CS230

Achieving Comparable Performance With Less Parameters in Segmentation of Melanoma Images Using Dense U-Nets

Charles Huang
Department of Bioengineering
Stanford University
chh105@stanford.edu

Gustavo Chau
Department of Bioengineering
Stanford University
gchau@stanford.edu

Abstract

Melanoma is one of the most dangerous skin cancers and the segmentation of skin lesions from dermatoscopic images is normally one of the first steps in its correct diagnosis. In this work, we propose the use of Dense U-nets for performing this task. Through experiments conducted in the dataset of the 2017 ISIC challenge, we show that by using dense U-nets we can achieve comparable performance to ordinary U-nets or the 2017 ISIC winning architecture, while reducing the number of needed parameters by as large as 91%.

1 Introduction

Skin cancer is one of the most common types of cancer, and, although not as common as other types of skin cancer, melanoma presents a higher mortality rate. It is estimated by the National Cancer Institute (NCI) that it represents the 5.3% of all new cancer cases (1). Additionally, there is a decreasing number of dermatologists per capita (2) making the need for automatic methods more imperative. Furthermore, in some areas of developing countries the lack of trained physicians jeopardizes the timed diagnosis of melanoma. Accordingly, there is growing interest in developing automatic methods for correct and well-timed diagnoses of Melanoma. Usually, the first step in this automatic pipeline is the correct segmentation of melanoma lesions, which enables the computation of some geometric and or texture features of the lesion that are helpful for radiologists or that helps classifier systems by removing unnecessary information.

This project centers on the segmentation of melanoma lesions from dermatoscopic images (Images of the skin obtained using a high quality magnifying lens and lighting system) using deep learning techniques. The dataset was obtained from Tasks 1 of the 2017 International Skin Imaging Collaboration (ISIC) challenge (2), where the objective is to predict a binary mask from a dermatoscopic image which contains only the skin lesion. An example of a typical input image and the groundtruth segmentation is shown in Figure2.

2 Related work

Different approaches for segmentation of melanoma or skin lesions have been proposed, which we can categorize into classical and learning-based methods. In the former, existing literature have proposed using different methods such as Fuzzy-logic thresholding (3), level-set segmentation (4), and deformable geometric models (5). As for learning approaches, the top three participants of

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

the ISIC challenge used deep learning based methods. The top result correspondent to a Fully convolutional-deconvolutional (6) that includes ensembling of 6 trained models. The second place (7) used an U-net and an ensemble of 10 models. The third place third (8) used a U-net with residual units and additional training data from another dataset. The reported Dice coefficients for these three works are 0.87, 0.86 and 0.86 (Respectively, Jaccard indices of 0.765, 0.762 and 0.760).

3 Dataset

The dataset was obtained from Task 1 of the 2017 International Skin Imaging Collaboration (ISIC) challenge, which contain approximately 2750 labeled examples consisting of a dermatoscopic image and the corresponding binary mask groundtruth manually delineated by physicians (9). The dataset is already divided into 2000 training examples, 150 development/validation examples, and 600 test examples. We decided to keep this distribution in order to have a common assessment with the previous winner of the competition. Although a 2018 version of the challenge exists, validation and test groundtruths, as well as full results disclosures, had not been yet released at the time the project was started. Accordingly, we decided to use the 2017 version. The input images are of varied resolution and thus were resized to a common size of 192 x 256 pixels along with their respective groundtruth masks. Each image was normalized by subtracting the mean and dividing by its standard deviation. We tried the inclusion of data augmentation including rotation, horizontal translation and zooming but the results were not significantly different from the not-augmented dataset.

4 Methods

We tested 4 architectures for solving the current problem: [i] U-net (10), [ii] the Winner of 2017 which corresponds to a Fully convolutional-deconvolutional network (6), [iii] an Small Dense U-net, and [iv] a larger Dense U-net. [iii] and [iv] are adapted from (11).

4.1 U-net

The ordinary U-net had 4 pooling layers and 2 convolution layers per pooling layer with approximately 7.8 million parameters (as defined in (10)). For initial comparisons we used a U-net with binary crossentropy (CE) loss, however, we noticed better results when using combined CE and dice loss (CE+D), so a combined loss was used for all other comparisons.

$$BCE = -\frac{1}{N}\sum_{m=0}^{M} y_m log(\hat{y}_m) + (1 - y_m) log(1 - \hat{y}_m) \tag{1}$$

$$D = -\frac{1}{N}\sum_{m=0}^{M} \frac{2\sum_{i=0}^{N} y_{m,i} * \hat{y}_{m,i}}{\sum_{i=0}^{N} y_{m,i} + \hat{y}_{m,i}} \tag{2}$$

$$Loss = BCE + D \tag{3}$$

4.2 Fully convolutional-deconvolutional network

The previous competition winner's architecture (6) has upsampling and deconvolution layers on the decoding side of the U-net (as opposed to deconvolution and convolution for all of the other U-nets tested here). As the winner did not provide a public implementation of their model, we followed the guidelines in (6) to the best of our knowledge. This network consists of 26 layers and approximately 5.0 M parameters.

4.3 Dense U-nets

Two additional dense U-net architectures (11) were compared as well. A standard dense U-net is shown in Figure 1). The small dense U-net has 4 pooling layers and 1 dense block per pooling layer in addition to 1 dense block at the center of the U-net. Each dense block then contains alternating 8 convolution layers with filter size (2,2) and 8 convolution layers with filter size (1,1) (i.e. bottleneck layers), all with a growth rate of 2. The larger dense U-net has 4 pooling layers and 1 dense block per

pooling layer. By contrast, each dense block has alternating 24 convolution layers with filter size (2,2) and 24 convolution layers with filter size (1,1) (i.e. bottleneck layers), all with a growth rate of 2. The specific amount of layers and growth rate parameters were heuristically chosen to test a dense U-net with approximately 10x less parameters than an ordinary U-net (in the case of the small U-net) and approximately 2-3x less parameters than a ordinary U-net (in the case of the larger U-net). Growth rate was reduced from 12 (as defined by the original densenet paper (11)) to 2 in order to keep the amount of parameters small while allowing for an enormous amount of depth in the tested dense U-nets (i.e. the small dense U-net has 375 layers and the larger dense U-net has 1095 layers).

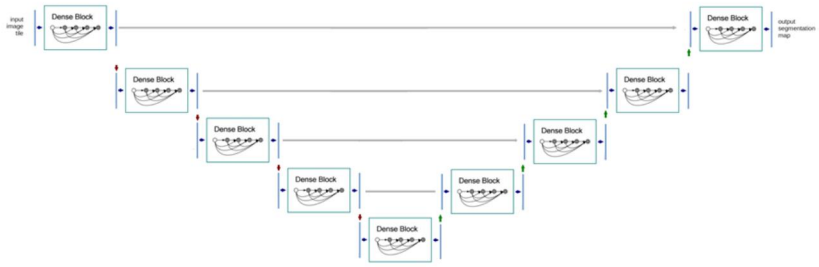


Figure 1: Generic dense U-net where each red downward arrow represents a pooling layer and each grey arrow represents concatenation. Each dense block varies in depth (alternating 8 convolution/8 bottleneck for the small dense U-net and alternating 24 convolution/24 bottleneck for the larger U-net) and has a growth rate of 2.

4.4 Post-processing

We also implemented the post processing algorithm described in (6), which consists on a initial lenient thresholding which allows to select the biggest predicted continuous region and its centroid. A second, more strict threshold is then applied to the original image and after dilation the region contained the centroid is selected. The postprocessing has as main objective conserving only one contiguous region in the predicted mask.

5 Experiments

All networks were programmed in the Keras framework (12) with a TensorFlow backend. All architectures were trained until convergence using Adam optimizer with a learning rate of 0.0001 which was heuristically chosen as we found that when using higher values the training did not converge. A mini-batch size of 16 was chosen as a compromise between computational speed and the GPU memory available. The validation dice was monitored during training and the model with maximum validation dice was used as the final model for testing. The segmentation task for each architecture was evaluated in terms of the Dice coefficient, defined for two set of points A and B as:

$$D(A, B) = \frac{2|A \cap B|}{|A| + |B|}. \tag{4}$$

In this case, A would be the predicted mask and B would be the groundtruth. Due to confidentiality of some sections of the code, instead of uploading the code, we have shared a private *gitlab* repository with the two co-head TAs of the course.

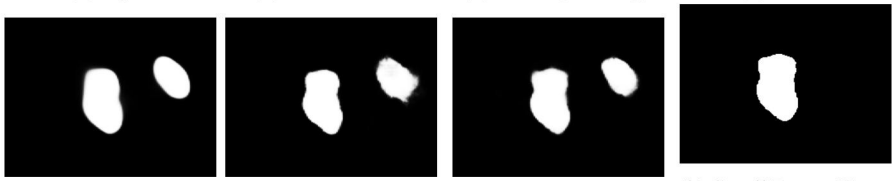
6 Results

The number of parameters and dice coefficient in train, validation and test of the different models are shown in Table 1. We observe that almost all models achieve a comparable dice coefficient of 0.81-0.83. Additionally, the addition of postprocessing provides a 0.01-0.02 increase in dice coefficient in all cases. Example predictions for one of the input test images is shown in Figure 2. It is observed that most predictions are equivalent and that postprocessing helps eliminate confounding regions.

Model	Loss Function	# parameters	Training Dice	Validation Dice	Test Dice	Test Dice (Postprocessing)
ordinary unet	CE	7.8M	0.85	0.90	0.77	0.80
ordinary unet	CE + D	7.8M	0.89	0.91	0.83	0.84
previous competition winner	CE + D	5.0M	0.85	0.92	0.81	0.83
small dense unet	CE + D	0.7M	0.96	0.93	0.79	0.81
dense unet	CE + D	2.7M	0.97	0.93	0.82	0.83

Table 1: Dice coefficient obtained for the different models.

(a) Input (b) Groundtruth (c) U-net (CE only) (d) U-net



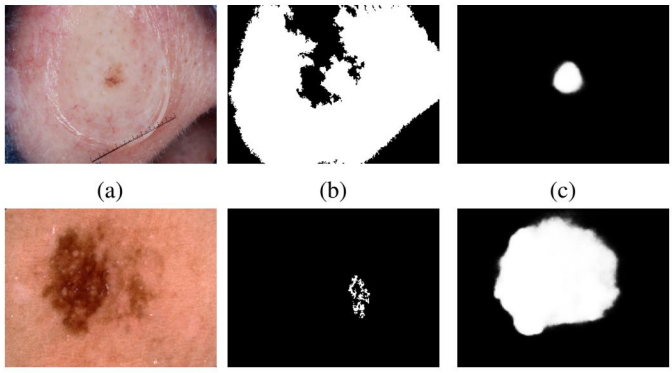
(e) previous winner (f) Dense U-net (g) Small Dense U-net (h) Small Dense U-net (with postprocessing)

Figure 2: Example of Predicted images with the different methods.

7 Discussion

As observed in the two first rows of Table 1, the use of a CE+D loss instead of only CE provides significant enhancements in performance. This can be explained by the fact that CE only compares in terms of a pixel-wise metric while CE+D takes into account the overlap of the full regions. We obtained similar performances in the range of a dice coefficient of 0.81-0.83 using ordinary U-nets, the 2017 winner's network and the dense U-nets, however the use of dense U-nets provide a reduction in the number of needed parameters of as large as 91% with respect to the vanilla U-net. The results also suggest that in most methods, training did not suffer from overfitting as Training and Validation dice coefficients are comparable (except for in the cases of both Dense U-Nets). There is, however, an apparent difference between the validation and test scores in all runs. This apparent difference is likely due to an underlying issue with the ISIC challenge data where validation and test images come from different distributions.

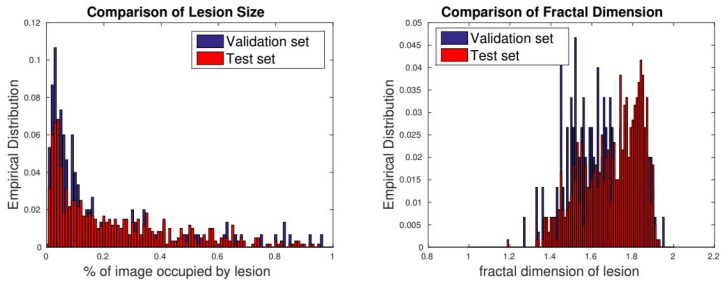
To test this hypothesis, from the groundtruth images, we computed the percentage area of image that the lesion occupies (PA) and the fractal dimension (FD) of the lesion for each of sample in the validation and test sets. The distributions are shown in Figure 4. As observed both the test and validation distributions seem very different. To formally corroborate this, we tested the null hypothesis that both sets come from the same distribution by an unpaired Mann–Whitney U test over the PA and FD values, obtaining, after correcting for multiple test, p-values of 0.00019 and and 0.000016. This indicates that it is very likely that the validation and test sets are indeed coming from different distributions. In Figure 3, we observe some images where all the models fail to predict accurately the lesion due to problems with the complexity of the shape or the abnormal color with respect to surrounding skin.



(a) (b) (c)

(d) (e) (f)

Figure 3: Example of discrepancy in the testing set of images caused by the significantly higher lesion border complexity (top row) or the colored skin being mostly not melanoma tissue (bottom row). Here (a) and (d) are the input, (b) and (e) are the ground truth, and (c) and (f) are the predicted segmentation.



(a) (b)

Figure 4: Distributions of PA and FD of validation and test set samples

8 Conclusion/Future Work

Training performance for both dense U-net architectures was better than both the ordinary U-net and previous winner's U-net, with higher training and comparable validation and test dice scores. The marked drop between validation dice and test dice in all trials and our statistical analysis shows that this particular ISIC challenge dataset has validation and test data that comes from different distributions. Both dense models had fewer parameters than the ordinary U-net and previous winner's U-net, perhaps making them more appropriate choices in a clinical setting where model storage space may be limited.

The segmentation problem posed in the ISIC Challenge may in fact be adequately solved using an ordinary U-net, as the differences in the underlying distributions of the provided validation and test datasets appears to cap any performance gains made in the training and validation dice scores. However, more complex problems might require U-net architectures with greater complexity and depth that can achieve better performance in training. In the case of image translation or reconstruction problems (i.e. reconstructing high resolution MRI images from undersampled K-space data), employing more complex U-net architectures is often necessary to achieve a satisfactory training and validation loss, and using a dense U-net for those purposes may provide substantial performance gains.

9 Contributions

The work in this project has been a collaboration with equal contributions by both team members of the project. The authors would like to acknowledge the code basis provided by the Zacharchuk lab of the RSL deparment. The authors would like to acknowledge and thank Yuan Xie, Yannan Yu, Enhao Gong, and Greg Zaharchuk, especially.

References

[1] N. C. Institute, "Cancer Stat Facts: Melanoma of the Skin," https://seer.cancer.gov/statfacts/html/melan.html, 2018, [Online; accessed 06-Oct-2018].

[2] I. S. I. Collaboration, "ISIC 2018: Skin Lesion Analysis Towards Melanoma Detection ," https://challenge2018.isic-archive.com/, 2018, [Online; accessed 06-Oct-2018].

[3] M. E. Yüksel and M. Borlu, "Accurate segmentation of dermoscopic images by image thresholding based on type-2 fuzzy logic," *IEEE Transactions on Fuzzy Systems*, vol. 17, no. 4, pp. 976–982, 2009.

[4] M. Silveira and J. S. Marques, "Level set segmentation of dermoscopy images," in *2008 5th IEEE International Symposium on Biomedical Imaging: From Nano to Macro*. IEEE, 2008, pp. 173–176.

[5] Z. Ma and J. M. R. Tavares, "A novel approach to segment skin lesions in dermoscopic images based on a deformable model," *IEEE journal of biomedical and health informatics*, vol. 20, no. 2, pp. 615–623, 2016.

[6] Y. Yuan, M. Chao, and Y. Lo, "Automatic skin lesion segmentation with fully convolutional-deconvolutional networks," *CoRR*, vol. abs/1703.05165, 2017. [Online]. Available: http://arxiv.org/abs/1703.05165

[7] M. Berseth, "ISIC 2017 - skin lesion analysis towards melanoma detection," *CoRR*, vol. abs/1703.00523, 2017. [Online]. Available: http://arxiv.org/abs/1703.00523

[8] L. Bi, J. Kim, E. Ahn, and D. Feng, "Automatic skin lesion analysis using large-scale dermoscopy images and deep residual networks," *CoRR*, vol. abs/1703.04197, 2017. [Online]. Available: http://arxiv.org/abs/1703.04197

[9] 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

[10] 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

[11] G. Huang, Z. Liu, and K. Q. Weinberger, "Densely connected convolutional networks," *CoRR*, vol. abs/1608.06993, 2016. [Online]. Available: http://arxiv.org/abs/1608.06993

[12] F. Chollet *et al.*, "Keras," https://keras.io, 2015.