A Visually Explainable Learning System for Skin Lesion Detection Using Multiscale Input with Attention U-Net

Duy Minh Ho Nguyen, Abraham Ezema, Fabrizio Nunnari(✉), and Daniel Sonntag

German Research Center for Artificial Intelligence, Saarbrücken, Germany
{Ho_Minh_Duy.Nguyen,Abraham_Obinwanne.Ezema,
Fabrizio.Nunnari,Daniel.Sonntag}@dfki.de

Abstract. In this work, we propose a new approach to automatically predict the locations of visual dermoscopic attributes for Task 2 of the ISIC 2018 Challenge. Our method is based on the Attention U-Net with multi-scale images as input. We apply a new strategy based on transfer learning, i.e., training the deep network for feature extraction by adapting the weights of the network trained for segmentation. Our tests show that, first, the proposed algorithm is on par or outperforms the best ISIC 2018 architectures (LeHealth and NMN) in the extraction of two visual features. Secondly, it uses only 1/30 of the training parameters; we observed less computation and memory requirements, which are particularly useful for future implementations on mobile devices. Finally, our approach generates visually explainable behaviour with uncertainty estimations to help doctors in diagnosis and treatment decisions.

Keywords: Skin lesion · Diagnose features · Attention U-Net

1 Introduction

Skin cancer is one of the most frequently occurring diseases with more than one million positive diagnoses in the United States each year. The most dangerous type of skin cancer is the melanoma, causing over 9,000 deaths, and 76,380 new cases according to the American Cancer Society per year [12]. While melanoma at an early stage can be treated successfully, it still demands rigorous manual evaluations by the dermatologist for several skin lesion patterns. Hence, partly automatizing skin cancer detection plays an important role in the early diagnosis of skin cancer.

In recent years, the International Skin Imaging Collaboration (ISIC) [4] organizes competitions to seek the best algorithm that can diagnose melanoma automatically. Our work utilized the ISIC challenge data in 2018, which was composed of 3 subtasks. The first task was to segment the lesion and skin boundaries, next was the lesion attribute detection to predict the positions of five skin lesion

U. Schmid et al. (Eds.): KI 2020, LNAI 12325, pp. 313–319, 2020.
https://doi.org/10.1007/978-3-030-58285-2_28

attributes as a negative network, pigment network, milia-like cysts, streaks, and globules [4] (Fig. 2). The last task was the classification of images as melanoma, basal cell carcinoma, melanocytic nevus, actinic keratosis, benign keratosis, vascular lesion, and dermatofibroma. While image classification (Task 3) can be seen as a black box, the segmentation (Task 1) and lesion detection (Task 2) steps give visual feedback of known features that doctors can visually inspect and evaluate. In particular, the lesion attribute detection supports doctors to identify whether a lesion is benign or malignant. These visual features are described as a global distribution spanning over a massive area, or a local distribution in a small area, or multiple spots in the lesion. Therefore, the automatic detection and visualization of skin lesion attributes are critical and can be of tremendous support to doctors when diagnosing melanoma in an early phase while explaining the machine learning decisions.

In this paper, we propose a visually explainable learning system with uncertainty estimations for Task 2 of the ISIC Challenge 2018. Our approach adheres to the mental model of the doctor by leveraging the predictive power of deep learning approaches to reduce the bias of a doctor for lesion classification.

2 Related Works and Our Contribution

Several methods have been proposed for the extraction of features of skin lesions, all based on variants of the convolutional neural networks (CNN) such as XceptionNet [3], ResNet [6], U-Net [11]. However, unlike Task 1 and Task 3, the best performance in feature extraction (Task 2) was very low. The highest score (Jaccard index) was just above 30% compared to over 80% mean accuracy and 88% J-index in Task 1 and Task 3, respectively. The reasons behind are the lack of annotated data, imbalanced datasets, and complex structures with varying appearances per patient. To deal with these issues, most of the approaches utilized the transfer learning strategy and fine-tuned large pre-trained deep learning models; afterwards, stacking the networks into an ensemble to make final predictions. For instance, the second-ranked team (LeHealth) [14] adapted the ResNet architecture for PSPNet [13] to simultaneously predict the positions of five lesion attributes. The best method (NMN) [8] constructed an ensemble network based on five baseline architectures: Densenet169 [7], two versions of ResNet [6], Xception [3], and DeepLab-v3 [2], each predicting the position of a separate attribute. Unlike those works, we propose a new strategy for predicting the five skin lesion attributes based on a single variant of the U-Net called Attention U-Net [9], which has consistently shown to improve the performance of the U-Net architecture across different datasets.

In particular, our new approach differs from previous work in two aspects. First, instead of using existing models such as ResNet or XceptionNet, which were trained on a huge dataset like ImageNet [5] to initialize network parameters, we present a novel approach for training the Attention U-Net based on a transfer learning that first trains on the Task 1 (image boundary segmentation) and then uses the trained weights as initialization for Task 2 (lesion attributes prediction).

This idea is motivated by the key point that most of the lesion structures are located inside the lesion boundary, so initializing weights in this manner can be considered as a step to reduce the impact of the surrounding foreground–thus, the model can converge faster. Besides, by employing only the Attention U-Net architecture, we can downgrade the amount of memory for storing models on each device, which allows us to train the network without the difficulty of finding compatible devices.

Secondly, we utilize multi-scale images as a sequence of inputs rather than a single image, as conventional approaches do. This exploits the intermediate feature representations better. Experimental results on the ISIC 2018 challenge dataset show that our proposed method outperforms the second-ranked team (LeHealth) [14] and attains a close margin with the best team (NMN), thereby producing a much better performance-explainability trade-off that can be evaluated by doctors in future experiments.

3 Method

There are two principal ways to detect lesion attributes. The first one tries to train a network that can predict all five attributes together, while the second type focuses on training separate networks for each type of lesion attribute. In this work, we apply the second strategy for two main reasons. The first reason is to avoid the negative impact of the class imbalance in the dataset, while the second reason is that this approach leads to an uncertainty property whereby a pixel can be assigned to several classes with different probabilities depending on the input images. In those cases, the doctor can examine them carefully and make a final decision by visualizing the corresponding regions.

Datasets: We used two datasets downloaded from the ISIC 2018 challenge website (https://challenge2018.isic-archive.com/). The first dataset (for Task 1) includes 2594 images with corresponding ground-truth mask images for segmentation. A mask image is a 2-color image, black/white, whose resolution matches with the corresponding sample image, where pixels associated to a positive case are marked white. The second dataset (for Task 2) comprises of 2594 images with 12,970 ground-truth masks (one separate mask for each attribute). However, since most of the attributes do not appear together in an image, the corresponding masks are empty (all black) for the absent attributes. Table 1 represents in detail a distribution for each attribute, where the highest-occurring attribute are the pigment network and milia-like cysts with 58.7% and 26.3%, respectively. Streaks is the lowest-occurring attribute, with 2.9%. This imbalance makes the prediction task more complicated, that is a trained network will be severely biased towards the attributes with a lot of training data points as compared to attributes with fewer samples. This challenge motivated our choice to employ a separate model for each attribute prediction.

Network Architecture: The proposed method to predict masks for segmentation and for the five lesion attributes is illustrated in Fig. 1, where the main component is the Attention U-Net [9] with multi-scale images as input.

Table 1. Distribution of mask images

Lesion attributes	Pigment network	Globules	Milia-like cysts	Negative network	Streaks	Total images
Mask count	1522	602	681	189	100	2594
Rate	58.7%	23.2%	26.3%	7.3%	2.9%	100%

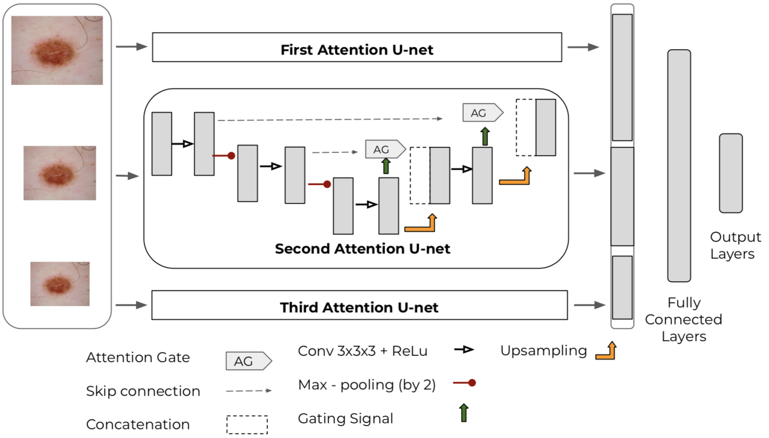

Fig. 1. Our proposed architecture with three blocks of Attention U-net.

The Attention U-Net is a modified version of U-Net [11], which has been proven to be very effective on small datasets. In particular, Attention U-net is equipped with Attention Gates (AG), which are used to recognize relevant spatial information from low-level features and passed to the decoding path. For each input feature map x^L at layer L, AG provides an attention coefficients α to transform the input feature map x^L to an output of semantical features $\hat{x}^L$, defined as: $\hat{x}^L = x^L \odot \alpha$, where $\odot$ denotes the element-wise product operator; α is the attention coefficient to identify salient image regions and prune feature responses to preserve only the activations relevant to the specific task. By leveraging AG, Attention U-Net focuses on target structures without additional supervision, thus enabling us to avoid an external object localization model.

We utilize a sequence of three images with different resolutions 180×180, 256×256, and 450×450 as the input of the three distinct Attention U-Nets. Such inputs are also referred to as the "Pyramid" feature [1], whereby the strength lies in the ability to search objects faster using a coarse-to-fine strategy, thus enabling the network to exploit more information of objects via the multiple resolution levels. Finally, we concatenate the feature vectors from the multi-scale images and pass them to a fully-connected layer before the final prediction layer. Pixel-level output is thresholded at 0.5 for black/white discrimination. The loss function is defined as the complement of the Jaccard index:

$$L = 1 - \frac{\sum y_{truth}\, y_{predict}}{\sum y_{truth}^2 + \sum y_{predict}^2 - \sum y_{truth}\, y_{predict} + \alpha} \tag{1}$$

where $y_{predict}$ and y_{truth} are the predicted pixel vector and its corresponding ground truth, and $\alpha = 1e - 05$ is a smoothing value to avoid divisions by zero.

Transfer Learning from Segmentation Task: While most recent works commonly initialized the parameters of their networks by transfer learning from ImageNet [5], we approach in a novel way through learning directly from the segmentation task. Specifically, we randomly sampled 70% of the total 2594 images of Task 1 as the training set, with the remaining 30% as the held-out set. For each image in the training set, we applied a pre-processing step to center the data by subtracting the mean per channel and constructing multi-scale versions with three corresponding sizes: 180×180, 256×256, and 450×450. At the next step, these images were fed into our architecture, as described in Sect. 3. We trained the proposed framework for 40 epochs with earlystopping. A Jaccard index score of about 76% was obtained on the test set of 780 images–closely matching the baseline results on the leaderboards from ISIC 2018 Task 1[1].

Lesion Attributes Detection: Given the trained segmentation network, we clone it into five new instances, one for each lesion attribute, thus each initialized with the segmentation task parameters. In other words, we model the prediction problem as five independent binary segmentation problems. Besides the advantages of avoiding the data imbalance problem (Table 1) and producing an uncertainty score; by further examination of the data, we discovered that most of the lesion attributes were located near the lesion boundary. Therefore, initializing weights from the segmentation network can be considered as a consequential preprocessing step to lessen the effect of the surrounding foreground. Consequently, the supporting model can predict more precisely the positions of lesion attributes.

4 Experiments and Results

During experiments, we build models with the Keras framework and using the AMSGrad optimisation algorithm [10] with a learning rate and weight decay of approximately 10^{-4}. A five-fold cross-validation scheme was applied for each lesion structure, with 60 epochs for each fold, then we computed the expected performance based on out-of-sample tests on the networks. To be consistent with the standard requirements of the ISIC challenge, we use the Jaccard index as the main score.

Figure 2 shows sample results in the detection of the globules lesion attribute.

We compare our cross-validation results with the two top methods: NMN's method [8] and LeHealth's method (see Table 2). Our average Jaccard index result is 0.278, which is 0.002 more compared to LeHealth's method and 0.029 less than the best approach. Nevertheless, our method surpasses both competitors in

[1] https://challenge2018.isic-archive.com/leaderboards/.

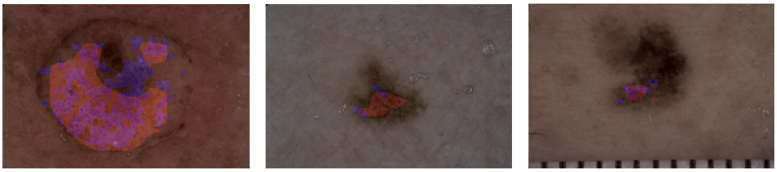

Fig. 2. Results for Globules where the blue regions indicate the ground-truth labels and the red regions indicate our visual predictions/explanations, respectively. (Color figure online)

Table 2. Comparing our results that uses network initialization from the segmentation network against the NMN and LeHealth team based on the Jaccard Index. The best-performing scores are in bold.

Method	Pigment network	Globules	Milia-like cysts	Negative network	Streaks	Average
Our method	0.535	**0.312**	0.162	0.187	**0.197**	0.278
Our method (without transfer)	0.493	0.221	0.145	0.156	0.118	0.227
NMN's method	**0.544**	0.252	**0.165**	**0.285**	0.123	**0.307**
LeHealth's method	0.482	0.239	0.132	0.225	0.145	0.276

two categories out of the five: globules and streaks. Furthermore, the experiment results prove the effectiveness of our transfer learning strategy as it improves the performance of all attributes; especially for the classes with the least data: Streaks (7.9%) and Negative Network (3.1%). Also, this approach improves our performance score from 0.227 to 0.278.

We quantify the computation and memory requirements. As a rough estimation, in our method each attention U-Net requires about $2320k$ parameters for each class. Hence, in total we trained approximately $2320k \times 5 = 11600k$, which is below 12 million parameters. On the other hand, ResNet [6] (the network architecture used by the winning team NWN) typically requires about $60344k$ parameters for a single class; hence $60344k \times 5 = 301721k$, which is more than 300 million parameters for five classes.

5 Conclusion

In this work, we proposed a novel approach for skin attributes detection based on Attention U-net with multi-scale image inputs. While our network only requires a small number of parameters compared to other state-of-the-art methods, it achieves performance on par or better compared to the best approaches for some classes. This advantage benefits from our effective transfer learning tactic that leverages the segmentation network as initialization. In term of the social impact, our system can contribute as a visually explainable system to doctors for the early diagnosis and treatment of skin cancer.

References

1. Bovik, A.C.: The Essential Guide to Image Processing. Academic Press, Cambridge (2009)
2. Chen, L.C., Papandreou, G., Schroff, F., Adam, H.: Rethinking atrous convolution for semantic image segmentation. arXiv preprint arXiv:1706.05587 (2017)
3. Chollet, F.: Xception: deep learning with depthwise separable convolutions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1251–1258 (2017)
4. Codella, N., et al.: Skin lesion analysis toward melanoma detection 2018: a challenge hosted by the International Skin Imaging Collaboration (ISIC). arXiv preprint arXiv:1902.03368 (2019)
5. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: ImageNet: a large-scale hierarchical image database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition, pp. 248–255. IEEE (2009)
6. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)
7. Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q.: Densely connected convolutional networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4700–4708 (2017)
8. Koohbanani, N.A., Jahanifar, M., Tajeddin, N.Z., Gooya, A., Rajpoot, N.: Leveraging transfer learning for segmenting lesions and their attributes in dermoscopy images. arXiv preprint arXiv:1809.10243 (2018)
9. Oktay, O., et al.: Attention U-Net: learning where to look for the pancreas. arXiv preprint arXiv:1804.03999 (2018)
10. Reddi, S.J., Kale, S., Kumar, S.: On the convergence of Adam and beyond. arXiv preprint arXiv:1904.09237 (2019)
11. Ronneberger, O., Fischer, P., Brox, T.: U-Net: convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) MICCAI 2015. LNCS, vol. 9351, pp. 234–241. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24574-4_28
12. Siegel, R.L., Miller, K.D., Jemal, A.: Cancer statistics, 2016. CA: Cancer J. Clin. **66**(1), 7–30 (2016)
13. Zhao, H., Shi, J., Qi, X., Wang, X., Jia, J.: Pyramid scene parsing network. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2881–2890 (2017)
14. Zou, J., Ma, X., Zhong, C., Zhang, Y.: Dermoscopic image analysis for ISIC challenge 2018. arXiv preprint arXiv:1807.08948 (2018)