Analyzing Skin Lesions in Dermoscopy Images Using Convolutional Neural Networks

Anonymous
Department of Computer Science
Golisano College of Computing and Information Sciences
Rochester Institute of Technology
Rochester, NY 14586

Abstract—In this paper, we discuss the problem of automatic skin lesion analysis, specifically melanoma detection and semantic segmentation. We accomplish this by using deep learning techniques to perform classification on publicly available dermoscopic images. Skin cancer, of which melanoma is a type, is the deadliest form of cancer in the US and more than four million cases are diagnosed in the US every year. In this work, we present our efforts towards an accessible, deep learning-based system that can be used for skin lesion classification, thus leading to an improved melanoma screening system. For classification, a deep convolutional neural network architecture is first implemented over the raw images. In addition, hand-coded features such as 166 dimensional color histogram distribution, edge histogram and multi-scale color local binary patterns are extracted from the images and fed to a random forest classifier. The average of the outputs from the two mentioned classifiers is taken as the final classification result. The classification task achieves an accuracy of 80.3%, AUC score of 0.69 and a precision score of 0.81. For segmentation, we implement a convolutional-deconvolutional architecture and the segmentation model achieves a Dice coefficient of 73.5%.

Index Terms—Learning systems, Dermatology, Classification algorithms, Convolutional Neural Networks (CNN)

I. INTRODUCTION

Melanoma is the most dangerous type of skin cancer[1]. It is estimated that 91,270 new cases of melanoma will be diagnosed in the U.S. in 2018[2][3]. According to the National Cancer Institute (NCI), around a million of Americans have some or the other form of Melanoma, currently¹.

Although the 5-year survival rate of melanoma is 98% when detected and treated early, yet an estimated 9,320 people will die of melanoma in 2018 due to late-stage diagnosis[2]. If the cancer is allowed to spread from local sites to regional, the survival rate drops to 63% and further drops to 17% in the later stages of the disease. Although less than one percent of skin cancer cases are melanoma, it accounts for the vast majority of skin cancer deaths.

Melanoma is a type of cancer that typically occurs on the skin and results from the pigment-containing cells known as melanocytes[4], hence is often perceived on the skin as a mole that is increasing in size, changing in color, having more irregular edges, or general skin breakdown. This kind of cancer, like other cancers, can be inspected visually by the help of device called dermatoscope, which is basically a microscope with a magnifier attached. Studies have shown that the diagnosis of melanoma done by specialists yields more accurate results than the diagnosis done by non-specialists of dermoscopy [5]. Dermoscopy images or video clips can be digitally captured and processed for further analysis.

In this work, we evaluate skin lesions from dermoscopy images using convolutional neural networks (CNNs). in recent years, CNNs has been shown to be a very effective technique for image classification tasks, such as is demonstrated in the ImageNet challenge to classify up to a thousand different natural image categories [6]. One major advantage of CNNs is that it is able to perform a form of transfer learning, where a network is originally trained for a specific task but the weights learned for that task can significantly improve the learning when applied to a very different task. Our proposed method takes advantage of this property of CNNs for our model training.

Also, we used low-level visual features of the lesion images like color histogram, edge histogram, multi-scale color local binary patterns to feed to a machine learning classifier.


Fig. 1. Sample images from the original dataset
We evaluate the skin lesions present in the subset of the large-scale publicly accessible dataset of dermoscopy images collected by The International Skin Imaging Collaboration (ISIC) [7], and used in the 2017 challenge on dermoscopic image analysis. As laid out by the challenge, there were three main tasks: - lesion segmentation, feature detection, and melanoma classification. Using deep learning architectures, we successfully perform two of the three tasks and describe our implementation and results in the rest of the paper. The results we obtained would place us in top 10 positions of the leaderboard for both the classification and the segmentation tasks. Figure 1 shows some samples of the dermoscopy images from the ISIC dataset.

II. PAST RELATED WORK

The ABCDEs are the conditions used by dermatologists to detect melanomas. These include Asymmetry, irregular edges, uneven skin surface and more than one Color present, or a large (greater than 6mm) Diameter, and finally, the Evolution of the moles (how the first four characteristics change over time) [8]. Other similar rules or guiding principles include the 3-point checklist[9], the 7-point checklist[10] and Color, Architecture, Symmetry and Homogeneity (CASH)[11]. But visual diagnosis can be complicated and can lead to subjective results. Skin lesion analysis involves challenging visual classification and segmentation tasks because of the wide variability in the characteristics of the disease and also because the skin every patient is different from one another, so is the growth of the lesion. Computer vision and machine learning scientists have attempted melanoma classification and general skin disease analysis, and in this section, we discuss some of these related past approaches.

Masood and Al-Jumaily[12] summarized the comparative analysis of different algorithms for feature selection and classification of Melanoma. Computer-aided systems are built using the data collected from the actual cases, and the model is expressed as a set of rules with the help of domain experts. This method was quite expensive as the system demanded very fast image processing and a detailed knowledge to develop the best set of rules. They did not apply deep learning techniques at this time.

Sabbaghi et. al[13] approached the problem of classifying melanomas in dermoscopy images from a different perspective. They applied stacked sparse auto-encoders for discovering the latent information features from input image pixel intensities. The learned high-level features were fed into a nonlinear classifier to detect Melanoma. They accomplished this by utilizing a bag-of-features representation for the images, thus improving the overall accuracy of their system. Their method was evaluated on a dataset of total 814 dermoscopy images, of which 640 were benign and 174 malignant. They tested on a total of 244 images. Unlike many other deep learning techniques applied to melanoma detections, this work did not implement convolutional neural networks.

Kawahara et. al[14] present a CNN architecture which they apply to the Dermofit Image Library provided by the American Cancer Society. The dataset had 1300 skin images with corresponding class labels and lesion segmentations with 10 lesion categories, including Melanoma. They improved the state-of-the-art results on that dataset to 85.8% from the previous results of 75.1%. No lesion segmentations were performed. Similarly, Majtner et. al[15] presented a technique for skin lesion classification, which combined CNN and handcrafted features, specifically RSurf features and local binary patterns (LBP). They also compared the results using their technique with the results from the melanoma classification challenge, hosted by the International Skin Imaging Collaboration (ISIC) - the same dataset we test our approach on. Their reported accuracies were between 79.4 and 80.5% on the ISIC dataset.

Yu et. al[16] placed first-place in the ISIC 2016 challenge, a very similar challenge to that on which we report our test results. The achieved this by using a very deep CNN architecture (having more than 50 layers). They applied residual learning to and then constructed a fully convolutional residual network (FCRN) for lesion segmentation process. They reported accuracies of 85.5% on the 2016 classification challenge, and report up to 95.3% accuracy on the segmentation task.

Other very successful deep learning techniques that have been applied to skin lesion analysis include works by Shimizu et. al [17], Schaefer et. al [18], Barata et. al [19]. Abbas et. al [20], Iyatomi et. al [21] and one of the most highly cited dermoscopy image analysis paper by Celebi et. al [22].

To perform a semantic segmentation Noh et. al [23] uses CNN. A multi-layer convolution-deconvolution is learned which is composed of convolution, deconvolution, unpooling and ReLu operations. Segmentation using a CNN is basically a pixel-wise classification and the network is based on the VGG16 layer net[24] architecture. They achieved an accuracy of 72.5% on PASCAL VOC dataset[25] using a similar architecture. Also, Ronneberger et. al [26] described a network architecture named U-net which is similar to the conv-deconv architecture described above. Additional details on U-net are provided in Section III-D

An article in the journal Nature[27] presented a CNN-trained end-to-end system where the disease labels were estimated from images directly, similar to the techniques described above. The classification task was implemented using a proprietary dermatologist-labeled dataset which had total 129,450 clinical images, out of which, 3,374 were dermoscopy related images. They utilized GoogLeNet Inception v3 CNN architecture[28] that was pre-trained on approximately 1.28 million images (1,000 object categories) from the 2014 ImageNet Large Scale Visual Recognition Challenge. They used this pre-trained network to train their skin lesions image dataset using transfer learning. For one of the partition, CNN achieves the accuracy of around 72.1% as compared to two dermatologists who attain the accuracy of 65.56% and 66% respectively.

To classify melanoma images, [29] used a very simple architecture of the convolutional neural network. They classify the lesion images without segmenting or cropping the lesions. Simple pre-processing techniques were applied to the images.
like resizing the images to 256 x 256 and subtract the mean to center the data. They arranged the labels of the images so that the learning algorithm doesn’t get same label images consecutively. The CNN classifier has 17 layers which make total 5 convolutional blocks. The testing error, after the classifier is trained, is 0.189.

A method based on which learns the Mahalanobis distance during training and reduces the feature dimensions is proposed by [30]. The approach works by reducing the dimensions of the features. This work is based on the idea that feature vectors with the dimensionality of a few hundred may not work well with a classifier. If the dimensionality is reduced, the performance of the classifier should be improved. Training data is used to learn Mahalanobis distance which is then later used for local manifold construction.

Codella et. al [31] works with ensemble of classifiers. They use a fully connected CNN, ResNet and Unet to extract features and feed the features to a Support Vector Machines for classification.

III. OUR APPROACH

Given a skin image $I$ with a corresponding class label $y$ representing the category of the skin lesion, our goal is to extract image features $f = \phi(I)$ that discriminate well between the class labels. We utilize a convolutional neural network (CNNs) architecture, along with other machine learning classification techniques for the segmentation and classification of skin lesions as malign or benign. Melanoma is a malignant skin tumor, while nevus and seborrheic keratosis are considered to be benign tumors. This work develops a binary image classification model: the classification model distinguishes between (a) melanoma and (b) nevus and seborrheic keratosis.

![Fig. 2. Process Flow Diagram](image)

Figure 2 shows how the classification and segmentation models are developed. The CNN used for segmentation are trained using the training images and the ground truth which is in the form of a corresponding binary mask. The output of the segmentation is a binary mask for the test image which is further used to segment out the lesion. For the CNN classifier, the training images are used to make the classifier learn. Next step is to extract some hand-coded features and feed them to a Random Forest classifier. The results from both the classifier are averaged to output the final classification result. Three hand-coded features which are used to train a Random Forest Classifier are:

- 166-D color histogram
- Edge histogram
- Multi-scale color local binary patterns

A. Preprocessing

in order to avoid overfitting of the CNN to the data, it is common to augment the data by translating, resizing or rotating an image already existing in the dataset by a random angle between -15 and 15 degrees. Hence, given the $i$th image in the original dataset, $I^{(i)}$, we can alter it to produce a $j$th augmentation of the $i$th image $\tilde{I}_{j}^{(i)}$. The augmented image is normalized as

$$\tilde{f}_{j}^{(i)} = \phi(\tilde{I}_{j}^{(i)} - \mu(\tilde{I}_{j}^{(i)}))$$

where $\mu$ is the mean value for each channel in the input image $I$ and $\phi(I)$ is the feature extracted from the fully connected layer of the CNN. The full set of augmented normalized feature vectors $\tilde{f}^{(i)}$ are added to the dataset to increase the training size.

B. Hand-Coded Feature Extraction

We extracted some of the low-level visual features of the lesion images as these features have proven to be very important in medical images. Color is an important feature in the lesion images. To find a 166-dimensional color histogram, images were converted to HSV colorspace. To produce a histogram of 166 colors, 18, 3, 3 and 4 bins of hue, saturation, value and grayscale respectively were created. As we are not cropping the lesions from the images, we use the edge as a feature which provides information about the location and the shape of the lesion on the skin. To do so, we extract edge histograms which contain the direction of the edge(8 bins) and magnitude of the edge(8 bins). To get the texture information, we extract multiscale color local binary patterns which are basically an extension of grayscale local binary patterns. They provide information about the texture of an image which is a very informative feature in medical images as the colors in the lesion grow to be uneven and the texture of the lesion varies from person to person.

After extracting the above-explained visual features, these are fed into a non-linear classifier for training. A random forest classifier[32] grows multiple decision trees. Each tree gives a classification output, the forest outputs the class which is chosen by most of the trees. After testing with different parameters, the final random forest classifier in this work consists of 20 decision trees which provide the most accurate results.
C. Classification

For the classification task, we first trained an end to end CNN whose architecture is similar to a ResNet [33]. Residual networks are deep but simple architectures of neural networks. They have shown state of the art results in image classification tasks. Figure 3 shows a residual block used in residual networks. In a residual network, the inputs of a lower layer are fed to a node in a higher layer. A residual block in a network ensures that in each back-propagation, layers are learning something new. This work uses an end to end 101-layer deep neural network with a softmax classifier to produce the final results.

D. Segmentation

This task involves the automatic prediction of lesion segmentation from dermoscopic images getting converted into the corresponding binary masks. To perform this task, we train an architecture similar to U-net[26] network.

Fig. 3. Residual block for CNN used in classification

![Residual Block Diagram]

Fig. 4. CNN Architecture for image segmentation

The architecture of this network is a conv-deconv structure. On the left side, there are five convolutional blocks which are downsampling the image. While on the right side, there are another five convolutional blocks which upsampling the features to finally generate a binary mask as the output. After each block of convolution, image is being cropped and finally producing a 64-D feature vector. This network consists of total 23 layers which include max-pooling and ReLu (Rectified Linear Unit).

Figure 4 shows the architecture of the network used for image segmentation. Convolution blocks from 1 to 9 use ReLU activation function. Convolution operations are followed by a downsampling process through a $2 \times 2$ max-pooling operation with stride 2. The deconvolution architecture includes an upsampling operation of the features. The last convolution layer uses Sigmoid as the output function to output the binary masks. Adam optimizer[34] is used during training the network. We experimented with two different color spaces for the input images (Grayscale and YCbCr) to the network and found out that converting the input images to YCbCr gives the best results for segmentation.

To train the network, the input images and their corresponding binary maps are fed into the network and trained using standard stochastic gradient descent. The prediction function is computed using pixel-wise soft-max over the last feature map and minimizing the weights using the cross-entropy loss function. The parameters for the original binary masks pre-computed as suggested by [26].

IV. EXPERIMENTS, RESULTS AND DISCUSSION

This work uses the dataset provided ISBI 2017 Challenge[35]. The dataset is publicly available. There are 2000 RGB images provided for training, 150 images for validation and 600 images for testing. Ground truth for classification is provided in a CSV file which provides class labels for every image. Ground truth for segmentation is provided as binary masks for each image. Figure 5 shows a sample image and its corresponding binary mask from the original dataset.
Data is heavily imbalanced and we balanced it by doing data augmentation. Details of data augmentation specific to classification task are provided in Section IV-A. Figure I shows the distribution of classes in the original training dataset and in the augmented data. For both segmentation and classification all the images were preprocessed using following steps:

- To center the data, mean of the images was subtracted from the data.
- The pixel values were normalized by dividing its standard deviation across the data.
- Input images were resized to 256x256 to be accepted by the architectures for segmentation and classification.

**A. Feature Extraction and Classification**

To balance the heavily imbalanced data, we did data augmentation to generate more images provided for training. The applied data augmentation techniques were: rotation 40 degrees, random horizontal and vertical shifts, shear intensity, random zoom and horizontal flip. 6 shows few examples of the images generated after data augmentation.

The balanced dataset was used to extract the hand-coded feature map using the computer vision techniques. The extracted features were fed to a Random Forest classifier. The CNN classifier and Random Forest classifier were tested on 600 images. Table II shows the scores for classification. Table III shows the confusion matrix for classification task.

**B. Segmentation**

For segmentation, 2000 images in YCbCr colorspace, were used to train the described conv-deconv architecture and 150 images for validation. We used 600 images to test the model.

The metric used to evaluate segmentation model is called Dice coefficient. It is computed by comparing the pixel-wise agreement between the ground truth binary mask(Y) and the corresponding predicted segmentation binary mask(X).

$$\text{Dice coefficient} = \frac{2 \times |X \cap Y|}{|X| + |Y|}$$

We achieved a Dice Coefficient of 73.5% on testing 600 images. Figure 7 shows some sample input and output binary masks produced by the segmentation model.

**V. Future Work**

We are currently in the process of porting the presented technique to an android application which will be tested firstly with a small population of Dermatologists in the clinic to access its effectiveness in detecting melanoma, with a high degree of accuracy. This will require significant testing before releasing it to the general population for everyday usage.
VI. CONCLUSION

This paper presented approaches for semantic segmentation and classification of skin lesions from dermoscopic images. The proposed model for segmentation achieved a Dice coefficient of 73.5% and the classification model achieved accuracies of 80.5%.

In future, we intend to continue to push the boundaries on the performance of the network including the accuracies currently reported. Our current work is limited by the lack of availability of more powerful hardware with high graphical processing unit (GPU) power. Accessibility to higher GPUs will allow us to experiment further with different architecture layouts including deeper, multiple scaled networks while still maintaining the underlying structure of our presented architecture.

REFERENCES


