=5r-5, batch size=6, input size=224 x 224



	Jaccard Index	# trainable parameters
ResNet50_TL	0.819	67,573,409
Attention UNet	0.815	130,020,007
Unet	0.812	31,378,945
BiggerLeaky_256	0.809	125,820,673
BiggerLeaky_512	0.778	125,820,673
VGG_TL	0.775	33,772,481

Figure 10: Validation Result

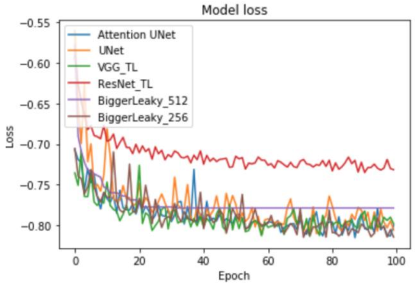


Figure 11: Validation Loss

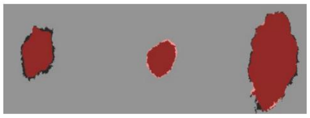


Figure 12: U-Net+ResNet50 Qualitative Overlay Visualization. Black - ground truth; Light pink - prediction; Dark red - overlap between ground truth and prediction.

5.4 Insights/Discussion

With a smaller amount of trainable parameters and the same level of performance, ResNet+U-Net was able to achieve the highest validation score. With more training and fine tuning, we are optimistic about improving performance. We also compared BatchNorm and LayerNorm and witnessed how they improved the performance in the VGG16 and Resnet based models, though they also increased training time. Notably, dropout did not regularize the original Attention U-Net's overfitting issue.

6 Conclusion/Future Work

Our modified U-Net and Attention U-Net achieved satisfying results compared to the ISIC leaderboard scores. The Resnet50 likely performed better than the VGG16 because of it's increased number of layers. With more time, we would explore stabilizing the Focal Loss function and fine tuning Resnet + U-Net model. We also started testing with another popular image segmentation model - Mask RCNN at the end of this project, and it is showing promising results as we finalize this report.

7 Contributions

- Ning Li: U-Net model building, Transfer learning, hyperparameter tuning, error analysis building, LayerNorm implementation in Keras, documentation and final report.
- Makena Low: Transfer learning, error analysis, documentation and final report.
- Grace Wang: Attention U-Net model tuning, hyperparameter tuning, error analysis, documentation and final report

Special thanks to TA Aarti Bagul for great suggestions and recommendations.

Source code is available @ https://github.com/louis-li/MelanomaSegmentation/

References

[1] S.E. Umbaugh, R.H. Moss, W.V. Stoecker, Automatic color segmentation of images with applications in detection of variegated coloring in skin tumors, IEEE Eng Med. Biol. 8 (1989) 43–52.

[2] A. Green, N. Martin, J. Pfitzner, M. O'Rouke, N. Knight, Computer image analysis in the diagnosis of melanoma, J. American Academy of Dermatology 31 (6) (1994) 958–964

[3] A.W. Kopf, T.G. Salopek, J. Slade, A.A. Marghood, R.S. Bart, Techniques of cutaneous examination for the detection of skin cancer, Cancer Supplement 75 (2) (1994) 684–690.

[4] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional networks for biomedical image segmentation. CoRR, abs/1505.04597, 2015. URL http://arxiv.org/abs/1505.04597.

[5] Ozan Oktay, Jo Schlemper, Loic Le Folgoc, Matthew Lee, Mattias Heinrich, Kazunari Misawa, Kensaku Mori, Steven McDonagh, Nils Y Hammerla, Bernhard Kainz, et al. Attention u-net: learning where to look for the pancreas. MIDL, 2018.

[6] K Simonyan, A Zisserman. Very Deep Convolutional Networks for Large-Scale Image Recognition [J]. Computer Science, 2014.

[7] K. He, X. Zhang, S. Ren, and J. Sun. Deep Residual Learning for Image Recognition. In Conference on Computer Vision and Pattern Recognition, pages 770–778, 2016.

[8] Layer Normalization Jimmy Lei Ba, Jamie Ryan Kiros, Geoffrey E. Hinton https://arxiv.org/abs/1607.06450

[9] Noel C. F. Codella, David Gutman, M. Emre Celebi, Brian Helba, Michael A. Marchetti, Stephen W. Dusza, Aadi Kalloo, Konstantinos Liopyris, Nabin Mishra, Harald Kittler, Allan 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)", 2017; arXiv:1710.05006.

[10] Tschandl, P., Rosendahl, C. and Kittler, H. The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. Sci. Data 5, 180161 doi:10.1038/sdata.2018.161 (2018).

[11] Yang, Dong, et al. "Automatic liver segmentation using an adversarial image-to-image network." International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, Cham, 2017.

[12] Pereira, Sérgio, et al. "Brain tumor segmentation using convolutional neural networks in MRI images." IEEE transactions on medical imaging 35.5 (2016): 1240-1251.

[13] "Skin Cancer | Skin Cancer Facts | Common Skin Cancer Types." American Cancer Society, www.cancer.org/cancer/skin-cancer.html.

[14] Ecoffet, Adrien. "Investigating Focal and Dice Loss for the Kaggle 2018 Data Science Bowl." Becoming Human: Artificial Intelligence Magazine, Becoming Human: Artificial Intelligence Magazine, 6 Mar. 2018, becominghuman.ai/investigating-focal-and-dice-loss-for-the-kaggle-2018-data-science-bowl-65fb9af4f36c.