Skin Disease Recognition Using Deep Saliency Features and Multimodal Learning of Dermoscopy and Clinical Images

Zongyuan Ge[1(✉)], Sergey Demyanov[1], Rajib Chakravorty[1], Adrian Bowling[2], and Rahil Garnavi[1]

[1] IBM Research, Melbourne, VIC, Australia
zongyuan@au1.ibm.com
[2] MoleMap NZ Ltd., Auckland, New Zealand

Abstract. Skin cancer is the most common cancer world-wide, among which Melanoma the most fatal cancer, accounts for more than 10,000 deaths annually in Australia and United States. The 5-year survival rate for Melanoma can be increased over 90% if detected in its early stage. However, intrinsic visual similarity across various skin conditions makes the diagnosis challenging both for clinicians and automated classification methods. Many automated skin cancer diagnostic systems have been proposed in literature, all of which consider solely dermoscopy images in their analysis. In reality, however, clinicians consider two modalities of imaging; an initial screening using clinical photography images to capture a macro view of the mole, followed by dermoscopy imaging which visualizes morphological structures within the skin lesion. Evidences show that these two modalities provide complementary visual features that can empower the decision making process. In this work, we propose a novel deep convolutional neural network (DCNN) architecture along with a saliency feature descriptor to capture discriminative features of the two modalities for skin lesions classification. The proposed DCNN accepts a pair images of clinical and dermoscopic view of a single lesion and is capable of learning single-modality and cross-modality representations, simultaneously. Using one of the largest collected skin lesion datasets, we demonstrate that the proposed multi-modality method significantly outperforms single-modality methods on three tasks; differentiation between 15 various skin diseases, distinguishing cancerous (3 cancer types including melanoma) from non-cancerous moles, and detecting melanoma from benign cases.

1 Introduction

Over 5 million skin cancer cases are diagnosed annually in America and Australia [13]. In Australia, the mean cost per patient for classification and staging of suspicious lesions (specialized surveillance and stage III in year 2) is over $3,000 [14]. Also, the availability of fully trained dermatologists worldwide is highly limited [4]. Shortage of experts and high costs make computer aided diagnosis (CAD) a necessary as an cost-effectiveness and data-driven skin disease diagnosis tool to fight against the increasing mortality of skin cancers.

M. Descoteaux et al. (Eds.): MICCAI 2017, Part III, LNCS 10435, pp. 250–258, 2017.
DOI: 10.1007/978-3-319-66179-7_29

A skin lesion is visually examined in two steps: clinical screening followed by dermoscopic analysis. Dermoscopy images are highly standardized images obtained through a high-resolution magnifying imaging device in contact with the skin. Clinical images, on the other hand, are taken by a standard digital camera and present more variations in view, angle and lighting. The majority of automated skin disease classification methods [7] could exhibit limited generalization capability when both dermoscopic and clinical modalities are being used because their domain-specific hand-crafted features are designed specifically for dermoscopy images [1]. Self feature learning scheme like deep convolutional neural networks (DCNNs) trained on very large datasets [11] has shown impressive performance in visual tasks such as object recognition and detection [12]. More importantly, those learned networks can be easily adapted to other domain tasks such as medical image segmentation [2] and skin cancer feature detection [5], all of which only cater for single image modality of dermoscopy.

To take advantage of multi-modality information embedded within dermoscopy and clinical images of the skin lesion, we develop a jointly-learned multi-modality DCNN along with a saliency-based feature descriptor to address the challenging problem of skin disease classification. The contributions of this paper are the following: (i) We propose and analyze several strategies to optimize DCNNs parameters learning of two image modalities. (ii) We propose a DCNN-based feature descriptor Class Activation Mapping-Bilinear Pooling (CAM-BP) which is able to locate saliency areas of skin images. During inference, CAM-BP assists the decision making process by producing probability maps, which improves the overall performance. (iii) We conduct comprehensive experiments and show the effectiveness of the proposed method on three diagnostic use cases: multi-class skin disease classification (across 15 disease categories), skin cancer recognition and melanoma detection.

2 Methods

In this work we explore the advantages of connecting two image modalities through a joint learning DCNN framework, and propose a novel saliency feature descriptor for multi-modality skin disease classification task. In Sect. 2.1, we first introduce two schemes for multi-modality learning (Sole-Net and Share-Net), then discuss our proposed framework Triple-Net. In Sect. 2.2, we introduce CAM-BP and explain how and why saliency information is important for discriminative feature pooling.

2.1 Cross-Modality DCNN Learning

Sole-Net: We first explore Sole-Net which is a fairly intuitive DCNN method combining information of two modalities. Each DCNN parameters are being learnt separately from each modality, and the final decision is obtained by averaging of outputs from the two trained models. The architecture of Sole-Net is illustrated in Fig. 1(a). We first denote (x_C, x_D) the pair training set where x_C

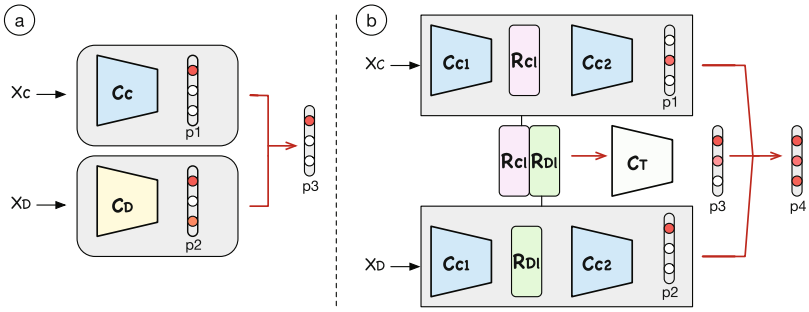

Fig. 1. Figure shows comparison of several DCNN frameworks which accept multi-modal inputs. (a) **Sole-Net**: Features from the two modalities are learnt in an dissociated manner with two separate loss functions (network blocks in two different colors). (b) **Triple-Net**: To improve upon the cross modelling ability, a new sub-network is trained on concatenated feature maps from middle layers.

and x_D are the clinical and dermoscopy images from one lesion. Each of those two DCNNs C_C and C_D contains a singe-modality learning sub-network with different parameters in different colors (blue and yellow). The cost function of each modality sub-network can be computed as[1]:

$$cost_C = ||p_C(x_C) - y_{C/D}||_2^2 \tag{1}$$

$$cost_D = ||p_D(x_D) - y_{C/D}||_2^2 \tag{2}$$

where $cost_C$ represents the cost for clinical image and $cost_D$ denotes the cost for dermoscopy image inputs. $p_C(x_C)$ and $p_D(x_D)$ ($p1$ and $p2$ in the Figure) are the outputs of each network. $y_{C/D}$ is the shared one-hot vector disease label of the observed lesion.

Share-Net: Then we explore the Share-Net where the architecture is similar to Sole-Net except C_C and C_D are sharing identical parameters. The gross cost function of Share-Net can be defined as:

$$cost_S = ||p_S(x_C) - y_{C/D}||_2^2 + ||p_S(x_D) - y_{C/D}||_2^2 \tag{3}$$

During training, the Share-Net allows its parameters across two sub-networks to be updated in a mirrored manner, The advantage of this architecture is that with the inputs of the same semantic meaning (i.e. both modalities belonging to same lesion), sharing weights across sub-networks means fewer parameters to train, which in turn means that less data required, and the model is less prone to overfitting [3].

Triple-Net: Sole-Net is capable to capturing single-modality information. However, it lacks the ability to generalize to other modalities (see Sect. 3.1). Share-net can obtain cross-modality knowledge to some extend, but it is limited by

[1] In the experiment, we observed minor overall performance difference between mean square loss and cross-entropy loss.

its capacity to learn discriminative cross-modality features because of sharing weights scheme. To exploit the merits of using cross-modality and single-modality information simultaneously, we propose Triple-Net. The proposed framework takes advantage of Sole-Net and Share-Net, but also contains extra sub-network and loss to improve discriminative cross-modality feature learning. As illustrated in Fig. 1, our proposed DCNN framework consists of three sub-networks. The first two sub-networks configure the same as the Share-Net. The third sub-network C_T takes in two corresponding convolutional feature maps R_{Cl} and R_{Dl} from a stage output (lth layer) of two sub-networks C_{Cl} and C_{Dl}. The Triple-Net has multiple cost functions and the cross-modality cost can be computed as:

$$cost_T = ||p_T(p_C^l(x_C), p_D^l(x_D)) - y_{C/D}||_2^2 \tag{4}$$

$p_{C/D}^l$ denotes the lth layer output of the network. p_T indicates the cross representation sub-network output. With the costs computed from Eq. (3) and (4), the overall Triple-Net cost is calculated as:

$$cost_{overall} = cost_S + \alpha * cost_T \tag{5}$$

where α is a hyper-parameter to setup the trade-off between single-modality and cross-modality learning rates. During prediction process, both single-modality and cross-modality are being used for decision making. The single-modality sub-network takes as $p_C(x_C) + p_D(x_D)$ an indicator for class prediction while cross-modality sub-network takes $p_T(p_C^l(x_C), p_D^l(x_D))$ as the evidence for decision. Triple-Net employs the combinations of those two.

2.2 Saliency Feature Learning

To take advantage of fine-grained information contained in the appearance of skin lesions, feature pooling method such as Bilinear Pooling (BP) [10] applied to DCNNs is a good candidate to capture fine-grained details within the image [6]. In short, it performs outer-product pair-wisely between two sub-feature maps from two DCNNs to generate distinctive representations (more details in [6,10]). However, the major disadvantage of BP is that grid-based local points are equally weighted (see Fig. 2) which leads to inability to catch saliency such as lesion area of skin images. To deal with this issue, we propose to pool BP features with spatial weights dependent on a saliency map.

Saliency map can be interpreted as the area that is most likely to belong the foreground and to contain crucial information of the image. Class activation map (CAM) is a technique to generate class activation maps using the global average pooling [15]. Each labeled category gets a class-based activation map which indicates the discriminative regions by the CNN to identify that class. CAM provides evidences which can be used to measure the probability to be a foreground object. In our proposed CAM-BP, we apply CAM as a saliency map to weight BP features. An illustration of CAM-BP is shown in Fig. 2. It can be formulated as:

$$\sum_C \frac{\sum_k w_k^c f_k(i,j)}{Z} \odot vec(f_k(i,j) f_k(i,j)^T) \tag{6}$$

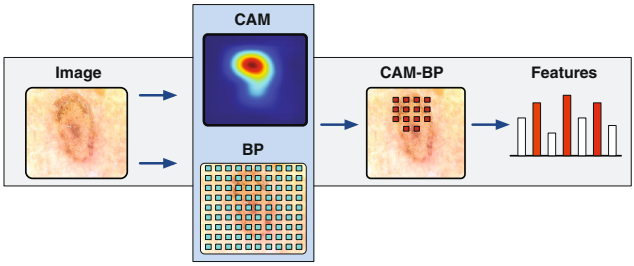

Fig. 2. Proposed saliency-based CAM-BP method: CAM activation map and BP are extracted separately. Then, the output of BP is spatially-weighted based on CAM to generate CAM-BP representation.

$f(i,j)_k \in \mathbb{R}^d$ denotes the activation of feature map k in one of the convolutional layer at location (i,j). Where w_k indicates the importance of the activation unit k at spatial location (i,j) driving to the final decision of class c. Z is a term to normalize the equation sums up to 1. The left side of element-wise production in Eq. 6 indicates how CAM is calculated and right side denotes BP. $vec()$ is the vectorization operation to compute the outer-product, thus $vec(f_k(i,j)f_k(i,j)^T) \in \mathbb{R}^{d^2}$. Average sum pooling is calculated to produce the final feature representation.

3 Experiments

Dataset: The dataset used in this work is provided by MoleMap[2]. The images are annotated by expert dermatologists with disease labels. To validate the effectiveness of our methods, we select a subset of 13,292 lesions which contains at least one image from each modality. We then randomly acquire two images from each lesion covering both modalities to prepare the dataset, resulting in 26,584 images from 15 skin conditions; 12 benign categories[3] and 3 types of skin cancer including melanoma, basal cell carcinoma and squamous cell carcinoma. We randomly partition the dataset into the ratio 7:3 for training and testing.

Network and training: We use VGG-16 CNN architecture [12] pre-trained to 92.6% top-five accuracy on the 2012 ImageNet Challenge as the base model for our evaluated frameworks. The extra sub-network in Triple-Net takes network blocks starting from the last *Conv* layer of VGG-16 and trained from scratch with batch normalization. We then use fine-tuning to optimize the parameters of the DCNNs given the amount of available training data. All layers of the network are fine-tuned with a learning rate of 0.001 and a decay factor of 0.95 every epoch.

[2] http://molemap.co.nz.

[3] Actinic Keratosis, Blue Naevus, Bowens Disease, Compound Naevus, Dermal Naevus, Dermatofibroma, Hemangioma, Junctional Naevus, Keratotic Lesion, Seborrheic Keratosis, Sebaceous Hyperplasia and Solar Lentigo.

Stochastic gradient decent (SGD) with momentum of 0.9 and decay of 5e–5 is used to train the network. During training, images are augmented with random mirroring. α in Eq. 5 is fixed to 1.5 to ensure a relatively high updating rate because of raw parameters. Following the training process as in [15], GoogLeNet is used as the base network to generate CAM and trained individually.

3.1 Analysis of Cross-Modality Learning

First, we validate the importance of cross-modality on three various of DCNNs described in Sect. 2.1 using 15-class skin disease classification task. The results are reported as overall accuracies. In this task from first two blocks of Table 1 we observe that: **(1)** Share-Net outperforms Sole-Net on both modalities, 54.1% vs 52.2% on clinical images and 55.0% vs 53.1% on dermoscopy images. **(2)** Cross-modality outputs boost the performance significantly. Compared with single-modality prediction, cross-modality predictions of Sole-Net and Share-Net results in nearly 16% and 15% improvement, respectively. **(3)** Triplet-Net outperforms Sole-Net and Share-net achieving 68.2% accuracy. Some classification samples of our proposed method are illustrated in Fig. 3.

The benefits of cross-modality learning can be further investigated in terms of swapping the modality inputs. Ideally, the performance of a well-regularised DCNN should be robust to modality swapping as the pair inputs represent the same semantic meaning (same lesion). From experimental results, we observed the performance drop is 7% less on Triple-net compared to Sole-Net, which shows that Triple-Net is more tolerable to modality swapping.

3.2 Results with CAM-BP

To conclusively evaluate the proposed CAM-BP, we apply it to both multi-modal approach of Share-Net and Triplet-Net, which reflect the generalization of this feature descriptor to various DCNNs. Figure 3 (bottom row) shows a few image samples demonstrating the effectiveness of CAM-BP in capturing

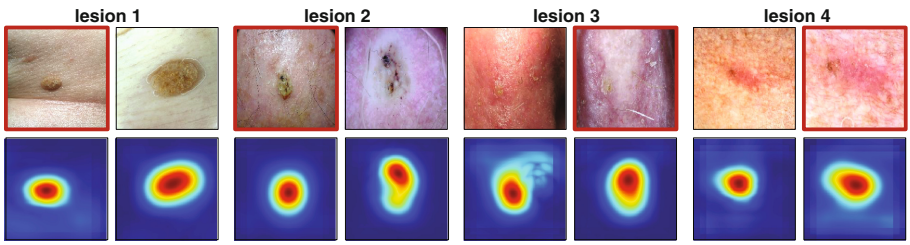

Fig. 3. The bottom row of the figure shows CAM-BP activation maps of two modalities *clinical* (left patch) vs. *dermoscopy* (right patch) for four different moles. The upper row shows samples where using only one modality has resulted in misclassification (marked in red block), but when both modalities are used in our proposed system the disease label is picked up correctly.

Table 1. Results on 15-disease classification

Methods	Modality	Accuracy
Sole-Net	Dermoscopic/Clinical	53.1%/52.2%
Share-Net	Dermoscopic/Clinical	55.0%/54.1%
Triple-Net	Dermoscopic/Clinical	**60.1%/59.4%**
Sole-Net	Cross	61.2%
Share-Net	Cross	62.9%
Triple-Net	Cross	**68.2%**
Share-Net + CAM-BP	Dermoscopic/Clinical	57.4%/58.1%
Triple-Net + CAM-BP	Dermoscopic/Clinical	61.3%/61.2%
Share-Net + CAM-BP	Cross	64.6%
Triple-Net + CAM-BP	Cross	**70.0%**

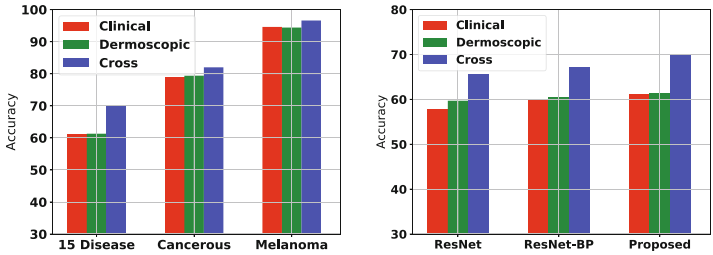

Fig. 4. Figure on the left shows our proposed method performance on three different skin disease detection tasks.

complementary saliency area from both modalities. This is important in clinical practice because visualizing the activation area provided by CAM-BP makes the model more interpretable. From last block of Table 1, the improvements across different DCNNs varies, but the overall performance improvement is consistent reaching 70% accuracy for 15-class skin disease classification.

3.3 Comparative Study and Other Detection Tasks

We have re-produced the results of the two other related DCNN-based methods modified on our image set: the residual network (ResNet) which achieved the state-of-the-arts on ImageNet 2015 challenge [9], and residual network with bilinear pooling (ResNet-BP) [8] which achieved the best performance on the ISBI 16 skin classification challenge. Figure 4 (right) shows the comparison results of our proposed method with previous competitive methods on 15 skin disease classification using single and cross modalities. Although the pre-trained network (VGG-16) being used in our method is smaller than ResNet in terms of number of

layers and parameters, we obtain 6.7% relative performance gain against ResNet-BP on 15 diseases classification task using multiple image modalities.

Moreover, we have examined the performance of our method on another two use cases including detecting 3 cancer types, and more specifically recognizing melanoma. In Fig. 4 (left), we observe that by combing two modalities, our proposed Triple-Net CAM-BP achieves impressive results on distinguishing between cancerous and non-cancerous moles with an accuracy of 82.0%, and detecting melanoma from benign lesions with 96.6% accuracy.

4 Conclusion

In this work, we demonstrate the effectiveness of cross-modality learning of DCNN for skin classification on a method accept both dermoscopy and clinical inputs. The key advantage of our method resides in two parts: (i) the use of cross-modality learning that extracts comprehensive features from sub-networks. (ii) the use of CAM-BP helps to locate the saliency area where the most important information can be retrieved, and produces discriminative features for inference.

References

1. Ballerini, L., Fisher, R.B., Aldridge, B., Rees, J.: A color and texture based hierarchical k-nn approach to the classification of non-melanoma skin lesions. In: Color Medical Image Analysis, pp. 63–86. IEEE (2013)
2. de Brebisson, A., Montana, G.: Deep neural networks for anatomical brain segmentation. In: CVPR Workshops (2015)
3. Chopra, S., Hadsell, R., LeCun, Y.: Learning a similarity metric discriminatively, with application to face verification. In: CVPR (2005)
4. Academic Grade Pay Commission: Productivity commission: Heal workforce (2014)
5. Demyanov, S., Chakravorty, R., Abedini, M., Halpern, A., Garnavi, R.: Classification of dermoscopy patterns using deep convolutional neural networks. In: ISBI (2016)
6. Gao, Y., Beijbom, O., Zhang, N., Darrell, T.: Compact bilinear pooling. In: CVPR (2016)
7. 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)
8. Ge, Z., Demyanov, S., Bozorgtabar, B., Mani, A., Chakravorty, R., Adrian, B., Garnavi, R.: Exploiting local and generic features for accurate skin lesions classification using clinical and dermoscopy imaging. In: ISBI (2017)
9. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR (2016)
10. Lin, T.Y., RoyChowdhury, A., Maji, S.: Bilinear CNN models for fine-grained visual recognition. In: ICCV (2015)
11. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al.: Imagenet large scale visual recognition challenge. Int. J. Comput. Vis. **115**(3), 211–252 (2014)

12. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition (2014). arXiv:1409.1556
13. American Cancer Society: Cancer facts & figures 2016 (2016)
14. Watts, C.G., Cust, A.E., Menzies, S.W., Mann, G.J., Morton, R.L.: Cost-effectiveness of skin surveillance through a specialized clinic for patients at high risk of melanoma. J. Clin. Oncol. **35**(1), 63–71 (2016)
15. Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., Torralba, A.: Learning deep features for discriminative localization. In: CVPR (2016)