DEEP AUTO-CONTEXT FULLY CONVOLUTIONAL NEURAL NETWORK FOR SKIN LESION SEGMENTATION

Zahra Mirikharaji, Saeed Izadi, Jeremy Kawahara, and Ghassan Hamarneh

Medical Image Analysis Lab, School of Computing Science, Simon Fraser University, Canada

ABSTRACT

Automatic segmentation of skin lesions in dermoscopy and clinical images is a common initial step in computer aided diagnosis. However, the low contrast of lesion boundaries and the existence of misleading image artifacts (e.g., hair), make segmenting skin lesions a challenging task. We propose a deep auto-context architecture that incorporates image appearance information as well as contextual information to predict the pixel-wise probability of a skin lesion. A sequence of fully convolutional networks is trained in a consecutive manner, where the input of each classifier is the original image concatenated with a degraded a posteriori probability estimated by the previous classifier. In contrast to common approaches that use morphological operations or thresholds to correct irregularities in the predicted lesion segmentation mask, our auto-context architecture efficiently refines the skin segmentation without any post-processing. Our experiments demonstrate that using our deep auto-context framework improves the segmentation performance of U-Net by 3.5% in terms of Dice similarity coefficient.

1. INTRODUCTION

Over the last three decades, the prevalence of skin cancer in the United States (U.S.) has been higher than all other cancers combined [1]. The most lethal type of skin cancer is melanoma with the mortality rate of one person per hour in the U.S. [2]. Early detection of melanoma plays an essential role in increasing the skin cancer survival rate. Even when utilizing dermoscopy with skin reflection suppression and widely used diagnostic criteria, like the 7-point checklist, diagnostic accuracy is still not perfect. Development of automatic approaches for skin lesion analysis has the potential to accelerate and improve skin cancer detection and improve survival prognosis.

Several efforts have been made towards automated and semi-automated analysis of dermoscopy and clinical skin images. Segmentation of skin lesions is an important precursor to lesion feature extraction and classification. Inter- and intra-subject lesion variability, image artifacts like hair and blood vessels, and low image quality are among the factors that make automatic skin lesion segmentation a challenging task.

Many classical methods based on image processing techniques, such as histogram thresholding, edge- and region-based methods, active contours and k-means clustering, have been applied to the skin lesion segmentation problem. Celebi et al. summarized skin lesion boundary detection approaches and discussed the difficulty and limitations of computerized approaches until 2009 [3]. Most of these techniques are based on low level and hand-crafted features tested on a small set of dataset. The complexity of classifying the presence of skin lesions at a pixel-level makes it challenging to accurately segment lesions when relying on low level and hand-crafted features like color, texture, shape, and size.

Recently, deep learning approaches, especially convolutional neural network (CNN), have been widely used for image understanding tasks [4]. Long et al. introduced fully convolutional networks (FCN) [5], where the fully connected layers are discarded and deconvolutional upsampling layers are learned to generate high-resolution feature maps used for semantic segmentation. Ronneberger et al. proposed U-Net, an end-to-end decoder-encoder fully convolutional architecture that concatenates higher resolution feature maps from the encoder's earlier layers with upsampled feature maps in the decoder, to perform semantic segmentation [6]. The U-Net architecture is widely used by the medical image analysis community, illustrating the effectiveness of using FCNs for image segmentation tasks.

Jafari et al. used a patch-based CNN to classify pixels of normal skin from lesion pixels [7]. They extracted two scale patches called local and global patches around each pixel, and used a parallel two-path CNN followed by a joint fully connected layer to classify each pixel. Feeding patches instead of the whole image into the network, in addition to a morphological post-processing step, makes their pipeline slow at test time. Recently Yu et al. leveraged a very deep fully convolutional residual network with more than 50 layers to segment skin lesions [8]. Their achieved performance confirms the advantage of very deep discriminative features for accurate skin lesion segmentation.

Tu et al. proposed auto-context, an iterative learning algorithm for structural refinement, which uses contextual in-

Thanks to NVIDIA Corporation for the donation of the Titan X Pascal GPU used for this research.

formation in addition to appearance information for image understanding models [9]. Auto-context takes as input appearance information as well as features from the predicted probability maps of the previous iteration into the current iteration. By iterating this process, classifiers are able to gradually correct earlier mistakes by using new contextual features. The original auto-context algorithm was originally proposed for patch-based segmentation with handcrafted features [9]. Salehi et al. recently proposed Auto-Net, an auto-context CNN to extract the fetal brain from 3D MRI [10].

In this work, we applied an auto-context deep framework that sequentially learns improved skin lesion segmentation maps given RGB skin images. We train a sequence of FCNs so that each take as input the original images as well as the degraded a posteriori probability map estimated by the previous early-stopped FCN. Compared to earlier patch-based auto-context approaches, feeding the whole contextual information into a CNN, leads to automatic learning of deep multi-scale contextual features. Also, in comparison to Auto-Net, during training we use the probability maps generated by early-stopped FCNs to prevent overfitting in the subsequent models, and use the fully converged FCNs for testing. The goal of this work is to show the advantage of applying deep architectures to the skin lesion segmentation task in an auto-context fashion. Our experimental results illustrate how deep auto-context framework and the early stopping technique refine the predicted probabilities when compared to a single FCN.

2. METHODOLOGY

Given a set of N images and their corresponding ground truth segmentations $\{(X(n), Y(n)); n = 1, 2, \ldots, N\}$, our goal is to learn the segmentation model parameters θ that generalize well on unseen samples. For an image with m-pixels $X = (x_1, x_2, \ldots, x_m)$ and a corresponding ground truth labelling $Y = (y_1, y_2, \ldots, y_m)$ such that $y_i \in \{0, 1\}$, we seek a dense prediction configuration Y^* which maximize the a posteriori probability given an observed image: $Y^* = \arg\max_Y p(Y|X; \theta)$.

One way of studying the a posteriori probability is to rewrite $p(Y|X)$ using Bayes rule as $p(Y|X) \propto p(X|Y)p(Y)$ and investigate its components separately. Alternatively, in this work we investigate the a posteriori distribution directly.

In fully convolutional networks (FCNs) the a posteriori probability $p(\cdot)$ is modeled by applying a softmax function on activation maps in the last layer as follows:

$$p(y_i = k|X(N_i); \theta) = \frac{exp(a_k(X(N_i)))}{\Sigma_c exp(a_c(X(N_i)))} \quad (1)$$

where $X(N_i)$ is a neighboring window around pixel x_i, and a_k is the output activation for class k. The a posteriori probability gets updated iteratively by measuring the compatibility

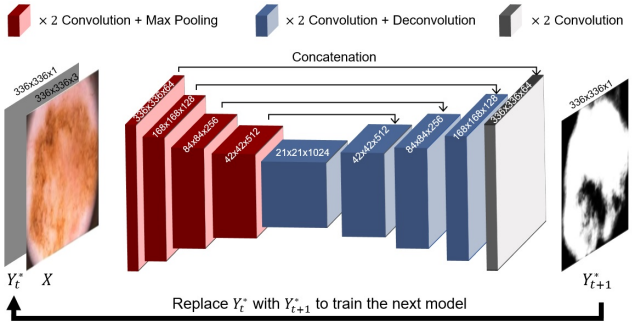

Fig. 1. Deep auto-context architecture schematic. The model $t+1$ is trained on the concatenation of the original image and the a posteriori probability from model t. Sizes on red and blue blocks show the feature map sizes before max pooling and deconvolution, respectively.

of Y^* and Y in a loss function and back propagating the error to update the set of model parameters θ. We note that although we write this equation in terms of individual pixels, we predict the entire dense segmentation in a single forward pass. In addition, the neighboring pixels N_i in equation 1, highlight how the output value y_i is dependent on the network receptive field, and not all the pixels in the input X.

As illustrated in equation 1, in conventional classification approaches like FCN, the likelihood of giving the class label k to pixel i, only depends on the deep features extracted from image appearance. However, considering the class labels of surrounding pixels is informative. Conditional random field [11] has been widely used to explicitly formulate the dependency of each class label to the neighboring pixels class labels. The common drawback of these approaches is the explosion of computational complexity if long-range contextual information from a large neighborhood is considered in the model.

To include a large scale of context information in the predicted segmentation, we adopt the auto-context architecture, which is composed of multiple FCN models. The idea is to design an iterative framework that predicts pixel-wise classification not only based on the image appearance but also considers the a posteriori probabilities estimated by the previous classifier. In the proposed approach, we have a sequence of T FCN models learned in a consecutive manner. The $t+1$-th model is trained given the training data $(X(n), Y(n), Y_t^*(n)); n = 1, 2, ..., N$, where $Y_t^*(n)$ is the a posteriori probability provided by:

$$Y_t^* = \arg\max_Y p(Y|X, Y_{t-1}^*; \theta_t). \quad (2)$$

For the first model ($t = 0$), the segmentation probability map, $Y_0^*(n)$, is a uniform distribution map. Since the uniform distribution does not contain actual contextual information, in the first iteration, the network gains no additional information

from it. Fig. 1 shows the proposed deep auto-context architecture. We start by training a fully convolutional network with an architecture similar to U-Net to segment the skin lesion. Once the first FCN is trained, it is applied to all training and validation data and a posteriori probability map is generated for each image. We concatenate the original RGB image channel with the a posteriori probability map and train a new FCN to refine the a posteriori probability estimated by previous network. The same procedure is repeated until the algorithm converges. At the test time, given a new image, FCNs are sequentially applied.

When training using the auto-context architecture, passing the training data sequentially to the subsequent FCN models may not ensure effective fine-tuning because the data and their ground truth were already shown to the previous models. One way to prevent overfitting, when training patch-based auto-context models, is to split the data in such a way that $t+1$-th model is not trained on data used in the first t models [12]. Splitting the data in this way may not be the best approach to deal with this overfitting. An alternative approach for dealing with this problem is degrading the a posteriori segmentation maps generated by the FCN to produce new maps that look more like the segmentation probability map encountered with novel test images. In this work, we hypothesize that using the parameters of the t-th auto-context model *before convergence* will result in degraded segmentation maps that in turn cause the $t+1$-th model to be trained on more realistic and challenging a posteriori probability maps (rather than ones already overfit to the training data). Thus, during training, we train each FCN until convergence but generate the a posteriori probabilities of training data using the network parameters before convergence. We feed the concatenation of these probability maps and the original image to the next deep model. At test time, we applied the sequence of fully converged FCNs to a new test image.

3. EXPERIMENTS

Data Description We validated the proposed method on ISBI 2016, *Skin Lesion Analysis Towards Melanoma Detection Challenge*, data set [13]. The data set is composed of 900 training images. We used 20% of the training data for validation, and to set model hyper-parameters. Another separate set of 379 test images and their ground truth, provided by the challenge organizers, is used to evaluate the model. We re-sized all images to 336×336 and normalized them using the mean and standard deviation of RGB pixels values computed over all training data. To increase training data and make the model more robust, we augment the training images with the rotations of 90, 180 and 270 degrees, and horizontal and vertical flipping without any replication.

Implementation We implemented and trained our deep auto-context networks using the PyTorch framework. All fully convolutional networks are initialized by a random Gaussian distribution and learned from scratch. We used stochastic gradient descent as a solver and a mini-batch of size 2, restricted by our GPU memory. A momentum of 0.99 and a weight decay of 0.0005 is used for all fully convolutional networks. The learning rate was tuned for each FCN on our validation set. The network trained for the first step of the auto-context architecture converges after approximately $90,000$ iterations while the second and third networks in the auto-context architecture take approximately $34,000$ and $11,000$, respectively. Training the whole deep auto-context architecture takes 2 days on our single 12 GB GPU memory.

Table 1. Segmentation quantitative performance comparison in U-Net and different auto-context iterations. Jacc., Acc., Spec. and Sens. indicate Jaccard, Accuracy, Specificity and Sensitivity metrics. Bold numbers indicate the best performance. All values are in percentages.

	Method	Dice	Jacc.	Acc.	Spec.	Sens.
A	U-Net [6]	86.86	78.29	93.63	93.51	93.05
B	Ours (T=1)	88.42	80.16	95.09	95.07	93.51
C	Ours (T=2)	88.98	80.76	**95.12**	92.04	**96.11**
D	Ours degraded (T=1)	**90.11**	**83.30**	95.02	**97.00**	90.15

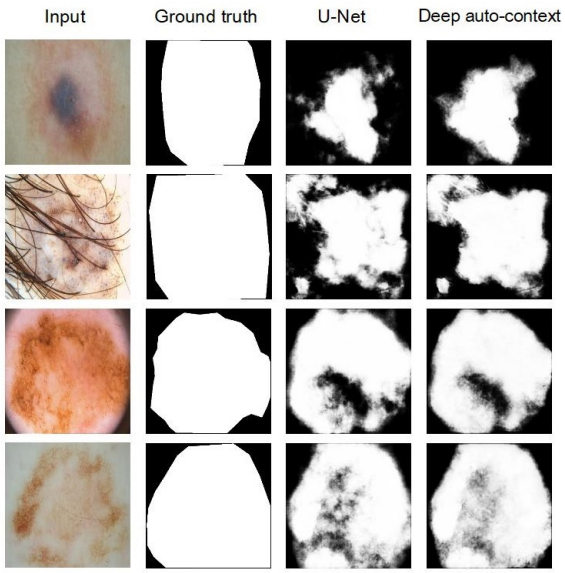

Fig. 2. Resulting segmentation masks over challenging cases.

Results We evaluate the contribution of stacking the FCNs in an auto-context by using the following pixel-level metrics used in the ISBI 2016 segmentation challenge: Sensitivity, Specificity, Accuracy, Jaccard and Dice. We calculated each of these metrics for each test image and reported the average value over all test samples. We use U-Net [6], as our baseline network architecture. For a fair comparison, we used the same architecture for sequential FCNs in the auto-context ar-

chitecture. Each model within the auto-context architecture is trained individually by optimizing a binary cross-entropy loss function. Table 1 indicates the performance of auto-context in different iterations in comparison to U-Net. Rows B and C confirm the advantage of using FCNs in an auto-context fashion. In comparison to U-Net, after one iteration of auto-context, Dice similarity coefficient increases by approximately 1.5% (Row B). Iterating the auto-context for the second iteration further improve the Dice similarity coefficient (0.5% as shown in Row C). In our experiments, we found that iterating the auto-context model beyond the second iteration did not improve results. To degrade the a posteriori probability maps of training data and make them more similar to the probability maps of unseen data at test time, we used the FCN parameters at the auto-context iteration 0 before convergence (after $34,000$ iterations) and generate the training data probability maps. Row D shows the result of training the FCN at iteration 1 using contextual information generated by the degraded probability maps. Iterating the new auto-context model after the first iteration didn't improve the predictions. We observe that degrading the a posteriori probability maps helps avoid overfitting and improves results, and thus recommend the degraded (T=1) as the best option. 28 teams have participated in the skin lesion segmentation part of the 2016 challenge. Based on these reported numbers[1], our performance ranked the second among the challenge participants.

Fig. 2 presents qualitative results of our proposed approach over some challenging cases. Comparing the results of our degraded deep auto-context with U-Net illustrates that by iterative training a fully convolutional network using the a posteriori probability segmentation map in addition to the original image, FCNs are able to gradually correct earlier mistakes by using new contextual features. While many previously proposed deep architectures for skin segmentation apply post-processing approaches (e.g., multi-thresholding, morphological operations) to filter false negative and positive gaps inside and outside the lesions [14], these post-processing operations are disconnected from the training step, require additional hyper-parameters, and are computationally expensive at test time.

4. CONCLUSIONS

We proposed to use a sequence of fully convolutional networks in an auto-context manner to sequentially refine the predicted skin lesion segmentation map of the previous network. The key contribution of this work is to incorporate contextual information into deep feature extraction models. Our proposed deep auto-context approach is a general, easy to implement framework, that is applicable regardless of the deep architecture, and is used to further refine segmentations.

[1] https://challenge.kitware.com/#challenge/560d7856cad3a57cfde481ba

5. REFERENCES

[1] Rogers et al., "Incidence estimate of nonmelanoma skin cancer (keratinocyte carcinomas) in the us population, 2012," *JAMA dermatology*, vol. 151, no. 10, pp. 1081–1086, 2015.

[2] "Cancer facts and figures 2017," http://www.cancer.org/acs/groups/content/@editorial/documents/document/acspc-048738.pdf, January 10, 2017.

[3] Celebi et al., "Lesion border detection in dermoscopy images," *CMIG*, vol. 33, no. 2, pp. 148–153, 2009.

[4] Krizhevsky et al., "ImageNet classification with deep convolutional neural networks," in *NIPS*, 2012, pp. 1097–1105.

[5] Long et al., "Fully convolutional networks for semantic segmentation," in *IEEE CVPR*, 2015, pp. 3431–3440.

[6] Ronneberger et al., "U-net: Convolutional networks for biomedical image segmentation," in *MICCAI*, 2015, pp. 234–241.

[7] Jafari et al., "Extraction of skin lesions from non-dermoscopic images for surgical excision of melanoma," *IJCARS*, pp. 1021–1030, 2017.

[8] Yu et al., "Automated melanoma recognition in dermoscopy images via very deep residual networks," *IEEE TMI*, vol. 36, no. 4, pp. 994–1004, 2017.

[9] Tu et al., "Auto-context and its application to high-level vision tasks and 3D brain image segmentation," *IEEE TPAMI*, vol. 32, no. 10, pp. 1744–1757, 2010.

[10] Salehi et al., "Auto-context convolutional neural network (auto-net) for brain extraction in magnetic resonance imaging," *IEEE TMI*, 2017.

[11] Kumar et al., "Discriminative random fields: A discriminative framework for contextual interaction in classification," in *IEEE ICCV*. IEEE, 2003, pp. 1150–1157.

[12] Kawahara et al., "Augmenting Auto-context with Global Geometric Features for Spinal Cord Segmentation," in *MLMI*, 2013, vol. 8184.

[13] Gutman et al., "Skin lesion analysis toward melanoma detection: A challenge at the international symposium on biomedical imaging (ISBI) 2016, hosted by the international skin imaging collaboration (ISIC)," *arXiv preprint arXiv:1605.01397*, 2016.

[14] Yuan et al., "Automatic skin lesion segmentation with fully convolutional-deconvolutional networks," *arXiv preprint arXiv:1703.05165*, 2017.