A Deep Bag-of-Features Model for the Classification of Melanomas in Dermoscopy Images

S. Sabbaghi, M. Aldeen, *Senior Member, IEEE,* R. Garnavi, *Member, IEEE*.

***Abstract*—Deep learning and unsupervised feature learning have received great attention in past years for their ability to transform input data into high level representations using machine learning techniques. Such interest has been growing steadily in the field of medical image diagnosis, particularly in melanoma classification. In this paper, a novel application of deep learning (stacked sparse auto-encoders) is presented for skin lesion classification task. The stacked sparse auto-encoder discovers latent information features in input images (pixel intensities). These high-level features are subsequently fed into a classifier for classifying dermoscopy images. In addition, we proposed a new deep neural network architecture based on bag-of-features (BoF) model, which learns high-level image representation and maps images into BoF space. Then, we examine how using this deep representation of BoF, compared with pixel intensities of images, can improve the classification accuracy. The proposed method is evaluated on a test set of 244 skin images. To test the performance of the proposed method, the area under the receiver operating characteristics curve (AUC) is utilized. The proposed method is found to achieve 95% accuracy.**

I. Introduction

Malignant Melanoma, caused by uncontrolled growth of pigment cells, called melanocytes, is the most serious form of skin cancer. Melanoma is the most prevalent type of cancer in young Australians (15–44 year olds)[1]. It is well-established that early detection of melanoma can greatly improve the chances for successful treatment; it can be removed by simple excision.

In order to diagnose skin lesions, clinicians have been using advanced imaging technology, especially dermoscopy, to enhance the clinical diagnosis of melanoma. Dermoscopy is a non-invasive diagnostic tool that is regulrarly used in vivo clinical examination, and has proven to be useful for the early detection of malignant melanoma. The diagnosis of pigmented skin lesions using dermoscopy is based on different rules such as: ABDC rule, 7-point checklist and Menzies' method which all reported in [16]. However, even with the use of these algorithms, which aim to make the diagnosis more reliable and reproducible, clinical diagnosis of melanoma is still challenging and suffer from inter- and intra-observer variability. In the last decade, several computer aided diagnosis (CAD) systems have been proposed to tackle this problem. Some systems attempt to mimic human perceptual properties by detecting and extracting several dermoscopic visual features and structures, such as presence of certain colour or total number of colours in the lesion [3, 4], blue-white veil detection [5], pigment network [4], irregular streaks [15], and regression structures [10]. These studies have played an important role in the diagnosis of melanomas. However, identifying all dermoscopic features that are diagnostically significant and combine them into a single algorithm might not be easy to perform.

The past decade has seen the growth of the bag-of-words (BoW) model in natural language processing, and due to its simplicity and performance, this approach has become well-established in other fields[2]. Interesting results have been reported in [5] for lesion classification, based on the use of bag-of-features (BoF) approach and machine learning techniques (kNN, SVM, etc.). The approach is inspired by the bag-of-words (BoW) [2], which is a very popular feature representation model in text retrieval algorithm. The study results in [5] showed that configuration with k-NN classifier leads to better results with Sensitivity (SE) and Specificity (SP) of 100% and 75%, respectively.

Recently, Deep Neural Networks (DNNs) has emerged as a new, active area of research in machine learning and, since the first deep auto-encoder network was proposed by Hinton et al. in [7], have been reported to result in significantly improved performance on a variety of pattern-recognition tasks. Deep neural network is a neural network with multiple hidden layers. DNNs aim at learning high-level features (e.g. film genres or edges) from low-level features (e.g. pixel intensities) for differentiating objects by a classifier.

It has been hypothesised that this kind of deep architecture would allow for learning powerful object representations and ultimately obtaining more accurate classification results. For instance, in [6], the authors employed a pre-trained convolutional neural network (CNN) to create feature descriptors of skin lesions and used sparse coding for dermoscopic image representation learning. Then a non-linear SVM is employed for classifying regions of cancer and non-cancer.

Unlike CNN-based feature representation, which performs convolutional and subsampling operations to find a set of locally connected neurons through local receptive fields for feature extraction, the approach presented in this paper employs full connection of Stacked Sparse Auto-encoder (SSA), which learns a single global weight matrix for high-level representation. Through this analysis, we introduce better understanding of learned features, and also examine how the discriminative power of deep neural networks can be improved via the use of BoF. In addition, in order to show the effectiveness of proposed model a comparison is performed against one of the state of the art performance [5], which uses the BoF approach based on constructing local descriptors of images for melanoma classification.

S. Sabbaghi and M. Aldeen are with the Department of Electrical and Electronic Engineering, University of Melbourne, Australia. E-mail: sabbaghis @student.unimelb.edu.au, aldeen@unimelb.edu.au
R.Garnavi is with the IBM Research Australia. E-mail: rahilgar@au1.ibm.com.

978-1-4577-0220-4/16/$31.00 ©2016 IEEE

The remainder of this paper is organised as follows: Section II provides an overview of the proposed method. The third Section concentrates on describing the BoF model through an experiment conducted on a set of melanoma images. In Section IV, a detailed description of two different model of sparse stacked auto-encoder is explained through experiments. Dataset and experimental setup is introduced in Section V. Finally, we conclude the paper with a summary of the work.

II. Method Overview

In this paper, we intend to examine the effect of BoF in training a deep neural network. We show how the deep representation of the network becomes more discriminative when the BoF is employed for classifying the skin images. Figure 1 illustrates the overall pipeline of our proposed method. We carried out three different experiments. The first experiment is implemented based on BoF approach alone for classifying dermoscopy images. The other experiments are two variants of the auto-encoder, one with input data as raw pixel intensity image and the other with the input being BoF. In both cases the output of the auto-encoder is a vector of two classes of lesions.

III. Bag-of-Features

Bag of feature (BoF) is one of the most popular and effective classification models that has been used in several image classification tasks, as well as in melanoma application. The main idea of this model is to represent an image as a set of independent local descriptors and then quantize these descriptors as a histogram vector. Our method for creation of the BoF is based on the work reported in [5], with some modification. Since the colour features have shown great impact in classifying the dermoscopy images[3, 14], it is preferable to add colour features (number of colours in skin lesion) as descriptors for the creation of the BoF. Generally, creation of BoF consists of two main steps.

First step is obtaining a set of descriptors from a set of training images. In order to extract the set of descriptors, a dermoscopy image is divided into smaller regions (patches) by using dense sampling technique. Then the feature descriptors are computed for each of the training image patches. These feature descriptors can be SIFT (Scale Invariant Feature Transform), LBP (Local Binary Patterns), and etc. However, in our BoF implementation the descriptors are based on a combination of well-known SIFT feature proposed by [11] and colour feature introduced in our previous work [14]. The aim is to add colour information to the original SIFT.

The second step involves constructing the visual vocabulary by clustering the extracted features from the training set using k-mean algorithms. Given a new image, the feature descriptors are extracted and then assigned to the closest visual vocabulary. Therefore, each image is represented by a histogram of the frequencies of each possible word from a given vocabulary. The histogram of BoF is then used as the feature vector for training the classifier. Figure 2 shows that the same class of images (benign) can be represented with a similar histogram of the BoF. In this paper, linear SVM classifier is employed for the

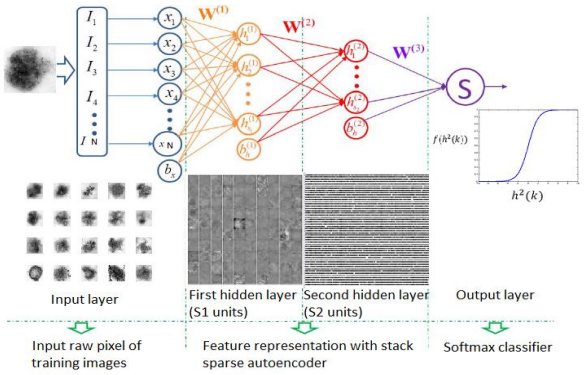


Figure1. Illustration of the architecture of Stack Sparse Auto-encoder on pixel intensity data

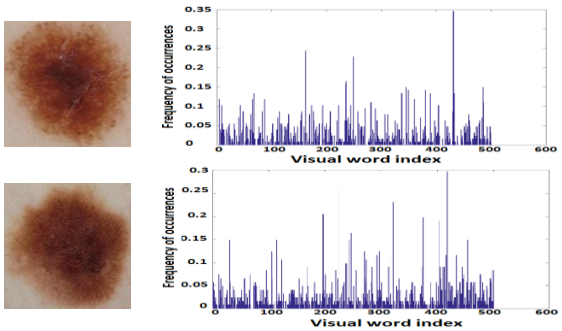


Figure 2. The visual colour word occurrences.

classification step, and the vocabulary size is set to 1000. In order to determine the ability of the method introduced in [5], we have implemented their approach on our dataset. Table I shows the performance measures for SIFT and colour features. We hypothesise that "combining different descriptors of the same class may improve the classification results". To test this hypothesis, the two descriptors (SIFT and colour) are combined, and fusion strategy is used, where the feature vectors are augmented into a single one. The results in table I show that the fusion of descriptors can improve the best classification accuracy.

IV. Stacked Sparse Auto-Encoder(SSA)

Neural networks with multiple hidden layers can be useful for solving classification problems with complex data, such as images [8]. However, training neural networks with multiple hidden layers can be difficult in practice. One way to effectively train a neural network with multiple layers is by training one layer at a time. This special type of network is known as sparse auto-encoder (SSA)[13]. Basically, the sparse auto-encoder attempts to replicate its input at its output.

In this paper we used stacked or deep auto-encoder to map images to binary codes. The stacked sparse auto-encoder (SSA) consists of two layers of basic sparse auto-encoder in which the outputs of each layer are wired to the inputs of each successive layer [13]. The first auto-encoder with a single hidden layer is trained on raw input images. Then, the learned feature representation (i.e. activations of the hidden units) is used as the input to the auto-encoder to the second layer. This learning process is repeated for subsequent layers. After layer-wise pre-training procedure is complete, the fine-tuning of the entire deep auto-encoder can be performed in a

supervised manner to improve its performance. Training a SSA involves finding the optimal parameters by minimizing the discrepancy between input and its reconstruction.

The optimal parameters are defined as a follows: s_1 and s_2 are the number of units in first and second hidden layers, λ regularization term in (4) and β the weights of the sparsity term in (4). For 570 images in the training set, we used 5-fold cross validation, where each fold consisted of 448 training and 122 validation images. Finally, the following optimal parameters are obtained based on empirical validation performance averaged across the different folds, which are: $(\lambda = 0.001, s_1 = 100, s_2 = 50, \beta = 4, \Omega = 0.1)$.

A. Training deep representation of raw input data for Skin Lesion Classification

Given a set of training images $X = [x_1, x_2, ..., x_n] \in R^d$. The training of SSA is accomplished based on optimization of a cost function, which minimizes the discrepancy between the input and its reconstruction. The SSA is composed of an encoding process and a decoding process.

During the encoding process, the input pixel intensities of the images are mapped to a new feature representation h_i through the activation function $g(\cdot)$:

$$h_i = g(W_1 x_i + b_1) \quad (1)$$

During the decoding process, the output $\hat{x_i}$ of the network is reconstructed by mapping h_i through the activation function $g(\cdot)$:

$$\hat{x_i} = g(W_2 h_i + b_2) = g(z) \quad (2)$$

The output layer has the same number of nodes as the input layer. Here, W_1 and W_2 represent the weight matrices for encoding and decoding layers, b_1 and b_2 denotes the bias vector and $g(z)$ is called activation function, which in this paper is a sigmoid function:

$$g(z) = \frac{1}{1+\exp(-z)} \quad (3)$$

Note that in the SSA learning procedure the label information is not used. Therefore, SSA learning is done through an unsupervised scheme. Finally, after the high-level feature learning procedure is completed, the learnt high-level representations of the images, as well as their labels (stored in a matrix, T) are fed to the final softmax layer to classify the dermoscopy images. Since, only two classes of classification (e.g. 0 and 1) are considered in this study, we use mean squared error function as the softmax layer in (4) to classify the lesion images.

$$E = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{k}(t_{ij} - \hat{y_{ij}})^2 + \lambda * \Omega_{weights} + \beta * \Omega_{sparsity} \quad (4)$$

Where n is the number of training example, and k is the number of classes. t_{ij} is the ij^{th} entry of the label matrix T and $\hat{y_{ij}}$ is the ith output from the auto-encoder when the input vector is pixel intensities x_j. This layer produces a value between 0 and 1 that can be interpreted as the probability of the input image being benign or malignant. The original vectors in the training data are 25600 in dimension. However, after passing them through the first encoder, the dimension is reduced to 100, and after using the second encoder, it is reduced further to 50. We then trained the final layer (softmax) to classify these 50 dimensional vector into two different classes of skin lesion. See figure 3.

TABLE I. Classification results on different local descriptors (BoF)

Features	Accuracy
SIFT	78%
Colour	80%
SIFT+Colour	85%

TABLE II. Classification results of deep auto-encoder with BoF for different folds on testset data

Fold #	Sensitivity	Specificity	Accuracy
1	94.23%	95.31%	95.08%
2	92.37%	93.70%	93.44%
3	94.23%	96.35%	95.90%
4	98.07%	93.70%	94.67%
5	98.07%	95.31%	95.90%
Average	**95.4%**	**94.9%**	**95%**

TABLE III. Comparison of results with different methods

Methods	SE	SP
Deep auto-encoder with raw input data	93.4%	92.8%
Deep auto-encoder with BoF	**95.4%**	**94.9%**
BoF [5]	93%	88%

B. Training deep representation of BoF for Skin Lesion Classification

In this section, we attempt to use the BoF as the input to the SSA to reduce the computation complexity. Also, we try to explore whether the performance of the trained network in the previous section could further be improved by utilizing this model.

The intent of our BoF based deep neural network is to map images into their corresponding BoF representation vectors. The procedure used to train the network is similar to the one described in section IV-*A*. The main difference is the input data used for feeding the auto-encoder . In here, each RGB dermoscopy image from a training set is converted to BoF model, based on the implementation described in Section III. Then, the generated BoF (SIFT+Colour) are fed into the stack auto-encoder for training. The input bag has a dimension of 10500, corresponding to number of extracted features. Finally, the SSA compress BoF vector into 2 classes through a sigmoid transforms that map onto the feature space.

V. EXPERIMENTAL SETUP

A. Dataset

A dataset of dermoscopy images obtained from the National Institutes of Health, USA [9]. The total image set consists of 814, of which 640 are benign and 174 are malignant. The images are randomly divided into two subgroups for training (570 images) and testing (244 images). They are 8-bit RGB colour images with a resolution of 768 × 560 pixels.

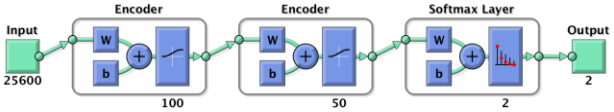


Figure 3. Semantic view of stack (deep) auto-encoders

Due to the differences in lighting conditions under which acquisition occurred, the images are not at the same level of contrast. A pre-processing step based on our previous work is then applied to enhance the contrast of images [12]. Since the input size of SSA should be fixed, all images are cropped at the center and then resized into 160×160 pixels. The labels for the images are stored in a 2-by-814 matrix T. Each column of T has a single element of 1 and the rest are 0s. The element with 1 indicates the class the corresponding lesion belongs to. Then the SSA model is employed to classify the skin lesion in the test set images. All the experiments are carried out on a PC with 3.4 GHz processor using MATLAB 2015b.

B. Results

One measure of how well the neural network has fited the data is the classification score. Sensitivity (SE) and Specificity (SP) are commonly used as evaluation metrics. Table II shows the results of deep BoF model performed on the test set data over five-fold cross validation.

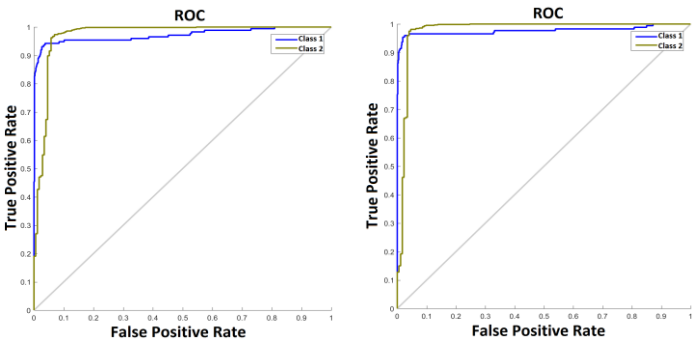
Table III summarizes the best performance achieved by our method and those by using the methods in [5]. It is concluded that deep BoF performs much better than that reported in [5], as it achieves higher classification scores with SE = 95.4% and SP = 94.9%. Also, the results confirm that using BoF as the input to the auto-encoder can easily improve the performance of neural network in comparison with the raw input images. Figure 4 compares the ROC curves of two stack auto-encoder network with different input used in the test set images. The best ROC is obtained by using deep BoF. (Fig4.b) shows the curve is closer to upper left corner.

VI. CONCLUSION

In this study, we have presented a Stacked Sparse Auto-encoder or deep framework for skin lesion classification task. The model learns a hierarchal high-level feature representation of skin image in an unsupervised manner. These high-level features enable the softmax classifier to perform efficiently for skin images classification. To show the effectiveness of the proposed framework, we compared the deep structure (Stacked Sparse Auto-encoder) with other state-of-art approach in this application. As future work, the results of this paper could be extended to use convolutional neural network (CNN) models that are popularly used in computer vision these days. Also, we intend to focus on using the proposed modeling approach to identify specific clinical patterns that may be indicative of a disease, in order to provide human verifiable evidence to support a disease diagnosis.

Acknowledgment

We thank Dr. William V. Stoecker from the Department of Dermatology, School of Medicine, University of Missouri, Columbia for providing us the NIH image set.



(a) (b)

Figure 4. (a) ROC plot of neural network with raw input image. (b) ROC plot of neural network with BoF.

REFERENCES

[1] Australia skin cancer facts and figures. 2014. p. Available at: http://www.cancer.org.au/.

[2] Baeza-Yates, R. and B. Ribeiro-Neto, *Modern information retrieval*. Vol. 463. 1999: ACM press New York.

[3] Barata, C., et al., *IMPROVING DERMOSCOPY IMAGE ANALYSIS USING COLOR CONSTANCY.*

[4] Barata, C., et al., *A system for the detection of pigment network in dermoscopy images using directional filters.* Biomedical Engineering, IEEE Transactions on, 2012. **59**(10): p. 2744-2754.

[5] Barata, C., et al., *A bag-of-features approach for the classification of melanomas in dermoscopy images: The role of color and texture descriptors*, in *Computer Vision Techniques for the Diagnosis of Skin Cancer*. 2014, Springer. p. 49-69.

[6] Codella, N., et al., *Deep Learning, Sparse Coding, and SVM for Melanoma Recognition in Dermoscopy Images*, in *Machine Learning in Medical Imaging*. 2015, Springer. p. 118-126.

[7] Hinton, G.E. and R.R. Salakhutdinov, *Reducing the dimensionality of data with neural networks.* Science, 2006. **313**(5786): p. 504-507.

[8] Hintz-Madsen, M., et al., *A probabilistic neural network framework for detection of malignant melanoma.* Artificial neural networks in cancer diagnosis, prognosis and patient management, 2001. **5**: p. 3262-3266.

[9] Kaushik V. S. N. Ghantasala, R.H.C., Uday Guntupalli, Jason R. Hagerty, Randy H. Moss, Ryan K. Rader, William V. Stoecker, *The Median Split Algorithm for Detection of Critical Melanoma Color Features.* In Proceedings of the International Conference on Computer Vision Theory and Applications (VISAPP), 2013: p. 492-495.

[10] Leo, G.D., et al. *Towards an automatic diagnosis system for skin lesions: estimation of blue-whitish veil and regression structures*. in *Systems, Signals and Devices, 2009. SSD'09. 6th International Multi-Conference on*. 2009. IEEE.

[11] Lowe, D.G., *Distinctive image features from scale-invariant keypoints.* International journal of computer vision, 2004. **60**(2): p. 91-110.

[12] Mahmouei, S., et al., *An improved colour detection method in skin lesions using colour enhancement.* Australian Biomedical Engineering Conference (ABEC 2015), 2015.

[13] Ng, A., *Sparse autoencoder.* CS294A Lecture notes, 2011. **72**: p. 1-19.

[14] Sabbaghi Mahmouei, S., et al. *Automated Colour Identification in Melanocytic Lesions*. in *Engineering in Medicine and Biology Society (EMBS), 2015 37th Annual International Conference of the IEEE*. 2015. IEEE.

[15] Sadeghi, M., et al., *Detection and analysis of irregular streaks in dermoscopic images of skin lesions.* Medical Imaging, IEEE Transactions on, 2013. **32**(5): p. 849-861.

[16] Soyer, H.P., et al., *Dermoscopy of pigmented skin lesions.* EJD, 2001. **11**(3): p. 270-276.