Deep Residual Neural Networks for Automated Basal Cell Carcinoma Detection

Hagar Maher

8th, October 2018

1 Abstract

- Current machine learning analyses of Basal Cell Carcinoma (BCC) dermoscopy images have failed to create a model viable for use in clinical applications.
- In this paper, They demonstrate a sensitivity and specificity that could make neural networks a realistic tool for dermatologists.
- Their algorithm follows a three step process:
    1. The original image is preprocessed and fed into the segmentation model
    2. A black and white lesion map is produced to extract the minimum area of the image
    3. The classification model is introduced for classifying whether an input image is BCC or not
- By building upon melanoma research performed by He, et al., They reached an overall weighted sensitivity and specificity of 96% and 89%, respectively.
- They demonstrated that deep residual neural networks (¿ 100 layers), carefully optimized, can surpass the limitations of depth one sees with more common convolutional neural networks.

2 Introduction

- Eighty percent of skin cancer occurrences are BCC, making it the most common type of cancer in the world.
- Despite the high prevalence of BCC, metastatic growth is quite uncommon, as is death.
- In order to get a clear image of a lesion, dermatologists often use dermoscopy imaging, a non-invasive method of visualizing micro-structures of the skin with high magni-fication.
- As Kittler, et al., has shown in, dermoscopy has improved the diagnostic accuracy of melanoma by 10-27% over simple naked eye examinations. Nevertheless, the accuracy of analyzing dermoscopy images still depends on the experience of a physician.
- A dermatologist not trained in reading dermoscopy images can be less accurate than naked eye analysis. Once the time required to obtain dermoscopy images is considered, it is clear that automated recognition systems are more efficient.
- Simple neural networks that achieve a modest level of accuracy are not difficult to create; however, a system that can achieve an accuracy viable for clinical use is quite challenging and depends on several features such as: vari-ations between size and shape, contrast between where the lesions begin and end, and artifacts such as hair color and veins.
- In addition, BCC itself is classified into 2 major sub-types:

  1. Nodular
  2. Superficial

  each with their own defining characteristics.
- There is little research on BCC and neural networks, and the works provided have been insufficient thus far.
- In 2011, Cheng, et al., published two papers related to BCC.

  1. The first paper diagnoses BCC via telangiectasia analysis. 96.7% accuracy was claimed, however, in actual clinical practice, true accuracy will be significantly less due to lack of telangiectasias in early lesions, and telangiectasias with similar features being in benign skin lesions.
  2. The second published paper used a novel technique of combining feature extraction from lesions with features of the patient's personal profile and physical exam characteristics of lesions. There was, however, no accuracy, specificity, or sensitivity values provided. As a result, no meaningful conclusions can be drawn from this paper

- In 2016, Kharazmi, et al., provided a method for detecting BCC based on the vascular features of a lesion. By generating a vessel mask based on pigmen-tation (specifically hemoglobin and melanin components) of the underlying skin, they were able to extract a set of 12 vascular features to use in diagnosis of BCC. While their system on its own would not be comprehensive enough, further improvements in our own system could incorporate Kharazmis' work in the future.

- To their knowledge, there is no previous work done on residual neural networks being trained on a dataset of BCC images.

- In this paper, they present, the first attempt to solve this problem. By using a fully convolutional residual neural network (FCRN) for segmentation and a deep residual neural network (NN) for classification, they seek to automatically diagnose a malignant lesion, providing a starting mode that can one day be used on images taken with an optical camera. their inspiration is drawn from Lequan, et al., paper on Automated Melanoma Recognition in Dermoscopy Images via Very Deep Residual Network. They base their work on improving upon the techniques they used

3 Method

3.1 Overview

They present the details of the two-stage deep residual NN for lesion segmentation and classification.

1. First, the input lesion's original image will be preprocessed and fed into the segmentation model.
2. Then, a black and white lesion map is generated based on the input image, which is then used to extract the segment of our original image.
3. Finally, our classification model is introduced for identifying if our input fundus image is BCC.

3.2 Residual Neural Networks

- Most biomedical imaging NN use convolutional neural networks (CNN) foranalysis,asthis hasdominatedthe field since the introduction of AlexNet in 2012. This approach poses several limitations:
    1. Med-ical datasets are often limited in quantity.
    2. Analyzing the images in these datasets requires the parsing of very discriminative features. For a NN to parse these features, the depth the NN reaches is most important. However, as a CNN's depth increases, accuracy eventually reaches a plateau and then degrades away from an optimal solution
    3. Other problems arise such as the exploding/vanishing gradient problem in which the gradient moves at a rate slower or faster than what is optimal in earlier layers of the network
- In many deep CNNs the gradient tends to be unstable and is a fundamental issue in gradient based learning. To avoid these issues, their model draws upon the work done by He, et al., on Residual Neural Networks (Resnets).
- ResNets are a new NN model built of tens to hundreds of residual blocks. Most residual blocks consist of two parts:
    1. The first being a set of nonlinear manifestations (e.g., convolutional layers, rectified linear unit layers, and batch normalization). The output of this block is summed with the second part of the residual block,
    2. The second part of the residual block, an identity mapping, which is a linear transformation that skips one or more layers. The identity block adds no extra complexity to the model as it has no extra parameters, but is key in solving the degradation problem. The result of these transformations is a NN capable of reaching hundreds (

possibly event housands) of layers deep and being increasingly more robust to the problems a standard CNN would experience when analyzing medical image datasets. A residual unit can be expressed as:

$$y_{n+1} = i(y_n) + RES(y_n; X_n)$$

where yn is the input feature set to Nth ϵ 1, ..., N residual unit and X_n is the weights for the Nth layer. RES(:) is the residual function itself, represented by a convolutional layer (weight), a batch normalisation (BN) and a rectified linear unit (ReLU), and iy_n is an identity mapping

3.2.1 Preprocessing

- Due to the nature of their dataset being a combination of two different databases, their first step was to normalize our inputs by re-sizing our dermoscopy images down to 480x480. This has the added benefit of allowing them to decrease training times by increasing their batch size, due to the increased free memory space.

- Their next step was data augmentation. Data augmentation allows them to create new dataset samples from existing ones while still maintaining labels for training. During training, they augmented the dataset by random rotations (maximal range of 270), and random flipping (vertical). This allowed them to increase their dataset from 1,520 images to 12,160 images. Those familiar with dermoscopy may make the argument that an additional preprocessing step is required due to the surrounding structures such as hair, moles, or droplets of water created when preparing the dermoscopy images. As seen in in Figure 1,

(a) Original Image (b) Output

Fig. 1: Segmentation Model with Hair Occlusion

an original BCC lesion is shown next to the output of their segmentation model demonstrating no irregularities when part of the lesion was occluded by hair follicles.

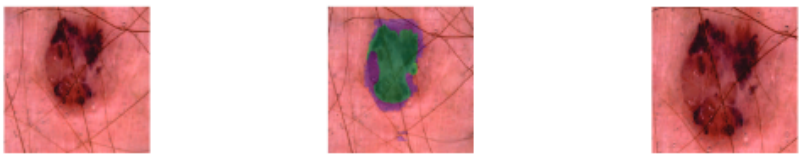
- Similar images in their dataset also suffered no loss in accuracy with occlusion, as they believe that their NN model is capable of interpreting extraneous information as non-pertinent to the lesion at hand.

3.3 Segmentation

- Despite the ResNet model presented in being de-signed for classification, only a small number of changes as described in were required to adapt for segmentation, the most important being using a fully convolutional network (FCN).
- As opposed to a traditional CNN, which is translation invariant, a FCN adds an expanding path composed of either efficient non linear transposed convolutions, or unpooling layers. FCN allows the NN to process spatial informa-tion that was missed from previous layers.
- Their automated segmentation model is based on the model presented in with the exceptions described below. Due to their utilization of Keras and Theano instead of Caffe, there were several steps which had to be performed before their model could be utilized. The main focus was recognizing how Caffe and Keras differ in terms of model network setup. The most pertinent difference was that many parameters do not translate directly over from Caffe to Keras. For this they used either defaults or their best intuition for rebuilding the model e.g. using a Glorot Uniform Initializer (a.k.a Xavier uniform initializer) for their deconvolution layers.

3.4 Integration

- After their segmentation algorithm processes an input im-age, the output is a set of matrix values between 0 and 1 that are converted to a single channel PNG and scaled between 0 to 255, with a value of 0 corresponding to a high probability of extraneous information and 255 indicating a high probability of lesion.
- They use this PNG to crop an overlay of the original lesion using a simple algorithm written in Python that draws a bounding box around the original lesion.
- Aside from the decrease in memory cost when loading a smaller image into the NN, there are added benefits, for example avoiding features that are not important to the lesion such as moles or hairs adjacent to the lesion, preventing accidental data leakage. The output of our integration stage is shown in figure 2.



(a) Original (b) Segmented (c) Cropped

Fig. 2: Segmentation Model from Beginning to End

3.5 Classification

- They employ a very deep residual network as their classifi-cation model, with an input from their integration state.
- The organization of the classification model is nearly identical to their segmentation model, except that the output is one hot encoded with a range from 0 to 1 representing a probability of whether an image is BCC or not.

4 Experimental Design and Results

4.1 Dataset

The dataset for their project draws upon two independent sources, composed of high resolution BCC lesions in ad-dition to a mixture of malignant and benign skin images.

1. Our first source is the public dataset used in the "Skin Lesion Analysis Towards Melanoma Detection" competition released with ISBI 2016.
2. The second source used is the International Skin Imaging Collaboration (ISIC) Archive, the largest online source of dermoscopy images.

4.2 Environment

- Their model was built with Python and Keras on top of Theano using a GTX 1080 Ti.
- The model was trained using ADAM with a batch size of 4, beta 1 of 0.9, beta 2 of 0.999, weight decay of 0.0, and a learning rate of 0.001.
- With this setup it took approximately 0.125 seconds to process one 480x480 image.

4.3 Results

- Their model was tested using three different depths: 53, 98 and 152. The depths were only adjusted in the classification model.
- Their best model was at a depth of 152, in which our overall weighted sensitivity and specificity for detecting BCC from non-BCC was 97% and 96%, respectively.
- To effectively judge the outputs of the model, they used four metrics:
    1. Sensitivity
    2. Specificity
    3. Accuracy
    4. Dice coefficient

- The results for their classification architectures is presented in table I.

TABLE I: Depth Comparisons

	AC	SE	SP	DI
53 Layers	.92	.98	.95	.88
98 Layers	.89	**.98**	.94	.85
152 Layers	**.93**	.97	**.96**	**.88**

- In figure 3 they plot their receiver operating characteristic graph for each model, demonstrating that their model with a depth of 152 performed slightly better than their 2 other models. They found that around epoch 45 their results tended to plateau. The AUC values for these curves are 0.96, 0.95, 0.96, respectively.

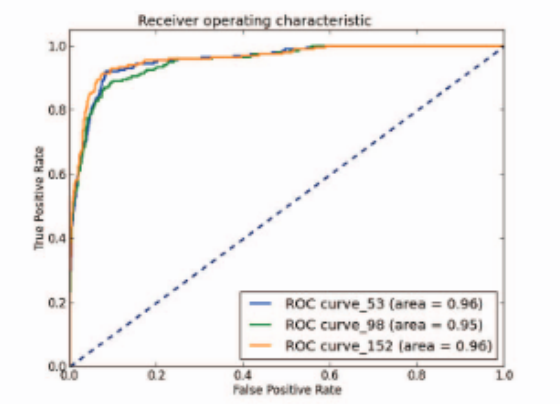


Fig. 3: ROC for models at 3 different depths

5 Discussion

5.1 Limitations

- First was acquiring access to a high quality dataset for BCC dermoscopy images, and dermoscopy images in general. Compared to models like Imagenet, which have millions of training images, they struggled to find several thousand images. This problem is not unique, and numerous other manuscripts, models based on medical imaging almost always suffer from either a lack of images or a low quality dataset.

- Due to the restrictions of their environment, re-sizing their images to 480x480 allowed them to train their models within a reasonable amount of time, and within memory limits of their GPU. However, down-sampling their image resulted in a loss of quality and features that their NN is no longer able to pick up.

- In the future, model optimization such as using non-fully connected layers in the beginning of their etwork or reducing the dimensionality of their etwork would allow them to keep the images at a higher resolution and train in a reasonable time frame.

5.2 Clinical Accuracy

- The last limitation of their model is their restricted un-derstanding of the clinical accuracy vs. quantitative dataset accuracy. The most common reason that research, just like the one presented in this paper, is not used in a clinical setting is because the metrics used to analyze a model do not fully encompass what we define as clinical accuracy.

- Their research shows an accuracy of 93%, however they cannot accurately report what the actual clinical accuracy is. Imagining their research was integrated into a clinical software tool, they may find that dermatologists may not use our software. The reason being, the model may be very good at diagnosing BCC in a lesion that is already obvious to an experienced dermatologist, but has much more false positives in complicated cases, where most of the dermatologists efforts are spent. Consider the following equations:

$$T_{wo} = T_d + T_i \tag{2}$$

$$T_{wm} = T_{fp} - T_s + T_i \tag{3}$$

where:

  - T_{wo} is the total time spent working without a model
  - T_d is the time spent diagnosing

- T_t is the time spent treating
- T_{wm} is the total time spent working with a model
- T_fp is the extra time spent reviewing false positives
- T_s, time saved using the model

Thus, they define clinical accuracy to be an accuracy that results in a statistically significant difference between T_{wm} and T_{wo} .

6 Conclusion

- In this paper They proposed a novel method for analyz-ing BCC dermoscopy images.
- By using deep ResNets their experimental results show that, on paper, the model could be used in practice as a screening tool.
- They built their model into two stages: segmentation and classification.
- Their segmentation model uses an FCRN capable of identifying a lesion in an image and eliminating extraneous information.
- Their classification model then takes this segment and analyzes it using a deep residual network 152 layers deep.
- This model works seamlessly from a single input image to a final output without any requirement of manual work.
- To improve upon this model in the future, researchers can employ more data to switch from binary classification to a categorical model, retrain the model on optical images, and possibly employ the use of transfer learning, drawing upon models pre-trained on different textures.

7 Acknowledgment

- 2018 IEEE EMBS International Conference on Biomedical Health Informatics (BHI) 4-7 March 2018 Las Vegas, Nevada, USA
- 978-1-5386-2405-0/18/$31.00 ©2018 IEEE
- This work was not supported by any organization 1 Ameer Kambod is with the Wayne State University School of Medicine, Detroit, MI 48202, USA akambod@med.wayne.edu 2 Mobeen Kambod is with the University of Michigan, Ann Arbor, MI 48109, USA mkambod@umich.edu

8 Summary

8.1 Motivation

- There is little research on BCC and neural networks, and the works provided have been insufficient thus far. To their knowledge, there is no previous work done on residual neural networks being trained on a dataset of BCC images. In this paper, they present, the first attempt to solve this problem. By using a fully convolutional residual neural network (FCRN) for segmentation and a deep residual neural network (NN) for classification, we seek to automatically diagnose a malignant lesion, providing a starting mode that can one day be used on images taken with an optical camera. Their inspiration is drawn from Lequan, et al., paper on Automated Melanoma Recognition in Dermoscopy Images via Very Deep Residual Network. They base their work on improving upon the techniques they used

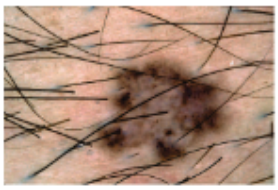
8.2 Framework

- They built their model into two stages:
  - Segmentation
  - Classification
- Their segmentation model uses an FCRN capable of identifying a lesion in an image and eliminating extraneous information.
- Their classification model then takes this segment and analyzes it using a deep residual network 152 layers deep.

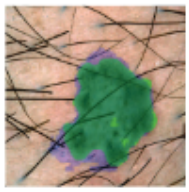
  This model works seamlessly from a single input image to a final output without any requirement of manual work.

8.3 Arguments

- Those familiar with dermoscopy may make the argument that an additional preprocessing step is required due to the surrounding structures such as hair, moles, or droplets of water created when preparing the dermoscopy images. As seen in in Figure 1,



(a) Original Image



(b) Output

Fig. 1: Segmentation Model with Hair Occlusion

an original BCC lesion is shown next to the output of their segmentation model demonstrating no irregularities when part of the lesion was occluded by hair follicles. Similar images in their dataset also suffered no loss in accuracy with occlusion, as they believe that their NN model is capable of interpreting extraneous information as non-pertinent to the lesion at hand.

8.4 Challenge

8.4.1 Solved Challenges

1. Simple neural networks that achieve a modest level of accuracy are not difficult to create; however, a system that can achieve an accuracy viable for clinical use is quite challenging and depends on several features such as: vari-ations between size and shape, contrast between where the lesions begin and end, and artifacts such as hair color and veins

2. In many deep CNNs the gradient tends to be unstable and is a fundamental issue in gradient based learning. To avoid these issues, their model draws upon the work done by He, et al., on Residual Neural Networks (Resnets).

3. Due to the nature of their dataset being a combination of two different databases, their first step was to normalize our inputs by resizing our dermoscopy images down to 480x480. This has the added benefit of allowing them to decrease training times by increasing their batch size, due to the increased free memory space

4. Due to their utilization of Keras and Theano instead of Caffe, there were several steps which had to be performed before their model could be utilized. The main focus was recognizing how Caffe and Keras differ in terms of model network setup. The most pertinent difference was that many parameters do not translate directly over from Caffe to Keras. For this they used either defaults or their best intuition for rebuilding the model e.g. using a Glorot Uniform Initializer (a.k.a Xavier uniform initializer) for their deconvolution layers.

8.4.2 Challenges Not Solved

1. Acquiring access to a high quality dataset for BCC dermoscopy images, and dermoscopy images in general. Compared to models like Imagenet, which have millions of training images, they struggled to find several thousand images. This problem is not unique, and numerous other manuscripts, models based on medical imaging almost always suffer from either a lack of images or a low quality dataset.

2. Due to the restrictions of their environment, re-sizing their images to 480x480 allowed them to train their models within a reasonable amount of time, and within memory limits of their GPU. However, down-sampling their image resulted in a loss of quality and features that their NN is no longer able to pick up.

3. The last limitation of their model is their restricted un-derstanding of the clinical accuracy vs. quantitative dataset accuracy. The most common reason that research, just like the one presented in this paper, is not used in a clinical setting is because the metrics used to analyze a model do not fully encompass what we define as clinical accuracy.

4. Model optimization such as using non-fully connected layers in the beginning of their etwork or reducing the dimensionality of their etwork would allow them to keep the images at a higher resolution and train in a reasonable time frame.

8.5 Results

Their research shows an accuracy of 93%. Their best model was at a depth of 152, in which our overall weighted sensitivity and specificity for detecting BCC from non-BCC was 97% and 96%, respectively.