U-Net Ensemble for Skin Lesion Analysis towards Melanoma Detection

Dandi Chen[1,*], Xu Min[1,2,*], Chao Che[1,3], Jingyuan Chou[1], Zhuoran Xiong[1,4], Fei Wang[1,♯]
[1]Department of Healthcare Policy and Research, Weill Cornell Medicine. New York. NY. USA.
[2]Department of Computer Science and Engineering. Tsinghua University. Beijing. China.
[3]Ministry of Education Key Lab of Advanced Design and Intelligent Computing. Dalian University. Dalian. China.
[4]Department of Electrical Engineering. Columbia University. New York. NY. USA.
[*]These authors contribute equally.
[♯]Corresponding Author. Email: few2001@med.cornell.edu

***Abstract*—The International Skin Imaging Collaboration (ISIC) is hosting a challenge to support development of automated melanoma diagnosis algorithms with three tasks. Our team participates lesion segmentation task (task 1) and lesion attribute detection task (task 2). We employ a U-Net ensemble model of three architectures for these two tasks. For lesion segmentation task, the ensemble includes 192 U-Net models with different architectures, data preprocessing methodologies, external data sources, images sizes and loss functions. The lesion attribution detection is treated as five binary classification problems. Each attribute is detected by a classifier integrating 60 U-Net models. To handle the data imbalance problem, we train a classifier to identify whether the image has any attributes before attribute detection. The experimental results on validation set confirm the effectiveness of our ensemble framework. We can achieve 0.820 as threshold Jaccard index in task 1 and 0.432 as Jaccard index in task 2 on provided validation set.**

I. Introduction

Skin cancer refers to the uncontrolled growth of abnormal skin cells. It can develop on skin exposed to the sun. Sometimes it may form on areas that are not exposed to sunlight. Melanoma is one of the deadliest form, which is response for over 100,000 new cases and over 9,000 deaths each year in the United States.

ISIC has been holding the skin lesion analysis challenge since 2016, where the participants need to develop image analysis tools to enable the automated diagnosis of melanoma. This year our team participate the lesion segmentation task (Task 1) and lesion attribute detection (Task 2). Since U-Net has been demonstrated to be effective in medical image segmentation, we construct an ensemble framework integrating different U-Net model variants in both tasks. We test three types of ensemble schemes, including unweighted average, weighted average and hierarchical average. Through experiment, we select 144 best models to integrate for the lesion segmentation task. For attribute detection task, we employ a hierarchical framework consisting two levels of classification. The level-1 classifier identifies images containing attributes, which are fed to the level-2 classifier to detect the specific dermoscopic attributes.

II. Methodology

The details of our methodology are introduced in this section.

A. The framework for lesion segmentation and attribute detection

The overall flow of our ensemble framework for lesion segmentation is illustrated as Figure 1. We integrate the segmentation results of different U-Net models based on three architectures using the different ensemble strategies.

- *Unweighted Average*. Each U-Net model will output a probability map for every skin image, with each pixel associated with a probability indicating how likely this pixel belongs to the lesion. The unweighted average scheme constructs a final probability map by simply taking the arithmetic average of all the base probability maps obtained from those U-Nets. Finally the probability map will be thresholded to get the segmentation results. This is an unsupervised ensemble approach.
- *Weighted Average*. For this scheme, we assume the final aggregated probability map for each image is a convexly weighted combination of the base probability maps, where all combination coefficients are nonnegative and the sum of all them equals one. The coefficients are learned through minimizing the cross entropy on a training set. This is a supervised ensemble approach.
- *Hierarchical Average*. During our experiments, we observe that complementary predictions can be made by several U-Net variants trained with three different loss functions - dice coefficient, Jaccard index and binary cross entropy. The hierarchical average scheme is taking two levels of weighted average on the base probability maps. The first level is obtained within each group of probability maps obtained from the U-Net models minimizing the same loss. The second level is the weighted average of the three probability maps obtained in the first level. All weighting coefficients are learned by minimizing the total cross entropy on the training images.

The goal of task 2 is to detect five attributes including pigment network, negative network, streaks, milia-like cysts and

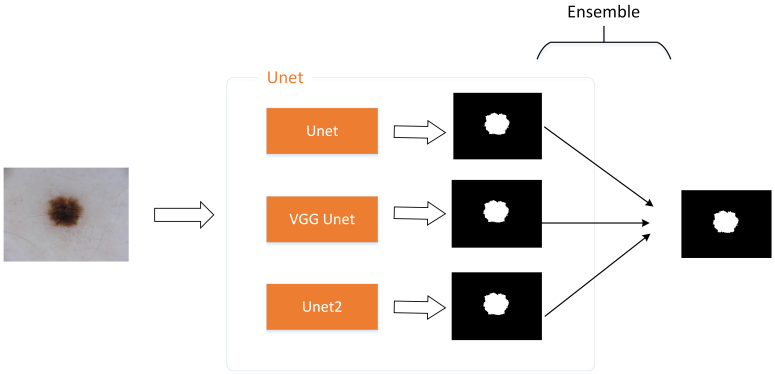

Fig. 1. The overall flow of our ensemble framework for lesion segmentation.

TABLE I
NUMBER OF SKIN IMAGES WITH DIFFERENT PATTERNS

Attribute	Number of Images
Pigment network	1518
Negative network	190
Streaks	97
Globules	603
Milia-like cyst	608
None	514

globules. This can be regarded as a 5-class classification problem. We solve this problem by the one-vs-rest scheme, where we construct five binary classifiers, one for each attribute (which are called attribute discriminators). During experiments we found that there are certain number of skin images that are without any of the five attributes. In order to account for this situation, we build another layer of classification on top of the attribute discriminators, whose goal is just to judge whether an image contain any of the five attributes (which are called attribute detectors). The overall ensemble framework for task 2 is demonstrated in Figure 2.

During the development of our methodologies, we also considered the relationship between task 1 and task 2. Basically the detected attributes in task 2 are supposed to fall in the lesion region identified in task 1. Therefore, we extracted the bounding boxes of the detected lesion region from task 1, which are then used as input image of task 2 instead of the original. This can greatly reduce the imbalance between the foreground and background pixels. The output images are finally restored to the original size for submission.

Table 1 summarizes the counts of images with different attributes (or without any attributes) in the training data. From the table we can observe that the distribution of the images with different attributes are highly imbalanced, and 514 images do not have any of the attributes. This also validates the rationale on the need of the attribute detectors.

B. U-Net Variants

U-Net [1] is a popular deep learning architecture for biomedical images segmentation, which can achieve fairly good performance with limited number of training samples. Our basic U-Net model is implemented based on the code[1] from Menegola *et al.* [2]. We constructed 192 U-Net models for lesion segmentation task and 60 U-Net models for lesion attribute detection task by different parameter settings. Those models are the different version of U-Net with three architectures varying in data preprocessing, external data, data resizing and loss function. The details are as follows.

1) Architecture: U-Nets we adopted are all based on Ronneberger, et al. [1]. The basic U-Net consists of a contracting path and an expansive path. The two paths have similar but different architectures and they are symmetric overall, which yields a U-shaped architecture. The contracting path follows the typical architecture of a convolutional network. The expansive path have a large number of feature channels in upsampling layers, which can propagate context information to higher resolution layers. The network does not have any fully connected layers and only uses the valid part of each convolution.

Three U-Net variants we adopted are U-Net, U-Net2 and VGG-UNet.

- *U-Net*. Similar as basic U-Net[1], encoder-decoder architecture is applied in [2]. The encoder has three convolution blocks, and each convolution block contains three convolution layers and one max-pooling layer. The decoder contains three upsampling convolution blocks with bypass from the encoder, which are used to transform the encoded feature to an image mask with same shape as input. And a fully-connected layer is used to connect the encoder and decoder.
  Sigmod activation is adopted in the last layer to output the probability map and ReLu activation is applied to all other layers in our implementation.
- *U-Net2*. The architecture of U-Net2 is very similar to U-Net, except that a Batch Normalization layer after each upsampling layer in the decoder.
- *VGG-UNet*. VGG-UNet is the combination of VGG[3] and U-Net. The first part of the model consists of VGG-16 layers removing the last max-pooling layer and fully-

[1]https://github.com/learningtitans/isbi2017-part1

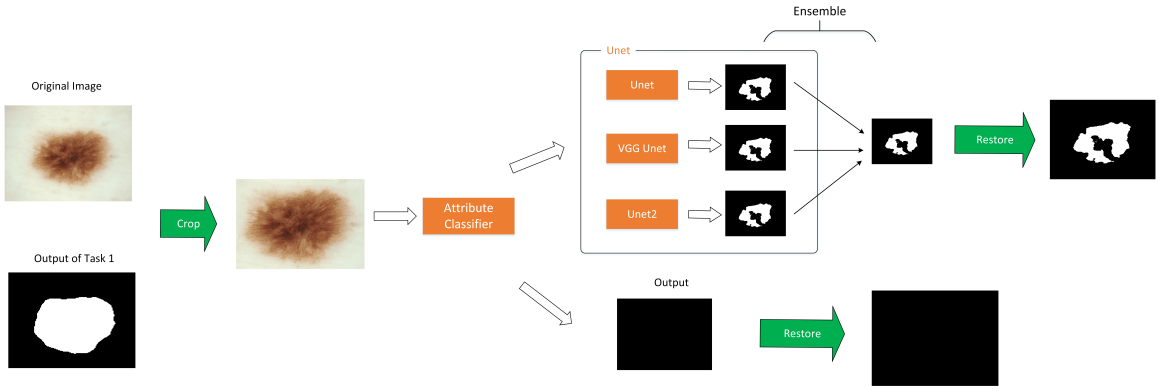

Fig. 2. The ensemble framework for lesion attribute detection.

connected layers. Then VGG-16 layers are concatenated with a basic U-Net model consisting of convolutional and upsampling layers. We added dropout layers after each of the five layers with copied features. Pre-trained weights from ImageNet dataset are used for the initialization of VGG-16 layers. Hence, in the initial stage of the training, VGG-16 layers are kept frozen and only the parameters of U-Net layers are updated. After 100 epochs, all the layers including VGG-16 layers are updated.

2) Data Preprocessing: Before training, all the images are preprocessed with standard normalization and contrast-limited adaptive histogram equalization(CLAHE). We adopt the real-time image augmentation procedure integrated in Keras. All the images are transformed by different augmentation operations, including rotation in a rang up to 270 degree, shifting horizontally and vertically up to 10% of width or height and zooming up to 10% size.

We also performed data augmentation on the testing set. After transforming each testing image into several copies using data augmentation, we get their corresponding predictions via trained model. Then inversely transformations are applied for the final prediction of single testing image.

3) External Data: We downloaded external dataset from ISIC archive[2] and PH2 dataset[4] for task 1, while no external data was used in task 2. Since multiple segmentations are provided in ISIC archive as groundtruth for several images, we filter them out and 11,516 images from ISIC archive are obtained finally for training.

4) Data Resizing: Images were resized to 128×128, 256×256 and 512×512 when we train U-Net variants. We observed that 128×128 size performed better in task 1 and 256×256 size performed better in task 2. Thus, 144 U-Net models trained with 128×128 resized images and 12 U-Net models trained with 256×256 images are kept for final ensemble.

5) Loss Function: We adopt four loss functions when training U-Net variants, including binary cross entropy, dice coefficient, Jaccard index and threshold Jaccard index. The last one is also used for final evaluation in task 1. A smooth parameter is added to make dice coefficient and Jaccard index differentiable. The loss function is minimized by ADAM.

[2]https://isic-archive.com/

TABLE II
PERFORMANCE OF DIFFERENT MODELS FOR LESION SEGMENTATION

Architecture	Prepro	External Data	TDA	Loss Function	Jaccard
U-Net2	Yes	Archive	No	Jaccard	0.797
U-Net2	Yes	Archive	Yes	Jaccard	0.795
VGG-UNet	No	No	No	Jaccard	0.794
VGG-UNet	No	No	Yes	Jaccard	0.793
VGG-UNet	No	Archive, PH2	No	Jaccard	0.787
U-Net	No	PH2	No	Jaccard	0.786
VGG-UNet	No	Archive, PH2	Yes	Jaccard	0.786
U-Net	No	PH2	Yes	Jaccard	0.784
U-Net2	Yes	Archive	No	Dice	0.782
U-Net	No	No	No	Jaccard	0.780

III. RESULTS

We randomly choose 10 % of provided training images of ISIC 2018 for detect earlystopping when training U-Net models. 100 provided validation images are used for held-out testing set.

Our U-Net variants were implemented using Keras with Theano backend. Batch size was set to 4 and maximum epochs is set to 220. Learning rate in Adam is set as 1e-5.

Following results are based on validation set provided by ISIC 2018 Challenge.

1) Results of Lesion Segmentation: Total 192 U-Net models were tested in task 1 and 10 best models were selected for ensemble. Performance of 10 best models on 100 validation images are illustrated in Table II. "Prepro" and "TDA" stand for preprocessing and test data augmentation, respectively. "Archive" and "PH2" indicate the model adopted ISIC Archive and PH2 dataset as external data, respectively.

Jaccard index of weighted average ensemble of 10 models in Table II can reach 0.811 although the best single model performance is 0.797. We also noticed that 5 and 20 best single models with weighted average ensemble can reach 0.802 and 0.807.

2) Results of Lesion Attribute Detection: For lesion attribute detection, we first cropped out region of interest from

an input image according to task 1 prediction. Then we analyzed the cropped image using the three-step pipeline. 1) We first classified whether a image contains any lesion attributes or not. This step is necessary since many images has none of the attributes at all. 2) We then classified whether a image contain one of the five attributes. Here we train five classifiers separately. 3) We finally used a U-Net model to locate lesion attribute if it is believed to have this attribute. Again, we need to train five U-Nets for five lesion attributes. We multiplied the probability map obtained in step three with the two probabilities predicted in the first two steps, therefore we need to choose a reasonable cutoff to get the final binary mask. We tried cutoff values from 0.1-0.9 on the unweighted average ensemble, and found that 0.5 can generate the best performance with Jaccard index 0.432. In fact, our model is not quite sensitive to the choice of cutoff, since even if we used cutoff 0.9, we can still achieve Jaccard index of 0.347.

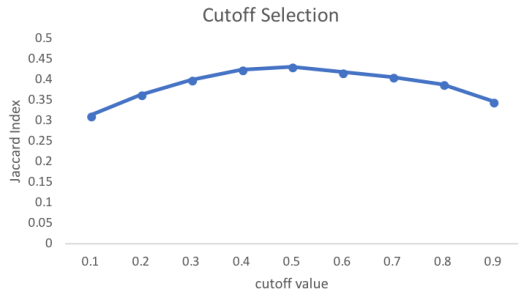

Fig. 3. The overall flow of our ensemble framework for lesion segmentation.

For attribute detection, we did not leverage any external data. We observed that, for task 1, models with input size 128 are generally better than 256, meanwhile models with input size 256 are better for task 2. Hence we finally chose those models with input size 256 for task 2. The performances of VGG-Unet under different settings are reported in Table III. In Table III, "Prepro" and "TDA" have the same meaning as in Table 3. "Column crop" indicates whether we crop the minimum bounding box of the segmentation of task 1 as the input. BCE stands for binary cross-entropy. The Jaccard scores of different ensemble method are shown as Table IV. In Table IV, "Ensemble Dice" indicates the ensemble method that only integrates four VGG-UNet models using dice as loss function. Similar strategy applied for "Ensemble Jaccard" and "Ensemble BCE".

IV. Conclusion

This report describes the ensemble framework that our team used to participate the lesion segmentation and attribute detection task of ISIC 2018. The framework integrates different U-Net models of three architectures using different data preprocessing, external data, data resizing and loss function strategies. Besides U-Net, we have also investigated other deep learning models such as DeepLab and CRF-as-RNN. Traditional machine learning classifiers based on superpixel and salient object detection methods are tried as well. It turns out that U-Net variants have outperformed all other models obviously in the experiments. As a result, we only adopt U-Net models in our ensemble framework.

TABLE III
Performance of different models for attribute detection

Achitecture	Preprocessing	Crop	TDA	Loss function	Jaccard
VGG-UNet	No	Yes	No	BCE	0.343
VGG-UNet	No	Yes	Yes	BCE	0.354
VGG-UNet	Yes	Yes	No	BCE	0.355
VGG-UNet	Yes	Yes	Yes	BCE	0.367
VGG-UNet	No	Yes	No	Dice	0.361
VGG-UNet	No	Yes	Yes	Dice	0.382
VGG-UNet	Yes	Yes	No	Dice	0.361
VGG-UNet	Yes	Yes	Yes	Dice	0.393
VGG-UNet	No	Yes	Yes	Jaccard	0.348
VGG-UNet	No	Yes	Yes	Jaccard	0.375
VGG-UNet	Yes	Yes	No	Jaccard	0.376
VGG-UNet	Yes	Yes	Yes	Jaccard	0.392

TABLE IV
Performance of different ensemble methods for attribute detection

Ensemble Method	Jaccard
Unweighted Average	0.432
Weighted Average	0.428
Hierarchical Average	0.428
Ensemble Dice	0.417
Ensemble Jaccard	0.408
Ensemble BCE	0.394

References

[1] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241.

[2] A. Menegola, J. Tavares, M. Fornaciali, L. T. Li, S. Avila, and E. Valle, "Recod titans at isic challenge 2017," *arXiv preprint arXiv:1703.04819*, 2017.

[3] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[4] T. Mendonça, P. M. Ferreira, J. S. Marques, A. R. Marcal, and J. Rozeira, "Ph 2-a dermoscopic image database for research and benchmarking," in *Engineering in Medicine and Biology Society (EMBC), 2013 35th Annual International Conference of the IEEE*. IEEE, 2013, pp. 5437–5440.