Deep Learning, Sparse Coding, and SVM for Melanoma Recognition in Dermoscopy Images

Noel Codella[1(✉)], Junjie Cai[1], Mani Abedini[2], Rahil Garnavi[2], Alan Halpern[3], and John R. Smith[1]

[1] IBM T.J. Watson Research Center, Yorktown Heights, NY, USA
nccodell@us.ibm.com
[2] IBM Australia Research Labs, Melbourne, VIC, Australia
[3] Memorial Sloan-Kettering Cancer Center, New York, NY, USA

Abstract. This work presents an approach for melanoma recognition in dermoscopy images that combines deep learning, sparse coding, and support vector machine (SVM) learning algorithms. One of the beneficial aspects of the proposed approach is that unsupervised learning within the domain, and feature transfer from the domain of natural photographs, eliminates the need of annotated data in the target task to learn good features. The applied feature transfer also allows the system to draw analogies between observations in dermoscopic images and observations in the natural world, mimicking the process clinical experts themselves employ to describe patterns in skin lesions. To evaluate the methodology, performance is measured on a dataset obtained from the International Skin Imaging Collaboration, containing 2624 clinical cases of melanoma (334), atypical nevi (144), and benign lesions (2146). The approach is compared to the prior state-of-art method on this dataset. Two-fold cross-validation is performed 20 times for evaluation (40 total experiments), and two discrimination tasks are examined: 1) melanoma vs. all non-melanoma lesions, and 2) melanoma vs. atypical lesions only. The presented approach achieves an accuracy of 93.1% (94.9% sensitivity, and 92.8% specificity) for the first task, and 73.9% accuracy (73.8% sensitivity, and 74.3% specificity) for the second task. In comparison, prior state-of-art ensemble modeling approaches alone yield 91.2% accuracy (93.0% sensitivity, and 91.0% specificity) first the first task, and 71.5% accuracy (72.7% sensitivity, and 68.9% specificity) for the second. Differences in performance were statistically significant ($p < 0.05$), suggesting the proposed approach is an effective improvement over prior state-of-art.

Keywords: Melanoma recognition · Dermoscopy · Dermatology · Deep learning · Sparse coding

1 Introduction

The United States saw an estimated 76,100 new cases of melanoma in 2014, and 9,710 melanoma related deaths [1]. The incidence of melanoma has doubled

L. Zhou et al. (Eds.): MLMI 2015, LNCS 9352, pp. 1–9, 2015.
DOI: 10.1007/978-3-319-24888-2_15

in a generation, and is increasing at a faster rate than any other type of solid tumor [2]. Early diagnosis is critical to combating this disease: when diagnosed in initial stages, treatments achieve a 98% 5-year survival rate. Once disease reaches lymphatics, survival rate drops to 62%. As the disease metastasizes to other areas of the body, survival drops even further to 16%. While non-invasive diagnostic methods with high sensitivity are necessary to curb the mortality rate, high-specificity is also required to prevent unnecessary medical costs, disfiguring procedures, and patient anxiety. Recent literature demonstrates that among a sampling of 20,000 skin lesions surgically excised to rule out melanoma, less than 0.1% of these tested positive for the disease [3].

Unaided visual inspection by expert dermatologists has been shown to yield diagnostic accuracy of about 60% [4]. In order to improve performance, dermoscopic imaging was introduced. Dermoscopy is a technique of placing a high-resolution magnifying imaging device in contact with the skin. Lighting is controlled, and a liquid interface or polarization filter is applied to remove surface skin reflectance, exposing underlying layers of skin to inspection. Assuming adequate levels of expertise by the interpreter, dermoscopic imaging has been shown to improve recognition performance over unaided visual inspection by approximately 50%, resulting in absolute accuracies between 75%-84% [5].

In an effort to standardize diagnostic methodologies and curb inter and intra-observer variation of dermoscopic image interpretation, procedural assessment algorithms were developed for clinicians to follow [6]. These include the ABCD rule, the 7-point checklist, the Menzies method, the CASH method, pattern analysis, and the revised pattern analysis. Among these clinical evaluation algorithms, studies have shown that pattern analysis yields better diagnostic performance over other approaches [7]. Pattern analysis involves the identification of predefined visual patterns in the lesions [6,8]. Often the descriptive terms referring to visual patterns are nicknamed in accordance with analogous entities in the natural world the patterns most resemble: i.e. “honeycomb,” “cobblestone,” or “moth eaten border”. This habit of analogous descriptions make intuitive sense, as the effectiveness of using analogies to describe and relate new knowledge to pre-existing knowledge has been well documented in education literature [9]. Studies also find that clinicians with the most experience in the field of dermoscopy tend to rely on that experience more-so than the results of any one particular analytic method [10].

Given melanoma recognition in the clinical setting has trended toward the use of pattern descriptions with analogies and expert experience, this work explores whether these same underlying principles could be used to improve the performance of automated approaches. Prior work toward automated melanoma recognition has followed classical computer vision approaches that extract hand-coded low-level visual features, combined with some form of classifier training [5,11–17]. Application of deep learning strategies, which have been successful for the task of recognition in natural photographs, have been limited by the relatively small size of the datasets. This work combines the use of deep convolutional networks trained in the domain of natural photographs, in addition

to specialized features learned via an efficient sparse coding algorithm [18], to eliminate the need of large collections of annotated data to learn good features, and allowing the system to draw analogies. Improvements in performance are demonstrated compared to previous state-of-art work.

2 Related Work

Many years of work surround the topic of melanoma recognition from dermoscopic imaging. Review articles covering a sampling of manuscripts in the most recent decade have been presented [5]. The diversity of approaches is fairly broad, but constricted within the space of classifical computer vision approaches, each work covering varying combinations of low-level visual feature extractions (color, edge, and texture descriptors, quantification of melanin based on color, etc), or machine learning techniques (kNN, SVM, etc.), and some also involving segmentation approaches [11–13]. A team from the Pedro Hispano Hospital of Portugal sought to evaluate the performance of several machine learning approaches [16,17], including global and local color, edge, and texture descriptors, training classifiers using an array of techniques, including SVM and KNN. The maximum accuracy achieved by algorithms studied was approximately 89%, on a limited dataset of 200 lesions containing 40 melanomas.

Aside from work to directly learn to recognize melanoma, there has also been work to learn to recognize specific skin lesion patterns that are indicative of melanoma, and are more easily visually verified by a clinician, such as "blue-white veil" [14,15]. While this line of modeling holds a great deal of promise to improve the accuracy of melanoma recognition, studies have again been limited by the availability of annotated data.

Recently, there has been activity to build large-scale public repository of dermoscopy data for the purposes of establishing a benchmark in the field of melanoma recognition, organized by the International Skin Imaging Collaboration (ISIC) [19]. Algorithms applied on this dataset used methods similar to other previous studies, involving a combination of low-level visual features, including color histogram, edge histogram, and a multiscale variant of color Local Binary Pattern (LBP), to achieve best performance [20]. In this work, data from the ISIC dermoscopy dataset is used for evaluation.

3 Methods

The following subsections describe the dataset used for performance evaluation, as well as the approaches for learning classifiers in this domain.

3.1 Dataset

The International Skin Imaging Collaboration (ISIC) dataset currently contains one of the largest collections of contact non-polarized dermoscopy images, complete with manual bounding boxes placed around the lesions for analysis. Examples are shown in Fig. 1. The dataset presents those cases that are most difficult

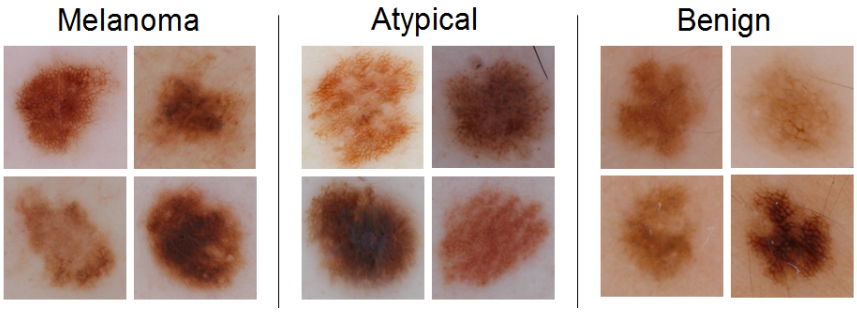

Fig. 1. Example images from the ISIC dermoscopy dataset, according to category.

for experts to distinguish, involving 334 images of melanoma and 144 images of atypical nevi, as well as 2146 clearly benign lesions (2624 total). Atypical nevi represent borderline cases: lesions that are not melanoma, but are visually similar to melanoma (as determined by expert analysis). Experiments of 2-fold cross-validation are performed 20 times (40 experiments total) for evaluation on this dataset. Two variants of the task are also performed: one task discriminating melanoma from both atypical and benign lesions (easier task), and one task discriminating melanoma from only atypical lesions (harder task).

3.2 Deep Learning Modeling Components

The presented deep learning approach uses two parallel paths: 1) transfer of convolutional neural network features learned from the domain of natural photographs, and 2) unsupervised feature learning, using sparse coding, within the domain of dermoscopy images. Classifiers are then subsequently trained for each using non-linear SVMs, and the models are then combined in late fusion (score averaging).

Convolutional Neural Network Features – The Caffe convolutional neural network (CNN) is a flexible and efficient deep learning architecture developed at Berkeley [21]. A pre-trained model from the Image Large Scale Visual Recognition Challenge (ILSVRC) 2012 is provided for download from the website. This pre-trained model includes 5 convolutional layers, 2 fully connected layers, and a final 1000 dimensional concept detector layer. In this work, the concept detector layer of this model (1000 dimensions, referred to as "FC8"), as well as the first fully connected layer (4096 dimensions, referred to as "FC6"), are used as visual descriptors for dermoscopy images.

Sparse Coding Features – Sparse coding is a class of unsupervised methods that seeks to learn a dictionary of sparse codes from which a given dataset can

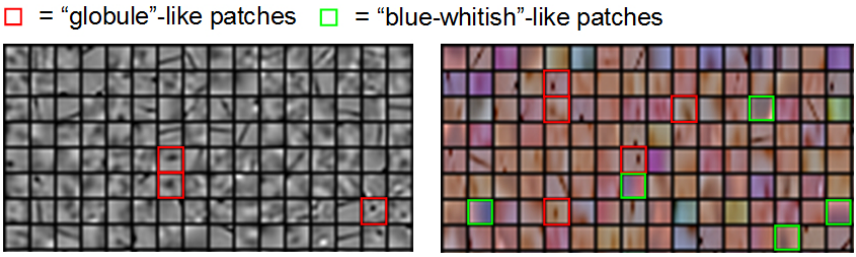

Fig. 2. Grayscale and RGB dictionary subsets learned from data. The method has identified common patterns that clinicians also search for, such as globules and blue-white structures.

be reconstructed. This is done by minimizing the following objective function:

$$\underset{D,\alpha}{min} \quad \frac{1}{n}\sum_{i=1}^{n}\left(\frac{1}{2}\left\|x_i - D\alpha_i\right\|_2^2 + \lambda\left\|\alpha_i\right\|_1\right) \tag{1}$$

Where D is the learned dictionary, α_i is the sparse representation for data sample x_i, n is the number of samples, and λ is a regularization parameter. The SPAMS sparse coding dictionary learning algorithm [18] is an online optimization approach for this objective function, based on stochastic approximations. Because of its good efficiency, the SPAMS algorithm was employed to learn dictionaries on this dataset. Two dictionaries are constructed in color (RGB) and grayscale color spaces. Images are rescaled to 128x128 pixel dimensions before extraction of 8x8 patches, to learn dictionaries of 1024 elements. A λ value of 0.15, and 1000 iterations (recommended defaults in the SPAMS implementation) were used for minimization of the objective function. Representative dictionaries are depicted in Fig. 2.

Classifier Learning – To train melanoma classifiers from various deep features under study, a non-linear SVM using a histogram intersection kernel and sigmoid feature normalization was employed. SVM scores were mapped to probabilities using logistic regression on training data [22]. A probability of 50% is used as the binary classification threshold. Fusion is done by unweighted SVM score averaging (late fusion).

Table 1. Performance Results: Melanoma vs. Atypical and Benign

	Hand Coded	Caffe CNN			Sparse Coding			Fusions	
	Ensemble	**4K FC6**	**1K FC8**	**Fusion**	**GRAY**	**RGB**	**Fusion**	**Deep**	**All**
ACC	0.912	0.919	0.853	0.910	0.825	0.903	0.907	0.923	**0.931**
SEN	0.930	0.903	0.805	0.893	0.823	0.885	0.905	0.925	**0.949**
SPE	0.910	0.921	0.860	0.912	0.825	0.906	0.907	0.923	**0.928**

Table 2. Performance Results: Melanoma vs. Atypical

	Hand Coded	Caffe CNN			Sparse Coding			Fusions	
	Ensemble	**4K FC6**	**1K FC8**	**Fusion**	**GRAY**	**RGB**	**Fusion**	**Deep**	**All**
ACC	0.715	0.723	0.654	0.725	0.651	0.681	0.695	0.728	**0.739**
SEN	0.727	0.724	0.664	0.725	0.643	0.685	0.691	0.728	**0.738**
SPE	0.689	0.722	0.632	0.723	0.670	0.673	0.706	0.729	**0.743**

3.3 Classical Modeling Approach

Low-level visual features involved in prior reports to achieve top performance in the ISIC dermatology dataset [20], as well as the ImageCLEF 2013 medical modality recognition benchmark [23], were used in this study as a comparison baseline. These include color histogram, edge histogram, a multiscale variant of color LBP [23,24], Gist, color wavelets, thumbnail vector, and various image statistics [20,23]. The strength of this approach is that features are optimally combined through an ensemble fusion algorithm – no prior assumptions about the effectiveness of a feature are made; features are tested and chosen based on performance on the data. 80% of training data was used for model learning over features, and 20% of training data was used for optimizing the late fusion of the features, in accordance with prior literature [20]. SVM scores were mapped to probabilities using logistic regression on training data [22]. A probability of 50% is used as the threshold.

4 Results

Classifier performance, in terms of the accuracy (ACC), sensitivity (SEN), and specificity (SPE), as measured by 20 experiments of two-fold cross-validation, is shown in Tables 1 and 2. The first displays results of experiments distinguishing melanoma from all non-melanoma lesions in the dataset. The second displays results of experiments distinguishing melanoma vs. atypical lesions only, a more difficult task. Each set of experiments are broken into 4 groups using different modeling approaches: ensemble models of low-level features (Ensemble), models from transferred convolutional neural network features (Caffe CNN), models from unsupervised sparse coding features (Sparse Coding), and late fusions of the previous models (Fusions).

4.1 Ensembles of Low-Level Visual Features

The "Ensemble" columns of Tables 1 & 2 show the performance results of the ensemble modeling approach over low-level visual features, which has previously achieved top performance on both the ISIC melanoma recognition dataset, and ImageCLEF medical modality recognition dataset [20,23]. This experiment serves as our baseline, to understand what the current performance standards are from the prior literature. Clearly, the task of distinguishing melanoma from only atypical lesions is the most challenging of the two tasks, as classifier accuracy drops by more than 10%.

4.2 Convolutional Neural Networks

The performance of SVMs trained on features extracted with convolutional neural networks transferred from natural photographs is shown in columns "4K FC6" (4096 dimensional fully connected FC6 layer of Caffe network), "1K FC8" (1000 dimensional concept output layer of Caffe network), and "Fusion" (late fusion of the two generated model outputs) under the "Caffe CNN" group. What is clear from this experiment is that this approach alone is on par with ensembles of low-level features. This is an important finding – the network has been optimized for natural photographs of real-world objects, which is a very different application domain to dermoscopy imaging, yet is performing similarly as well. Prior work has demonstrated transfer of deep networks between related domains [25]. In this application, the only shared quality between domains is that both are acquired from natural light, using similar image capture hardware. The content, however, is different: dermoscopy images contain no perspective distortion, all colors are restricted to those possible in skin tones, and sharp edges are scarce. Nevertheless, the features remain effective for discrimination.

4.3 Sparse Coding

The performance of classifiers trained over sparse representations of the dataset are shown in columns "GRAY" (grayscale), "RGB" (color), and "Fusion" (score averaging) under the "Sparse Coding" group. Sparse coding is a feature type that is similar to the first layer of a convolutional neural network, though the learning process is unsupervised. The sparse codes represent patterns with which the system can reconstruct the images with minimized error. Therefore, they are specifically tuned to this task and dataset. What is clear from these experiments is that color information is important to diagnosis. On average, these two features alone perform similarly well to an ensemble of several hand-coded low-level visual features, demonstrating its ability to quickly and efficiently adapt to the recognition task. The performance of the method was found to be robust to the number of sparse codes: experiments were also done for 512 and 4096 dictionary elements. The first produced similar results, though the second produced a significant performance drop, possibly due to overfitting (data not shown for brevity).

4.4 Fusions

The performance of simple late fusions of models is shown in the "Fusions" group of Tables 1 & 2. "Deep" represents the simple averaging of all Caffe and sparse coding features. "All" represents this averaging across all model types, ensembles of low-level features included, and achieved best accuracy in all tasks. What is clear from this experiment is that the combination of networks trained from natural photos, in addition to sparse codes trained directly within the task domain, leads to performance gains. The further fusion with ensembles

of low-level features brings additional performance gains, as the low-level features represent complimentary information (non-convolutional-based statistics and pattern analyses). The performance improvements are similar in both discrimination tasks of melanoma vs. all non-melanoma lesions, and melanoma vs. atypical only.

5 Conclusion

Dermatologists describe lesions using terms corresponding to natural world entities or patterns that most resemble the skin structures exhibited (such as "honeycomb" or "moth-eaten border"), as well as using specialized terms and overall experience gained for the task. This work investigates an automated method that attempts to mimic this process using convolutional neural networks trained on images of natural photographs to create feature descriptors of lesions that relate them to patterns in the natural world, combined with sparse coding representations that are highly specialized to the task. The method is compared to ensemble approaches using only hand-coded low-level features, which have been the previous state-of-art in this field. Statistically significant performance gains are observed, suggesting the proposed approach may be useful for recognizing disease. Future work may focus on using the proposed modeling approach to identify specific clinical patterns that may be indicative of disease, in order to provide human verifiable evidence to support a disease diagnosis.

References

1. Cancer Facts & Figures 2014. American Cancer Society (2014)
2. Melanoma Research Gathers Momentum. The Lancet **385**(9985), 2323
3. Oliveria, S.A., Selvam, N., Mehregan, D., Marchetti, M.A., Divan, H.A., Dasgeb, B., Halpern, A.C.: Biopsies of Nevi in Children and Adolescents in the United States, 2009 Through 2013. JAMA Dermatology, December 2014
4. Kittler, H., Pehamberger, H., Wolff, K., Binder, M.: Diagnostic accuracy of dermoscopy. The Lancet Oncology **3**(3), 159–165 (2002)
5. Abder-Rahman, A.A., Deserno, T.M.: A systematic review of automated melanoma detection in dermatoscopic images and its ground truth data. In: Proc. SPIE, Medical Imaging 2012: Image Perception, Observer Performance, and Technology Assessment, vol. 8318 (2012)
6. Braun, R.P., Rabinovitz, H.S., Oliviero, M., Kopf, A.W., Saurat, J.H.: Dermoscopy of pigmented skin lesions. J. Am. Acad. Dermatol. **52**(1), 109–121 (2005)
7. Carli, P., Quercioli, E., Sestini, S., Stante, M., Ricci, L., Brunasso, G., De Giorgi, V.: Pattern analysis, not simplified algorithms, is the most reliable method for teaching dermoscopy for melanoma diagnosis to residents in dermatology. Br. J. Dermatol. **148**(5), 981–984 (2003)
8. Rezze, G.G., Soares de Sá, B.C., Neves, R.I.: Dermoscopy: the pattern analysis. An. Bras. Dermatol. **3**, 261–268 (2006)
9. Aubusson, P.J., Harrison A.G., Ritchie S.M.: Metaphor and Analogy in Science and Education. Springer Science & Technology Education Library, vol. 30 (2006)

10. Gachon, J., et al.: First Prospective Study of the Recognition Process of Melanoma in Dermatological Practice. Arch. Dermatol. **141**(4), 434–438 (2005)
11. Garnavi, R., Aldeen, M., Bailey, J.: Computer-aided diagnosis of melanoma using border and wavelet-based texture analysis. IEEE Trans. Inf. Technol. Biomed. **16**(6), 1239–1252 (2012)
12. Ganster, H., Pinz, A., Röhrer, R., Wildling, E., Binder, M., Kittler, H.: Automated Melanoma Recognition. IEEE Transactions on Medical Imaging **20**(3) (2001)
13. Colot, O., Devinoy, R., Sombo, A., de Brucq, D.: A colour image processing method for melanoma detection. In: Wells, W.M., Colchester, A.C.F., Delp, S.L. (eds.) MICCAI 1998. LNCS, vol. 1496, p. 562. Springer, Heidelberg (1998)
14. Madooei, A., Drew, M.S., Sadeghi, M., Stella Atkins, M.: Automatic Detection of Blue-White Veil by Discrete Colour Matching in Dermoscopy Images. Medical Image Computing and Computer-Assisted Intervention, 453–460 (2013)
15. Celebi, M.E., Iyatomi, H., Stoecker, W.V., Moss, R.H., Rabinovitz, H.S., Argenziano, G., Soyer, H.P.: Automatic detection of blue-white veil and related structures in dermoscopy images. Comput. Med. Imaging Graph. **32**(8), 670–677 (2008)
16. Mendonca, T., Ferreira, P.M., Marques, J.S., Marcal, A.R., Rozeira, J.: PH2 - a dermoscopic image database for research and benchmarking. In: Conf. Proc. IEEE Eng. Med. Biol. Soc., pp. 5437–5440 (2013)
17. Barata, C., Ruela, M., et al.: Two Systems for the Detection of Melanomas in Dermoscopy Images using Texture and Color Features. IEEE Systems Journal **99**, 1–15 (2013)
18. Mairal, J., Bach, F., Ponce, J.: Sparse Modeling for Image and Vision Processing. Foundations and Trends in Computer Graphics and Vision **8**(2/3), 85–283 (2014)
19. International Skin Imaging Collaboration Website. http://www.isdis.net/index.php/isic-project
20. Abedini, M., Codella, N.C.F., Connell, J.H., Garnavi, R., Merler, M., Pankanti, S., Smith, J.R., Syeda-Mahmood, T.: A generalized framework for medical image classification and recognition. IBM Journal of Research and Development **59**(2/3) (2015)
21. Jia, Y., Shelhamer, E, Donahue, J., Karayev, S., Long, J., Girshick, R., Guadarrama, S., Darrell, T.: Caffe: Convolutional Architecture for Fast Feature Embedding (2014). arXiv preprint arXiv:1408.5093
22. Kender, J.R.: Separability and refinement of hierarchical semantic video labels and their ground truth. In: 2008 IEEE International Conference on Multimedia and Expo, pp. 673–676, 23 June 2008
23. Codella, N., Connell, J., Pankanti, S., Merler, M., Smith, J.R.: Automated medical image modality recognition by fusion of visual and text information. In: Golland, P., Hata, N., Barillot, C., Hornegger, J., Howe, R. (eds.) MICCAI 2014, Part II. LNCS, vol. 8674, pp. 487–495. Springer, Heidelberg (2014)
24. Zhu, C., Bichot, C., Chen, L.: Multi-scale color local binary patterns for visual object classes recognition. In: 20th IAPR International Conference on Pattern Recognition (ICPR), pp. 3065–3068. IEEE Press, New York (2010)
25. Yosinski, J., Clune, J., Bengio, Y., Lipson, H.: How transferable are features in deep neural networks? In: Advances in Neural Information Processing Systems, vol. 27, pp. 3320–3328 (2014)