Proceedings of the IASTED International Conference
Biomedical Engineering (BioMed 2017)
February 20 - 21, 2017 Innsbruck, Austria

SKIN LESION CLASSIFICATION FROM DERMOSCOPIC IMAGES USING DEEP LEARNING TECHNIQUES

Adria Romero Lopez, Xavier Giro-i-Nieto
Universitat Politecnica de Catalunya
Barcelona, Catalunya, Spain
{adria.romero@alu-etsetb., xavier.giro@}upc.edu

Jack Burdick, Oge Marques
Florida Atlantic University
Boca Raton, FL, USA
{jburdick2015, omarques}@fau.edu

ABSTRACT
The recent emergence of deep learning methods for medical image analysis has enabled the development of intelligent medical imaging-based diagnosis systems that can assist the human expert in making better decisions about a patients health. In this paper we focus on the problem of skin lesion classification, particularly early melanoma detection, and present a deep-learning based approach to solve the problem of classifying a dermoscopic image containing a skin lesion as malignant or benign. The proposed solution is built around the VGGNet convolutional neural network architecture and uses the transfer learning paradigm. Experimental results are encouraging: on the ISIC Archive dataset, the proposed method achieves a sensitivity value of 78.66%, which is significantly higher than the current state of the art on that dataset.

KEY WORDS
Medical Image Analysis, Deep Learning, Medical Decision Support Systems, Convolutional Neural Networks, Machine Learning, Skin Lesions

1 Introduction

Melanoma is a fatal form of skin cancer which is often undiagnosed or misdiagnosed as a benign skin lesion. There are an estimated 76,380 new cases of melanoma and an estimated 6,750 deaths each year in the United States [1]. Early detection is imperative: the lives of melanoma patients depend on accurate and early diagnosis. Physicians often rely on personal experience and evaluate each patients lesions on a case-by-case basis by taking into account the patient's local lesion patterns in comparison to that of the entire body [2].

Without computer-based assistance, the clinical diagnosis accuracy for melanoma detection is reported to be between 65 and 80% [3]. Use of dermoscopic images improves diagnostic accuracy of skin lesions by 49% [4]. However, the visual differences between melanoma and benign skin lesions can be very subtle (Figure 1), making it difficult to distinguish the two cases, even for trained medical experts.

For the reasons described above, an intelligent medical imaging-based skin lesion diagnosis system can be a welcome tool to assist a physician in classifying skin lesions. In this work, we are interested in a specific two-class classification problem, namely: determine whether a dermoscopic image containing a skin lesion contains a melanoma or a benign lesion.

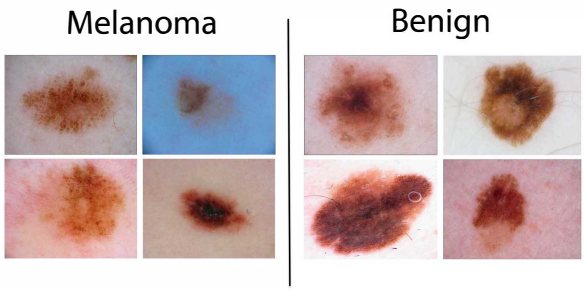

Figure 1: Sample images created from the ISIC Archive dataset [5]

In this paper, a novel method for skin lesion classification using deep learning is proposed, implemented, and successfully benchmarked against a publicly available skin lesion dermoscopic image dataset (the ISIC Archive dataset [5]). It uses an existing convolutional neural network (CNN) architecture – VGGNet (Very Deep Convolutional Network for Large-Scale Visual Recognition) developed by the Visual Geometry Group of the University of Oxford [6] in three different ways: (i) training the CNN from scratch; (ii) using the transfer learning paradigm to leverage features from a VGGNet pre-trained on a larger dataset (ImageNet [7]); and (iii) keeping the transfer learning paradigm and fine-tuning the CNNs architecture. This paper is structured as follows: Section 2 reviews related work and associated datasets and challenges; Section 3 described the proposed solution (and its variants) and associated methods and tools; Section 4 presents the results of experiments and discusses their significance; finally, Section 5 offers concluding remarks and directions for future work.

2 Background

In this section we provide a summary of relevant recent work in this field, as well as associated datasets and challenges.

DOI: 10.2316/P.2017.852-053

2.1 Related work

Most current methods in the field of melanoma classification rely on hand-crafted features, such as: lesion type (primary morphology), lesion configuration (secondary morphology), color, distribution, shape, texture, and border irregularity measures [8]. After feature extraction, machine learning methods such as k-nearest neighbors (kNN), Artificial Neural Networks (ANNs), logistic regression, decision trees and support vector machines (SVMs) can be used to perform the classification task with moderate success [9]. Examples of related work using hand-crafted features and popular classifiers include:

- Codella et al. [10] utilize hand-coded feature extraction techniques including color histogram, edge histogram, and a multi-scale variant of color local binary patterns (LBP).

- The approach proposed by Barata et al. [11] utilizes two different methods for the detection of melanoma in dermoscopy images based on global and local features. The global method uses segmentation and wavelets, Laplacian pyramids or linear filters followed by a gradient histogram are used to extract features such as texture, shape, and color from the entire lesion. After that, a binary classifier is trained from the data. The second method of local features uses a Bag of Features (BoF) classifier for image processing tasks (i.e. object recognition). Barata et al. conclude that color features perform much better than texture features alone.

More recently, the emergence of a machine learning paradigm known as deep learning has enabled the development of medical image analysis systems that can display remarkable accuracy, to the point of raising concerns about the future of the human radiologist [12][13]. Convolutional neural networks have produced promising results when classifying skin lesions. Examples of related work using deep learning include:

- The work of Kawahara et al. [14] explores the idea of using a pretrained ConvNet as a feature extractor rather than training a CNN from scratch. Furthermore, it demonstrates the use filters from a CNN pretrained on natural images generalize to classifying 10 classes of non-dermoscopic skin images.

- Liao's [15] work attempted to construct a universal skin disease classification by applying transfer learning on a deep CNN and fine-tuned its weights by continuing the backpropagation.

- In Codella et al. [10], the authors report new state-of-the-art performance using ConvNets to extract image descriptors by using a pre-trained model from the Image Large Scale Visual Recognition Challenge (ILSVRC) 2012 dataset [7]. They also investigate the most recent network structure to win the ImageNet recognition challenge called Deep Residual Network (DRN) [16].

2.2 Datasets and challenges

There are relatively few datasets in the general field of dermatology and even fewer datasets of skin lesion images. Moreover, most of these datasets are too small and/or not publicly available, which provides an additional obstacle to performing reproducible research in the area. Examples of dermatology-related image datasets used in recent research include:

- Dermofit Image Library [17] is a dataset that contains 1,300 high quality skin lesion images collected across 10 different classes.

- Dermnet [18] is a skin disease atlas with website support that contains over 23,000 skin images separated into 23 classes.

In the beginning of 2016, the International Symposium on Biomedical Imaging (ISBI) [19] released a challenge dataset for Skin lesion analysis towards melanoma detection. Photos in this dataset were obtained from the ISIC (International Skin Imaging Collaboration) [5].

3 Proposed solution

In this section we describe the selected convolutional network (ConvNet) architecture and discuss associated design choices and implementation aspects.

3.1 ConvNet architecture

VGGNet is a well documented and commonly used architecture for convolutional neural networks [6]. This ConvNet became popular by achieving excellent performance on the ImageNet [7] dataset. It comes in several variations of which the two best-performing (with 16 and 19 weight layers) have been made publicly available. In this work, the VGG16 architecture (Figure 2) was selected, since it has been shown to generalize well to other datasets. The input layer of the network expects a 224×224 pixel RGB image. The input image is passed through five convolutional blocks. Small convolutional filters with a receptive field of 3×3 are used. Each convolutional block includes a 2D convolution layer operation (the number of filters changes between blocks). All hidden layers are equipped with a ReLU (Rectified Linear Unit) as the activation function layer (nonlinearity operation) and include spatial pooling through use of a max-pooling layer. The network is concluded with a classifier block consisting of three Fully-Connected (FC) layers.

3.2 Design considerations

The original VGG16 must be modified to suit our needs, as follows:

- The final fully-connected output layer must perform a binary classification (benign vs. malignant), not 1000 classes.

- The activation function in the modified layer is modified from a softmax to sigmoidal.

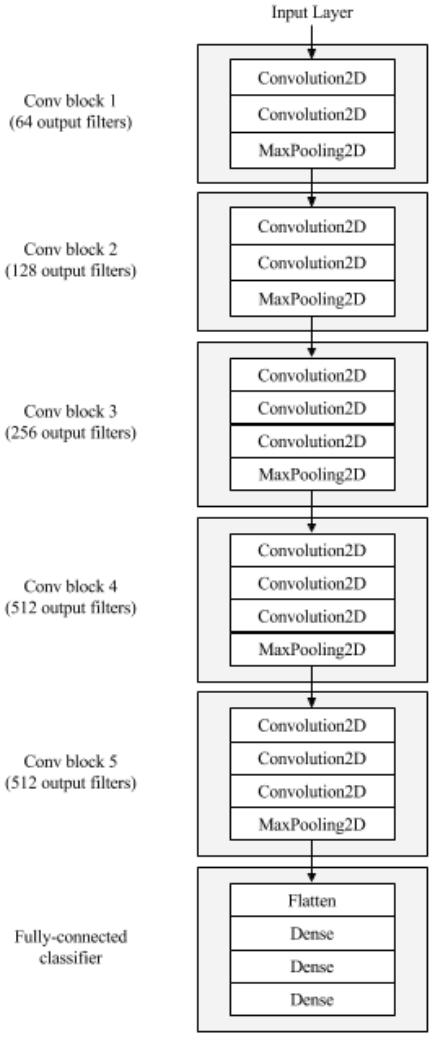

Figure 2: Original VGG16 architecture (adapted from [6])

3.2.1 Preprocessing

Input images must be preprocessed by: (i) normalizing the pixel values to a [0,1] range; (ii) cropping the image to square aspect ratio (if necessary); and (iii) resizing the image to the expected size of 224×224 pixels.

3.2.2 Data augmentation

In order to make the most of our few training examples and increase the accuracy of the model, we augmented the data via a number of random transformations. The selected data augmentation techniques were: size re-scaling, rotations of 40°, horizontal shift, image zooming, and horizontal flipping. Furthermore, it is expected that data augmentation should also help prevent overfitting (a common problem with small datasets, when the model, exposed to too few examples, learns patterns that do not generalize to new data) and, for this reason, improving the models ability to generalize.

3.3 One problem, three possible solutions

The modified VGG16 ConvNet can be used in three different ways: (i) training the ConvNet from scratch; (ii) using the transfer learning paradigm to leverage features from a pre-trained VGG16 on a larger dataset; and (iii) keeping the transfer learning paradigm and fine-tuning the ConvNets architecture. These variants (named *Method 1*, *Method 2*, and *Method 3*, respectively) are described next.

3.3.1 Method 1 - Training from scratch

The architecture is initialized with random weights and trained for a number of epochs. After each epoch, the model learns features from data and computes weights through backpropagation. This method is unlikely to produce the most accurate results if the dataset is not significantly large. However, it still can serve as a baseline for comparison against the two other methods.

3.3.2 Method 2 - ConvNet as feature extractor

Due to the relatively small number of images of skin lesion in most dermatology datasets, this method initializes the model with weights from the VGG16 trained on a larger dataset (such as ImageNet [7]), a process known as *transfer learning*. The underlying assumption behind transfer learning is that the pre-trained model has already learned features that might be useful for the classification task at hand. This corresponds, in practice, to using selected layer(s) of the pre-trained ConvNet as a fixed feature extractor, which can be achieved by freezing all the convolutional blocks and only training the fully connected layers with the new dataset.

3.3.3 Method 3 - Fine-tuning the ConvNet

Another common transfer learning technique consists of not only retraining the classifier on the top of the network with the new dataset, but also applying a fine-tuning of the network by training only the higher-level portion of the convolutional layers and continuing the backpropagation. In this work, we propose to freeze the lower level layers

of the network because they contain more generic features of the dataset. We are interested in training only the top layers of the network due to their ability to perform extraction of more specific features. In this method, the first four convolutional layers in the final architecture are initialized with weights from the ImageNet dataset. The fifth, and final, convolutional block is initialized with weights saved and loaded from the corresponding convolutional layer in Method 1.

3.3.4 Implementation aspects

Keras [20], a deep learning framework for Python, was utilized to implement the neural network architecture. Keras provides a layer of abstraction on top of Theano [21], which is used as the main neural network framework. Keras allows for: (1) modularity: users can create their network following a sequence which is a linear stack of layers; (2) minimalism: functions included in the library allow the user to create and modify network layers easily; and (3) extensibility: daily updates provide solutions to ongoing challenges faced by deep learning researchers. Moreover, Keras works on a Python environment, which gives users the freedom to use additional Python dependencies, including SciPy [22] and PIL [23].

In addition to Keras, CUDA libraries [24] were required to drive the NVidia GeForce GTX TITAN X GPUs (Graphics Processing Units) used to train and evaluate the implementation [25].

4 Experiments and Results

This section discusses the results of experiments using the proposed methods and the selected implementation.

4.1 Dataset

The ISBI 2016 Challenge dataset for Skin Lesion Analysis towards melanoma detection (described in Section 2.2) was used for our experiments.

The dataset contains a representative mix of images labeled as benign or malignant, pre-partitioned into sets of 900 training images and 379 test images [5].

4.2 Parameters

All methods were implemented in Keras. The optimizing function is RMSProp [26]. The loss function is described in [27]. A value of 0.5 is used for a dropout optimization in the fully connected layers. A batch size of 16 images is selected due to the small size of our dataset.

The dataset is balanced through undersampling. Listed alphabetically, the first 173 images from each class in the training dataset were selected and the first 75 images in each class from the testing dataset were selected. In total, the final dataset was composed of 346 training images and 150 testing images.

For data augmentation we used the following variations: size re-scaling, rotations (angles), horizontal shift, zooming (factor), and horizontal flipping.

4.3 Results

The model evaluation is performed using the same training and testing partition used in the ISIC dataset.

The metrics used are:

- *loss*, defined as the quantification of the agreement between the predicted images and the groundtruth labels;
- *sensitivity*, the fraction of true positives that are correctly identified;
- *precision*, the fraction of retrieved instances that are relevant;
- *specificity*, the fraction of true negatives that are correctly identified; and
- *accuracy*, the number of correct predictions divided by the total number of predictions.

The number of epochs for each method (chosen based on examining the behavior of accuracy/loss plots vs. number of epochs) was: 20 epochs for Method 1, 50 epochs for Method 2, and 20 epochs for Method 3.

Training and testing results for each method are shown in Tables 1 and 2, respectively. Best values are highlighted (in **bold**).

Table 1: Model evaluation: training dataset

	Loss	Accuracy	Sensitivity	Precision
M1	0.5637	71.87%	0.7087	0.6990
M2	**0.1203**	**95.95%**	**0.9621**	**0.9560**
M3	0.4891	76.88%	0.6903	0.8259

Table 2: Model evaluation: test dataset

	Loss	Accuracy	Sensitivity	Precision
M1	0.6743	66.00%	0.5799	0.6777
M2	1.0306	68.67%	0.3311	0.4958
M3	**0.4337**	**81.33%**	**0.7866**	**0.7974**

Figure 3 shows representative examples of prediction errors made by the classifier (false positives and false negatives, respectively). For contrast, Figure 4 shows examples of correct prediction results (malignant and benign, respectively).

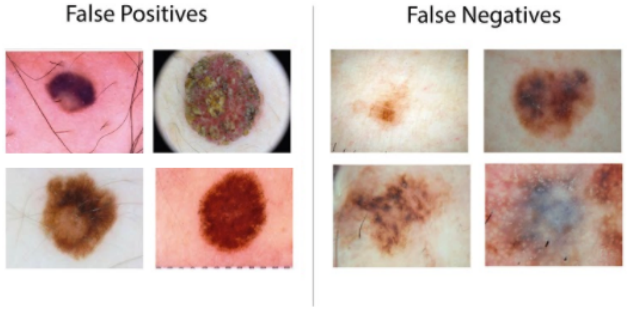

Figure 3: Examples of False Positives and False Negatives

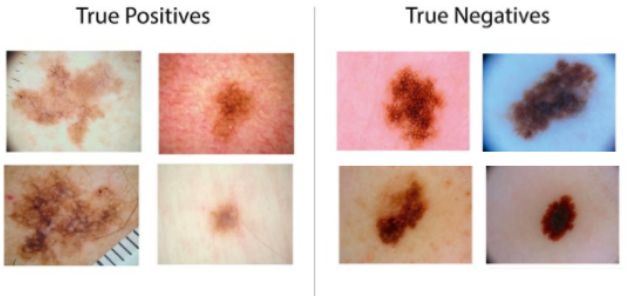

Figure 4: Examples of True Positives and True Negatives

4.4 Discussion

The results obtained using the first method are acceptable and the accuracy value is substantially above chance. Moreover, the minor difference between training and test sets suggests that the model neither overfits nor underfits.

Using the results for Method 1 as a baseline, it was not surprising to notice that Method 3 produced superior results for loss, accuracy, sensitivity, and precision in the test data. The performance of Method 2 – with exceptionally good numbers for the training set and the worst combination of loss, sensitivity, and accuracy values for the test set – presents a classical example of overfitting.

Transfer learning offers the benefit of faster epoch processing times since the layers are frozen and loaded from a previously trained network. Though decreasing processing time is prefered, the trade off is a result of the learned features potentially being unrelated to intended classification task. This observation could explain, to some extent, the inferior results of Method 2, since the ImageNet dataset is trained on 15 million labeled high resolution images from 22,000 different categories.

True		Malignant	Benign
	Malignant	59	16
	Benign	12	63
		Malignant	Benign
		Predicted	

Figure 5: Confusion Matrix for Method 3

Method 3 (whose confusion matrix appears in Figure 5) showed the best performance of all, due to reduced dependency on the ImageNet initialization weights. When using the dataset partitioned identically to the ISBI 2016 Challenge [19], it achieved an accuracy value of 81.33%, which would place the proposed approach in the top three in that challenge. Most importantly, when considering sensitivity, we achieve 78.66%, a result that is significantly better than the one reported by the current leader (50.7%). Our precision value, 79.74%, is also superior to the the current best result, of 63.7%.

In the context of medical images, *sensitivity* refers to the percent of true positives (malignant lesions) that are correctly identified whereas *specificity* measures how many samples predicted as benign (negative) are actually so. Our results for Method 3 (78.66% for sensitivity and 84.00% for specificity) are good indicators of the quality of the predictions made by the proposed model.

5 Conclusion

We propose a solution for assisting dermatologists during the diagnosis of skin lesions. More specifically, we have designed and implemented a two-class classifier that takes skin lesion images labeled as benign or malignant as an input, builds a model using deep convolutional neural networks, and uses this model to predict whether a (previously unseen) image of a skin lesion is either benign or malignant. The proposed approach achieves promising results – most notably, a sensitivity value of 78.66% and a precision of 79.74% – which are significantly higher than the current state of the art on this dataset (50.7% and 63.7%, respectively).

Avenues for future work include: (i) using a larger dataset to help lessen the risk of overfitting; (ii) performing additional regularization tweaks and fine-tuning of hyperparameters; and (iii) training the architecture with Dermnet – a skin related dataset – rather than Imagenet, a general dataset.

Acknowledgements

The authors gratefully acknowledge funding from NSF Award No. 1464537, Industry/University Cooperative Research Center, Phase II under NSF 13-542. We are also thankful to the 33 corporations that are the members of the Center for their active participation and funding.

We also thank Janet Weinthal (FAU) for her insights and assistance during the preparation of this manuscript.

References

[1] R. L. Siegel, K. D. Miller, and A. Jemal, "Cancer statistics, 2016," *CA: a cancer journal for clinicians*, vol. 66, no. 1, pp. 7–30, 2016.

[2] J. Gachon, P. Beaulieu, J. F. Sei, J. Gouvernet, J. P. Claudel, M. Lemaitre, M. A. Richard, and J. J. Grob, "First prospective study of the recognition process of melanoma in dermatological practice," *Archives of dermatology*, vol. 141, no. 4, pp. 434–438, 2005.

[3] G. Argenziano and H. P. Soyer, "Dermoscopy of pigmented skin lesions–a valuable tool for early diagnosis of melanoma," *The Lancet Oncology*, vol. 2, no. 7, pp. 443–449, 2001.

[4] H. Kittler, H. Pehamberger, K. Wolff, and M. Binder, "Diagnostic accuracy of dermoscopy," *The lancet oncology*, vol. 3, no. 3, pp. 159–165, 2002.

[5] "International Skin Imaging Collaboration: Melanoma Project Website," https://isic-archive.com/.

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

[7] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein *et al.*, "Imagenet large scale visual recognition challenge," *International Journal of Computer Vision*, vol. 115, no. 3, pp. 211–252, 2015.

[8] E. H. Page, "Description of skin lesions," https://goo.gl/m9ybFp.

[9] S. Dreiseitl, L. Ohno-Machado, H. Kittler, S. Vinterbo, H. Billhardt, and M. Binder, "A comparison of machine learning methods for the diagnosis of pigmented skin lesions," *Journal of biomedical informatics*, vol. 34, no. 1, pp. 28–36, 2001.

[10] N. Codella, Q.-B. Nguyen, S. Pankanti, D. Gutman, B. Helba, A. Halpern, and J. R. Smith, "Deep learning ensembles for melanoma recognition in dermoscopy images," *arXiv preprint arXiv:1610.04662*, 2016.

[11] C. Barata, M. Ruela, M. Francisco, T. Mendonça, and J. S. Marques, "Two systems for the detection of melanomas in dermoscopy images using texture and color features," *IEEE Systems Journal*, vol. 8, no. 3, pp. 965–979, 2014.

[12] M. Walter, "Is this the end? machine learning and 2 other threats to radiologys future," goo.gl/M9X3SF, 2016.

[13] S. Jha, "Will computers replace radiologists?" http://www.medscape.com/viewarticle/863127, 2016.

[14] J. Kawahara, A. BenTaieb, and G. Hamarneh, "Deep features to classify skin lesions," *IEEE International Symposium on Biomedical Imaging (IEEE ISBI)*, pp. 1397–1400.

[15] H. Liao, "A deep learning approach to universal skin disease classification," https://www.cs.rochester.edu/u/hliao6/projects/other/skin_project_report.pdf.

[16] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *arXiv preprint arXiv:1512.03385*, 2015.

[17] "Dermofit image library," https://licensing.eri.ed.ac.uk/i/software/dermofit-image-library.html.

[18] "Dermnet - skin disease atlas," http://www.dermnet.com/.

[19] "IEEE International Symposium on Biomedical Imaging," http://biomedicalimaging.org/.

[20] "Keras documentation," https://keras.io/.

[21] "Theano 0.8.2. documentation," http://deeplearning.net/software/theano/.

[22] "Scipy Python Library," https://www.scipy.org/.

[23] "Python Imaging Library (PIL)," http://www.pythonware.com/products/pil/.

[24] "CUDA, Nvidia," http://www.nvidia.com/object/cuda_home_new.html.

[25] "NVidia GeForce GTX TITAN X," http://www.geforce.com/hardware/desktop-gpus/geforce-gtx-titan-x.

[26] "RMSProp Optimizer," https://keras.io/optimizers/#rmsprop.

[27] "Usage of objectives," https://keras.io/objectives/.