DEEP FEATURES TO CLASSIFY SKIN LESIONS

Jeremy Kawahara, Aïcha BenTaieb, and Ghassan Hamarneh

Medical Image Analysis Lab, School of Computing Science, Simon Fraser University, Canada

ABSTRACT

Diagnosing an unknown skin lesion is the first step to determine appropriate treatment. We demonstrate that a linear classifier, trained on features extracted from a convolutional neural network pretrained on natural images, distinguishes among up to ten skin lesions with a higher accuracy than previously published state-of-the-art results on the same dataset. Further, in contrast to competing works, our approach requires no lesion segmentations nor complex preprocessing. We gain consistent additional improvements to accuracy using a per image normalization, a fully convolutional network to extract multi-scale features, and by pooling over an augmented feature space. Compared to state-of-the-art, our proposed approach achieves a favourable accuracy of 85.8% over 5-classes (compared to 75.1%) with noticeable improvements in accuracy for underrepresented classes (e.g., 60% compared to 15.6%). Over the entire 10-class dataset of 1300 images captured from a standard (non-dermoscopic) camera, our method achieves an accuracy of 81.8% outperforming the 67% accuracy previously reported.

1. INTRODUCTION

Skin cancers are commonly grouped into either melanoma or non-melanoma skin cancers. Melanoma skin cancers have a higher mortality rate, while non-melanoma skin cancers have a higher incidence rate. Early detection is important for treatment, which can differ based on the cancer type [1]. This makes systems to automatically classify types of skin lesions a potentially useful screening tool for initial referrals or as an additional supporting/safety-net expert system. As melanoma has a higher mortality rate than non-melanoma skin cancer, distinguishing between cancer and noncancerous melanoma skin images has attracted considerable research [2]. However, non-melanoma skin cancer is the most common cancer in light skin populations and, while it has a lower mortality rate than melanoma skin cancer, it places a large burden on quality of life and health care services [3]. Thus distinguishing among melanoma, non-melanoma and other types of benign skin lesions are an important component of a practical skin diagnosis tool and is a focus of this work.

Thanks to the Natural Sciences and Engineering Research Council (NSERC) of Canada for funding.

Focusing on non-melanoma skin cancers, Ballerini et al. [4] used a hierarchical K-nearest neighbors based approach to classify among 5-classes of skin lesions using images captured with a colour camera. Leo et al. [5] extended this approach to classify 10-classes of skin lesions that contained both melanoma and non-melanoma as well as benign skin lesions. This dataset [4, 5] (Fig. 1) of non-dermoscopic images is publicly available and allows us to compare methods. Shimizu et al. [6] used a similar strategy to classify among 4-classes of dermoscopic skin lesions that included both melanoma and non-melanoma lesions. These approaches [4, 5, 6] followed the similar pipeline of: preprocess the image; segment the lesion; extract a candidate set of conventional image and shape features; select a class discriminate subset of these features; and, train a classifier.

In this work, we prefer to avoid lesion segmentations and complex preprocessing as these are non-trivial steps where errors can propagate (e.g., poor segmentations gives poor features) and may require subjective human intervention. For example, the skin lesion segmentation approach of Li et al. [7] required manual initialization, post-processing, and utilized depth information. Further, they report large variations in the manual lesion segmentations done by dermatologists, which may indicate segmentations are subjective. Thus, we focus on a state-of-the-art image feature extractor, that does not require lesions' segmentations nor complex preprocessing, in the form of a pretrained fully-convolutional neural network.

Convolutional neural networks (CNNs) have emerged as a powerful classification tool and are consistently used in competitions such as the ImageNet challenge, which has researchers compete to classify hundreds of different natural objects [8]. CNNs not only give state-of-the-art results when trained for a specific task, but experiments have shown that the filters learned over the ImageNet dataset are generic and useful for other image tasks that the CNN was not originally trained for [9, 10, 11]. For example, Donahue et al. [9] used the AlexNet [12] architecture trained on ImageNet and found the responses from the first fully connected layer, FC6 (the sixth network layer), used to train a linear classifier outperformed conventional engineered features across a variety of benchmarked datasets of natural objects. Very recently, Codella et al. [10] used pretrained CNNs to extract deep features from dermoscopy images to perform 2-class classification of two tasks (melanoma vs. non-melanoma and

978-1-4799-2349-6/16/$31.00 ©2016 IEEE

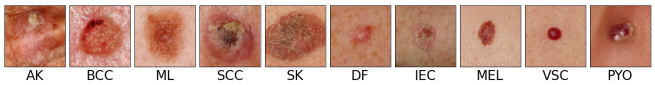

Fig. 1. Dermofit images from each of the 10-classes randomly sampled. The first 5-classes make up the 5-class experiments.

melanoma vs. atypical lesions) and found this approach to outperform conventional low-level visual features.

We convert a CNN into a fully-convolutional neural network (full-CNN) by converting the fully connected layers to convolutional layers [13]. This is an efficient approach to computing features over different spatial locations as it reuses the common convolutions done early in the network and can be used to compute features over multiple scales [13]. Aggregating features over the spatial dimensions has shown to improve predictive performance and the resulting feature vectors generalize well to other natural object image tasks [11].

Using a pretrained CNN as a feature extractor rather than training a CNN from scratch is attractive as it transfers learning (i.e. filters) from other domains where more training data is available, and avoids a time consuming training process. However, it is not obvious if the filters learned in a CNN trained on natural images will generalize well to those found on closeup skin images. Aside from recent dermoscopic work [10], most CNN-based works [9, 11] have focused on benchmarking similar natural objects. Thus we investigate whether CNN filters trained on natural objects generalize to multi-class (greater than two) classification of *non-dermoscopic* (i.e. without requiring a dermoscope) skin lesion images. We find that these features do generalize well and outperform previously published results over the same dataset, *without the aid of the corresponding lesion segmentations* used in previous approaches. We improve on the standard CNN as a feature extractor approach by using per-image normalization, a pretrained full-CNN to extract features from multiple scales, and by pooling across an augmented feature space, all of which yield classification improvements.

2. METHODS

Given a skin image x with a corresponding class label y representing the skin lesion class, we want to extract image features $f = \Phi(x)$ that discriminate well among the different class labels. To extract image features, we use the architecture of AlexNet [12] pretrained on the natural images found in ImageNet [8]. To extract features at multiple scales, we follow a similar approach to Sermanet et al. [13] to convert the CNN to a full-CNN. We convert the fully connected layers of AlexNet to convolutional layers, where these pretrained weights from the fully connected layers now act as convolutional filters. These filters can now be convolved with larger inputs (i.e. larger images) to efficiently extract responses at different scales. A skin image is passed through the network, and we extract the features from FC6 (now a convolutional layer) as FC6 has been shown to generally yield generic feature vectors [9, 10]. The responses from FC6 (i.e. the deep skin features) are used to train a logistic regression classifier to classify the skin lesions. We compute features at different scales by changing the size of the image. Thus, when the image is larger than the CNN's original receptive field, we get a feature vector with a height and width dimension, which corresponds to spatial locations in the larger input image. In order to reduce dimensionality and to be invariant to the spatial locations of the responses, we max-pool across the *spatial dimensions* (see Eq. 2) to get a single feature vector for the entire image.

Image normalization and preprocessing Typically, images are normalized by subtracting the averaged activity over the training set to center the RGB values around zero [12]. As the CNN was trained over ImageNet images, we subtract from our skin images the averaged pixel activity of the ImageNet training images. We explore other normalization approaches. To provide some invariance to differences in lighting and skin tone, we hypothesis that subtracting the mean RGB pixel values computed over each *individual* image (per-image-mean) will improve the discriminant values in the resulting feature vector. We report results over different image normalization options in Table 1. Aside from resizing images, this is the only preprocessing we perform. We contrast this simple preprocessing to other competing approaches that require more complex preprocessing such as lesion segmentation, and specular highlight removal [4, 5, 6].

Pooled deep features for augmented images A common approach to improve a CNN's classification accuracy is to augment the images [12]. As skin lesions can potentially be imaged from a variety of camera rotations, we augment the images using a rotation by 0, 90, and 270 degrees as well as a left-right flip. Given the ith image $x^{(i)}$, we augment and resize it to produce a jth augmentation of the ith image $\tilde{x}_j^{(i)}$. We normalize the augmented image and compute a feature vector by extracting the responses at FC6. For example, normalizing the image using the per-image-mean subtraction, we compute an augmented feature vector as,

$$\tilde{\boldsymbol{f}}_j^{(i)} = \Phi(\tilde{x}_j^{(i)} - \mu(\tilde{x}_j^{(i)})) \quad (1)$$

where $\mu(x)$ returns the mean value for each colour channel in x, and $\Phi(x)$ extracts the FC6 responses.

These augmented feature vectors (Eq. 1) could be used as additional samples to train a classifier and as additional image views during testing. However, there are additional time and memory costs associated with training and testing a classifier on more samples. Pooling across feature space creates a single representative feature vector for all the augmentations that allows us to keep the same time and memory benefits of having a single feature vector per image. We use a similar approach as [11], where instead of averaging across only left-right flips, we pool across m augmentations. Combining the

max-pooling over the full-CNN's spatial (height h and width w) dimensions with the mean-pooling in augmented feature space, we compute our augmented feature vector as,

$$\hat{f}_k^{(i)} = \frac{1}{m}\sum_j^m \max_{h,w}\left(\tilde{f}_{h,w,k,j}^{(i)}\right) \quad (2)$$

where $max_{h,w}\left(\tilde{f}_{h,w,k,j}^{(i)}\right)$ computes the max spatial response of the kth feature for the jth augmentation of image i. These pooled augmented feature vectors summarize the augmentations of each image, while keeping the time and memory benefits of using a lower number of training/testing samples.

3. RESULTS

We validate our approach on the Dermofit Image Library[1]. This dataset is composed of 1300 skin images with corresponding class labels and lesion segmentations. There are 10 lesion categories (Fig. 1) in this dataset: Actinic Keratosis (AK), Basal Cell Carcinoma (BCC), Melanocytic Nevus/Mole (ML), Squamous Cell Carcinoma (SCC), Seborrhoeic Keratosis (SK), Intraepithelial Carcinoma (IEC), Pyogenic Granuloma (PYO), Haemangioma (VSC), Dermatofibroma (DF), and Malignant Melanoma (MEL). As Ballerini et al. [4] report detailed results and experiments over 5 of these classes, we focus our comparison on these 5-classes (AK, BCC, ML, SCC, SK), but also benchmark over the entire 10-class dataset to compare with Leo et al. [5].

To divide our data, we follow the approach of Ballerini et al. [4], where we randomly split the dataset into three approximately equal sets with approximately the same distribution of class labels. We show cross validated results where two sets are used to train and one is held out to test. For ease of comparison, we report a single accuracy over all n (n=960 for 5-class and n=1300 for 10-class) tested images, $\frac{1}{n}\sum_i^n \delta(y_{\text{pred}} - y_{\text{true}})$ where $\delta(y_{\text{pred}} - y_{\text{true}})$ returns 1 if both the true and predicted labels are equal to each other, else 0. For a fair comparison, we compute results for [4] with this measure of accuracy using their confusion matrix (Table 1 *row a*). To better indicate the performance per class, we report the confusion matrix across all classes. Following [4], we also report the results of grouping our 5-class predictions into a 2-class problem. Specifically, we group our 5-class predictions for BCC, SCC and AK together to form a *cancer and potential risk lesion* class and group our 5-class predictions for ML and SK together to form a *benign lesion* class.

In all experiments, we train a logistic regression classifier (using the default parameters) on deep features to classify the skin lesions from a single scale. We use the Caffe [14] implementation of AlexNet [12] to extract the CNN pretrained feature vectors at FC6.

[1] https://licensing.eri.ed.ac.uk/i/software/dermofit-image-library.html

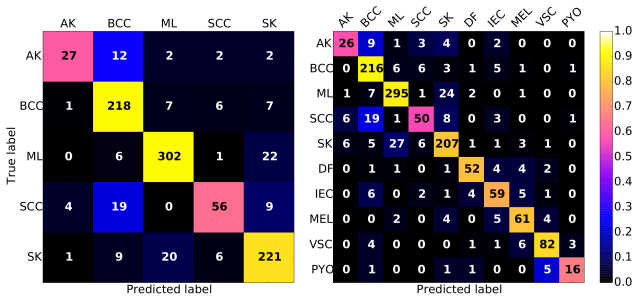

Fig. 2. Confusion matrices for our proposed approach when trained for the 5-class (*left*) and 10-class (*right*) problem. Each cell shows the number of samples predicted to belong to each class. Colours show per-class accuracy values normalized across rows by the number of images in each class.

We start by examining the effect of normalizing the images prior to extracting features. We observe that on images resized to 227×227, subtracting the dermofit-mean-pixel (subtract the mean response over our Dermofit training images) and the per-image-mean (*row c,d*) yield more accurate results than subtracting the ImageNet-pixel (*row b*). Thus, we leave out the ImageNet-pixel subtraction approach from further experiments. We also experimented with subtracting the mean ImageNet and mean Dermofit images, but neither improved results. While we found improvements to the 5-class problem when compared with [4], on this experiment we reach slightly lower 2-class accuracy than previous work.

The next set of experiments (*row e,f*) use features computed at a higher resolution. Images are resized to 339×339 and 5×5×4096 dimensional feature vectors are extracted from the full-CNN. We max-pool across the spatial domain to get a single 4096 feature vector to train our classifier. We find that the per-image-mean subtraction works slightly better than the others and thus use it for the rest of the experiments.

We examine the effect of augmentation (*row g,h*) by mean-pooling the augmented feature vectors across feature space (Eq. 2) and find this yields consistent improvements across both scales. We then examine the performance of feature vectors computed at two scales by concatenating the feature vectors (*row i*), yielding further improvements to accuracy over a single scale. We concatenated the multi-scale feature vectors (instead of pooling) in order to capture differences in lesion scales as all images are taken at roughly 50 cm [5] from the skin. Concatenating the pooled-augmented feature vectors (*row j*) yields the highest results in both the 2- and 5-class accuracy. We highlight we improve accuracy to 85.8% over the previous results of 75.1% without using segmentations. We run our proposed approach over the full 10-class dataset without any further tuning (*row l*) and find this generalizes well with an accuracy of 81.8%, outperforming the 67% accuracy reported by Leo et al. [5] (*row k*).

These above experiments indicate that deep features do

Table 1. Accuracy over all predictions. Rows with 1×1 indicate images of size 227×227 are convolved with the full-CNN and rows with 5×5 indicate images of size 339×339 were used. The plus sign (+) indicates concatenation of two feature vectors. The *aug* column indicates if image augmentation was used. The *norm* column indicates how images were normalized (e.g., subtracting the per-image-mean).

	method	aug.	norm.	5-class	2-class
(a)	[4]	✗	lesion seg.	75.1%	92.7%
(b)	1×1	✗	ImageNet-pixel	77.7%	90.6%
(c)	1×1	✗	dermofit-pixel	81.3%	92.1%
(d)	1×1	✗	per-image-mean	81.6%	92.2%
(e)	5×5	✗	dermofit-pixel	81.3%	93.1%
(f)	5×5	✗	per-image-mean	82.3%	91.9%
(g)	1×1	✓	per-image-mean	82.9%	93.0%
(h)	5×5	✓	per-image-mean	83.8%	94.7%
(i)	1×1 + 5×5	✗	per-image-mean	84.3%	93.0%
(j)	1×1 + 5×5	✓	per-image-mean	**85.8**%	**94.8**%
(k)	[5]	✗	lesion seg.	10-class = 67%	
(l)	1×1 + 5×5	✓	per-image-mean	10-class = **81.8**%	

generalize well to these skin images and outperform competing approaches [4, 5], *despite our approach not using (nor requiring) any lesion segmentations*. We found this result surprisingly remarkable as the pretrained CNN was optimized for natural images with considerably different appearance than closeup skin lesion images. A similar result was also found in the recent work of Codella et al. [10] (who reported *2-class* results over a *dermoscopy* dataset, in contrast to our *10-class* results over a *non-dermoscopy* dataset), and our findings further confirm the generalizability of pretrained CNNs to the skin domain (as opposed to work showing generalizability on more natural images [9, 11]).

We note that reporting accuracy over all images hides some large improvements to those classes with a small number of images. In particular, previous work reported 15.6% accuracy for the AK class where here we improve it to 60%. The confusion matrix for our full approach is shown in Fig. 2 showing a breakdown of accuracy by class. Finally, we highlight that our approach is fast. For a single image, the features can be extracted, augmented and classified within 0.4 seconds using a GPU implementation.

4. CONCLUSIONS

We demonstrated how filters from a CNN pretrained on natural images generalize to classifying 10 classes of non-dermoscopic skin images, outperforming previously published results. Our pipeline of using per-image-mean subtracted images, pooled-multi-scale feature extraction, and pooling across augmented feature space yielded consistent improvements to classification accuracy.

5. REFERENCES

[1] American Cancer Society, "Melanoma skin cancer," http://cancer.org/melanoma-skin-cancer-pdf, accessed: 2015-10-23.

[2] K. Korotkov and R. Garcia, "Computerized analysis of pigmented skin lesions: A review," *Artificial Intelligence in Medicine*, vol. 56, no. 2, pp. 69–90, 2012.

[3] A. Lomas, J. Leonardi-Bee, and F. Bath-Hextall, "A systematic review of worldwide incidence of nonmelanoma skin cancer," *Br. J. Dermatol.*, vol. 166, no. 5, pp. 1069–1080, 2012.

[4] L. Ballerini, R. B. Fisher, B. Aldridge, and J. Rees, "A color and texture based hierarchical K-NN approach to the classification of non-melanoma skin lesions," *Color Medical Image Analysis*, vol. 6, pp. 63–86, 2013.

[5] C. D. Leo, V. Bevilacqua, L. Ballerini, R. Fisher, B. Aldridge, and J. Rees, "Hierarchical classification of ten skin lesion classes," *Proc. Dundee Medical Image Analysis Workshop*, 2015.

[6] K. Shimizu *et al.*, "Four-class classification of skin lesions with task decomposition strategy," *IEEE TBE*, vol. 62, no. 1, pp. 274–283, 2015.

[7] X. Li *et al.*, "Depth data improves skin lesion segmentation," in *MICCAI*, vol. 5762, 2009, pp. 1100–1107.

[8] O. Russakovsky *et al.*, "ImageNet large scale visual recognition challenge," *IJCV*, vol. 115, no. 3, pp. 211–252, 2015.

[9] J. Donahue *et al.*, "DeCAF: A deep convolutional activation feature for generic visual recognition," *ICML*, vol. 32, pp. 647–655, 2014.

[10] N. Codella, J. Cai, M. Abedini, R. Garnavi, A. Halpern, and J. R. Smith, "Deep learning, sparse coding, and SVM for melanoma recognition in dermoscopy images," in *MICCAI MLMI*, vol. 9352, 2015, pp. 118–126.

[11] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *ICLR*, 2015.

[12] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *NIPS*, 2012, pp. 1097–1105.

[13] P. Sermanet *et al.*, "OverFeat: Integrated recognition, localization and detection using convolutional networks," *ICLR*, 2014.

[14] Y. Jia *et al.*, "Caffe: Convolutional architecture for fast feature embedding," *ACM Conference on Multimedia*, pp. 675–678, 2014.