EXPLOITING LOCAL AND GENERIC FEATURES FOR ACCURATE SKIN LESIONS CLASSIFICATION USING CLINICAL AND DERMOSCOPY IMAGING

Zongyuan Ge[⋆], *Sergey Demyanov*[⋆], *Behzad Bozorgtabar*[⋆], *Mani Abedini*[⋆], *Rajib Chakravorty*[⋆]
Adrian Bowling[†], *Rahil Garnavi*[⋆]
[⋆]IBM Research - Australia, Melbourne, VIC, Australia
[zongyuan,sergeyde,sydb,rachakra,mabedini,rahilgar]@au1.ibm.com
[†]MoleMap NZ Ltd., Auckland, New Zealand
adrian.bowling@molemap.co.nz

ABSTRACT

Similarity in appearance between various skin diseases, often makes it challenging for clinicians to identify the type of skin condition, and the accuracy is highly reliant on the level of expertise. There is also a great degree of subjectivity and inter/intra observer variability found in the clinical practices. In this paper, we propose a method for automatic skin diseases recognition that combines two different types of deep convolutional neural network features. We hold the hypothesis that it is equally important to capture global features such as color and lesion shape, as well as local features such as local patterns within the lesion area. The proposed method leverages deep residual network to represent global information, and bilinear pooling technique which allows to extract local features to differentiate between skin conditions with subtle visual differences in local regions. We have evaluated our proposed method on MoleMap dataset with 32,195 and ISBI-2016 challenge dataset with 1,279 skin images. Without any lesion localisation or segmentation, our proposed method has achieved state-of-the-art results on the large-scale MoleMap datasets with 15 various disease categories and multiple imaging modalities, and compares favorably with the best method on ISBI-2016 Melanoma challenge dataset.

Index Terms— Skin disease recognition, deep convolutional neural network (DCNN), bilinear pooling, feature fusion

1. INTRODUCTION

Skin disease causes mortalities and affects people's socio-economic life. Despite public awareness programs and efforts to address its prevalence, current statistics [1] indicate that two in three people in Australia will be diagnosed with skin cancer before the age of 70, yet 95% to 99% of all skin cancers can be easily treated if detected early. Thus, early diagnosis of skin cancer is critical for survival rates.

There have been numerous machine learning methods (SVM, kNN, etc.) applied to build an automatic system for skin lesion assessment. Traditional methods such as [2, 3, 4, 5] treated skin classification as a texture recognition problem. Most of the methods were directly transferred from the general image classification problems [6] and adapted to skin diseases classification task. Kawahara et al. [3] proposed a method to extract convolutional features from a pretrained general multi-class network, combined them with a simple linear classifier and achieved state-of-the-art result on the publicly available skin lesion dataset [7]. Codella et al. [5] presented an approach to mimic the process of how dermatologists describe patterns in skin lesions by using feature transfer and unsupervised learning within the same domain. Performances obtained from the International Skin Imaging Collaboration are state-of-the-art on two different melanoma recognition tasks.

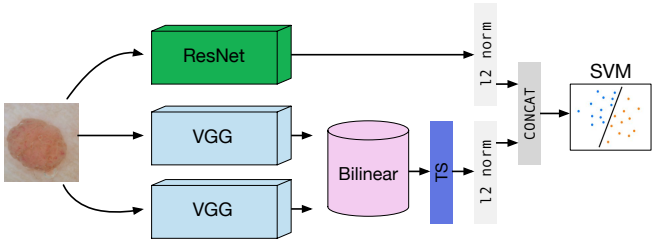

Fig. 1: Figure shows the overall architecture of our proposed framework, which uses two parallel paths to extract generic and local discriminative representations. Bilinear pooling is applied to the output of two VGG networks and Tensor Sketch (TS) is utilised to reduce the dimension of the representation. Features from two paths are normalised separately and concatenated before putting into a multi-class SVM. Blocks with the same colour share identical parameters, for example, two VGG networks in yellow boxes are duplicates to each other.

Here, we provide new perspective into skin lesion classification problem: skin lesion classification has a strong similarity with a research field called fine-grained image classification. The fine-grained image classification is defined as a problem to categorise visually similar sub-categories, for example, birds species [8] and food classification [9]. These are challenging problems due to large intra-class variations in viewpoint and appearance changes, as well as small inter-class variation due to all classes belong to the same sub-category. Commonalities can be found between skin disease recognition and fine-grained classification as most skin diseases look visually similar to each other in terms of colour,

shape and texture. Images from the same disease class might look significantly different to each other due to nuisance factors such as skin colour, noise from the background or location of the lesion area (different body areas) in the image. Those factors bring the motivation to explore and re-adapt methods from fine-grained classification field to skin lesion classification problem.

This paper examines in details how to best learn DCNN features for multi-class skin-lesion classification task. In doing so, we proposed a novel method of combining features from two different DCNNs. One of them is the deep residual network (ResNet) which claimed state-of-the-art performance on the ImageNet classification task [10]. This allows our framework to learn global features from the target dataset by using transfer learning. The other one is adapted from the fine-grained image classification task called bilinear pooling [11] which enables a network to learn discriminative local features.

2. METHODS

In this section, we first describe the architecture of the proposed method. Then we outline the details of two key components: residual network and bilinear linear pooling.

The overview of the proposed method can be seen in Figure. 1. The presented approach has two parallel paths: 1) transfer learning of deep residual network to extract global features learned from a large-scale general dataset, and 2) adapt the recently proposed bilinear DCNN approach as a local feature extractor. The local features extracted from the bilinear pooling layer are combined with the pooling layer features from the ResNet. Combining features from two sources allows us to take the advantage of information at both global and local level. The resultant features are fed into an SVM classifier.

2.1. Deep Residual Networks

Increasing the number of layers has proven to be a useful way to learn more complex representations for DCNN. Such as deep residual networks [12] which can be trained to contain more than one hundred layers and produced state-of-the-art performance on the recent ImageNet challenge [10]. The general form of the deep residual network can be represented as:

$$y^l = x^l + F(x^l, w^l) \qquad (1)$$
$$x^{l+1} = ReLU(y^l)$$

where x^l is input for the l-layer, F is a residual mapping which includes a few convolutional layers for mapping with weights w^l, and ReLU [13] is the rectifier linear function. The key idea of building a very deep neural network in ResNet is the "shortcut" path which allows a "direct" path for propagating information during training.

2.2. Bilinear Pooling

Bilinear based method with DCNN has achieved good results on several fine-grained tasks, such as face recognition [14], fine-grained image classification [11]. In one of the pooling layers, outer-product is performed at each spatial location of two pre-trained networks to generate discriminative local feature representations. The bilinear pooling can be calculated in a pooling layer as follows :

$$p_{i,j} = vec(x_{i,j} x_{i,j}^T) \qquad (2)$$

where $x_{i,j} \in \mathbb{R}^d$ is a local feature descriptor from one of the pooling layer, $p_{i,j}$ is the outer product of two vectors, $vec()$ is the vectorisation operation, and $p \in \mathbb{R}^{d^2}$.

2.3. Feature Fusion

In this work, we use ResNet and bilinear pooling as feature extractors ϕ_{res} and ϕ_{bp} respectively. For deep residual network, we take the output of the layer "pool" before the softmax classifier as feature representation. For the bilinear pooling method, the dimensionality of bilinear features normally are on the order of thousands to a few million. To reduce the dimension of the final representation for easy train and analysis, we make use of the compact bilinear pooling method proposed by Yang et al. [15], where the bilinear layer is parametrised by a pre-defined projection dimension d on the Tensor Sketch (TF) [16] method. The random projection weights can be learned as part of the end-to-end back-propagation in the network training. The final feature representation for an image I is the concatenation of the features obtained from $\phi_{res}(I)$ and bilinear pooling features $\phi_{bp}(I)$.

$$fea = (||\phi_{res}||_2^2, ||\phi_{bp}||_2^2) \qquad (3)$$

Both features are $l2$ normalised separately before employing a one-versus-all linear SVM as the classifier with the learning hyperparameter setting for C =1.

3. EXPERIMENT

To evaluate our proposed method and compare it with other state-of-the-art DCNNs, we produce results on MoleMap and ISBI-2016 challenge dataset [17] along with four deep learning baselines. These baseline methods are ResNet, ResNet with Bilinear Pooling, VGG and VGG with bilinear pooling.

VGG-16 Network (VGG). We choose to use VGG-16 model [18] as the first baseline which consists of 13 convolutional and 3 fully-connected layers. This model also serves as the base network for VGG bilinear pooling. It uses fixed small filter size (3×3) with padding to retain the same feature map size as the input signal at each layer to increase the total number of layers. In this baseline we follow a fine-tuning technique, where the network is trained on a general large-scale dataset (ImageNet), and replace the last fully-connected layer with a random initialised k-way fully-connected layer depending on the number of categories in the dataset, followed by retraining the entire model using back-propagation.

VGG Bilinear Pooling (VGG-BL). VGG-BL is the network equipped with bilinear pooling. The bilinear pooling layer is inserted after the last convolutional layer (with

(a) "Clinical" (b) "Clinical" (c) "Clinical"

(d) "Dermoscopic" (e) "Dermoscopic" (f) "Dermoscopic"

Fig. 2: Example images of two different modalities "Clinical" and "Dermoscopic". Significant variation in viewpoint due to the changed imaging conditions is evident for "Clinical" images.

ReLU). There are 512 feature kernels in the last VGG-16 convolutional layer resulting in the original bilinear feature dimension as $512 \times 512 \approx 250K$. After applying Tensor Sketch for feature dimension reduction, the output of the bilinear pooling layer is directly connected to the fully-connected layer which is the classification layer.

Residual Network (ResNet). We employ ResNet with pre-trained parameters from ImageNet models. Considering performance accuracy and computational efficiency, we use ResNet-50 as the default ResNet architecture for the rest of the experiments. Preliminary results and more detailed explanation can be found in Sec 3.2.

Residual Bilinear Pooling (ResNet-BL). We keep the ImageNet trained parameters and insert bilinear pooling layer after the last ReLU layer. The average pooling layer is removed as it is utilised for the same purpose of TF which is redundant. Like the normal ResNet, outputs from soft-max layer are directly used as final predictions.

3.1. Implementation details

Regarding feature normalisation and hyper-parameter configurations with bilinear pooling, we follow the setting provided by [15] where signed square root is applied to the output of the bilinear pooling layer. The dimension reduction ratio for TF is 32 as this provides the most appropriate results which is sufficient for reaching close-to maximum accuracy in [15].

Training. We only used mirroring technique for training data augmentation, no other augmentation techniques are used in our settings. For methods with ResNet and ResNet-Bilinear, the learning rate starts at 0.001 and is divided by 10 every 10 epochs. For methods with VGG and VGG-Bilinear, we employ a larger learning rate starting at 0.01 with the same decaying rate as ResNet. The weight decay is fixed at 0.0005 and 0.00005 for ResNet-based methods and VGG-based methods respectively. The training loss usually converges between 10 to 30 epochs depending on the model. Our code is implemented by Caffe [19] and LibLinear [20].

Table 1: VGG-BL 15-category classification results with various input resolutions on the MoleMap "Dermoscopic" modality.

Methods	Size	Mean Accuracy
VGG-BL	224/448	56.0%/63.6%

Table 2: 15-category classification results with architectures of ResNet-50, ResNet-101, ResNet-151 on the MoleMap dataset based on two imaging modalities.

ResNet Layers	Modality	Mean Accuracy
ResNet-50	Dermoscopic/Clinical	**62.8%**/62.0%
ResNet-101	Dermoscopic/Clinical	62.6%/**62.3%**
ResNet-151	Dermoscopic/Clinical	62.1%/62.0%

3.2. MoleMap

In this section, we explore the performance of our proposed method on the MoleMap dataset.

Data. The MoleMap NZ Ltd [1] dataset has been collected over the period of 2003 to 2015. The subset used in this paper contains 32,195 images of 14,754 lesions from 8,882 patients, images from 15 disease categories and spanning acrss three types of skin cancer and twelve benign disease groups, what makes it one of the largest labeled datasets of skin lesion images ever analyzed automatically and reported in the literature. The size of the images varies from 800x600 to 1600x1200 pixels. This dataset has two modalities defined as "Clinical" and "Dermoscopic". The "Clinical" is the clinical photography taken by a normal digital camera while "Dermoscopic" image is taken through a high-resolution magnifying imaging device in contact with the skin[2], see Fig. 2. We split the MoleMap dataset into 24,182 training images and 8,012 testing images. The train set has 10,725 for "Clinical" and 13,457 for "Dermoscopic" while the test set contains 3,572 "Clinical" images and 4,441 "Dermoscopic" images.

Resolution: Since image size may affect the performance on bilinear-based method [11], first we evaluate the performance with two input sizes 224×224 and 448×448 on the MoleMap "Dermoscopic" modality. The results presented in Table 1 show that higher resolution for the input size leads to a notable increase in accuracy for VGG-BL. This observation supports the use of bilinear pooling as higher input size allows catching the nuances in the local area.

Number of layers. For comparison to three different settings for ResNet 50, 101 and 152 on our task, we produced preliminary results for all those three architectures following the same setup as fine-tuning the VGG network. The results are shown in Table 2. Unlike ResNet on ImageNet, where more layers lead to better performance, the performance on skin lesion classification almost saturated with 50 layers.

Results. As can be seen in Table 3, ResNet is a strong baseline as it performs better than VGG for more than 10%

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

[2] Both clinical and dermoscopy images are recorded using Pentax EI2000, Canon G6, Dermlite DLCam.

Table 3: 15-category disease classification results on the MoleMap dataset.

Methods	Modality	**Mean Accuracy**
VGG	Dermoscopic/Clinical	53.0%/50.2%
ResNet	Dermoscopic/Clinical	63.2%/62.0%
VGG-BL	Dermoscopic/Clinical	63.6%/60.0%
ResNet-BL	Dermoscopic/Clinical	41.8%/39.7%
Proposed	Dermoscopic/Clinical	**66.4%/64.0%**
Proposed	Both	**71%**

accuracy improvement on both imaging modalities of the MoleMap dataset. Transfer learning with ResNet shows superior performance and robustness on a new task. Then, by purely adapting bilinear pooling on VGG, VGG-BL surpasses ResNet with a small margin on "Dermoscopic" but with significant less number of layers (16 vs 51). However, it performs worse on the "Clinical" protocol than ResNet which infers the less tolerant nature of viewpoint changes and background noise for bilinear pooling based network. Anecdotally, bilinear pooling severely deteriorates the performance of ResNet under both scenarios. The reason for the unsuccessful adaptation of bilinear pooling to the ResNet can be overfitting of the algorithm. As can be seen from the original paper [12], all ResNet structures face strong overfitting with deep layers, bilinear pooling equals to applying an extra polynomial kernel to the existing architecture, resulting in increased non-linearity and complexity of the model which aggravates this issue. Finally, using our proposed method yields an increase in accuracy for "Dermoscopic" from 63.2% to 66.4% compared to the ResNet. The same trend can be seen for "Clinical" from 62.0% to 64.0%. It is important to note that even though "Clinical" images show more noise and exhibit less visual details compared to the "Dermoscopic" images, the result 64% achieved by our method shows effectiveness of our method. Moreover, it is obvious that two frameworks ResNet and VGG-BL contribute complementary information to the best performed result.

To evaluate the complementary information provided by different modalities, we extended our experiments by summing the results (predicted probability of each test case) from both modalities of our proposed method together according to the lesion. It increases the overall accuracy to 71.0%. This in turn highlights the usefulness of fusing multiple modalities. In all cases, our proposed method yields the best performance, confirming that two sets of features local from bilinear pooling and generic from ResNet are both important to achieve good performance.

3.3. ISBI 2016 Challenge

Finally, we examine the performance and set-up a benchmark on the public available ISBI 2016 Skin Analysis towards Melanoma Detection challenge for other methods to compare with in the future.

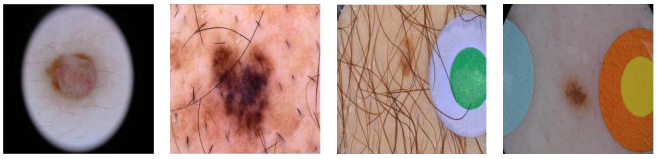

Fig. 3: Test images from ISBI dataset that are misclassified.

Table 4: Melanoma classification (2-class) accuracy on the ISBI-2016 dataset.

Methods	Modality	**Mean Accuracy**	**AP**
ResNet-50	Dermoscopic	83.6%	0.553
VGG-BL	Dermoscopic	81.2%	0.543
Proposed	Dermoscopic	**85.0%**	**0.625**

Data. The ISBI 2016 challenge dataset [17] comes from the International Skin Imaging Collaboration (ISIC) data archive and contains 1,279 contact non-polarised dermoscopy images, including 173 Melanoma and 727 Benign cases. The training and testing is split into 900 and 379.

Results. The performance of our proposed system and two other deep learning baselines are shown in Table 4. Both VGG-BL and ResNet have provided reasonable good performance over 80% on the ISBI dataset. It can be observed that the fusion of two different feature types brings additional performance gains. Our method is very competitive with state-of-the-art results on the ISBI 2016 challenge with 0.625 average prevision. In addition, we provide some insights into the test samples being misclassified. As can be seen in Fig. 3, those four images fail to be classified correctly due to several reasons. The first reason can be the noise from the background, such as the first image's surrounding (from left to right) in Fig. 3 is covered by the dark areas. The second reason is the relatively small lesion area being shown in the image like the third image. The last reason is the classifier is likely confused by the fiducial marks in the last two samples.

4. CONCLUSION

Inspired by the similarity between skin classification and fine-grained image classification, we have proposed a novel feature learning framework for skin disease classification. The proposed method incorporates two parallel deep convolutional neural networks to capture both global and local visual characteristics of skin lesions and learn discriminative features of various skin conditions. Our experiments results demonstrated that we have achieved state-of-the-art results on both MoleMap (with 15 various skin conditions) and ISBI (melanoma vs. benign nevus) datasets. Our proposed method shows great performance on different modalities such as "Dermoscopic" and "Clinical".

Acknowledgement

The authors acknowledge Alastair Sharfe, Molemap NZ Ltd. for assisting us in preparing the dataset for experiments.

5. REFERENCES

[1] Australian Institute of Health and Welfare & Australasian Association of Cancer Registries, "Cancer in australia: an overview, 2012.," *Cancer series no. 74.*, 2012.

[2] Rahil Garnavi, Mohammad Aldeen, and James Bailey, "Computer-aided diagnosis of melanoma using border- and wavelet-based texture analysis," *IEEE Transactions on Information Technology in Biomedicine*, vol. 16, no. 6, pp. 1239–1252, 2012.

[3] Jeremy Kawahara, Aıcha BenTaieb, and Ghassan Hamarneh, "Deep features to classify skin lesions," *ISBI*, 2016.

[4] Sergey Demyanov, Rajib Chakravorty, Mani Abedini, Alan Halpern, and Rahil Garnavi, "Classification of dermoscopy patterns using deep convolutional neural networks," in *ISBI*, 2016.

[5] Noel Codella, Junjie Cai, Mani Abedini, Rahil Garnavi, Alan Halpern, and John R Smith, "Deep learning, sparse coding, and svm for melanoma recognition in dermoscopy images," in *International Workshop on Machine Learning in Medical Imaging*, 2015.

[6] Jeff Donahue, Yangqing Jia, Oriol Vinyals, Judy Hoffman, Ning Zhang, Eric Tzeng, and Trevor Darrell, "DeCAF: A deep convolutional activation feature for generic visual recognition," *ICML*, 2014.

[7] Lucia Ballerini, Robert B Fisher, Ben Aldridge, and Jonathan Rees, "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. 2013.

[8] Ning Zhang, Jeff Donahue, Ross Girshick, and Trevor Darrell, "Part-based r-cnns for fine-grained category detection," in *ECCV*, 2014.

[9] ZongYuan Ge, Chris McCool, Conrad Sanderson, and Peter Corke, "Modelling local deep convolutional neural network features to improve fine-grained image classification," in *ICIP*, 2015.

[10] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al., "Imagenet large scale visual recognition challenge," *International Journal of Computer Vision*, vol. 115, no. 3, pp. 211–252.

[11] Tsung-Yu Lin, Aruni RoyChowdhury, and Subhransu Maji, "Bilinear CNN models for fine-grained visual recognition," in *ICCV*, 2015.

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," *CVPR*, 2016.

[13] Vinod Nair and Geoffrey E Hinton, "Rectified linear units improve restricted boltzmann machines," in *ICML*, 2010.

[14] Aruni RoyChowdhury, Tsung-Yu Lin, Subhranshu Maji, and Erik Learned-Miller, "Face identification with bilinear cnns," *WACV*, 2016.

[15] Yang Gao, Oscar Beijbom, Ning Zhang, and Trevor Darrell, "Compact bilinear pooling," *CVPR*, 2016.

[16] Ninh Pham and Rasmus Pagh, "Fast and scalable polynomial kernels via explicit feature maps," in *ACM SIGKDD*, 2013.

[17] David Gutman, Noel CF Codella, Emre Celebi, Brian Helba, Michael Marchetti, Nabin Mishra, and Allan Halpern, "Skin lesion analysis toward melanoma detection: A challenge at the international symposium on biomedical imaging (isbi) 2016, hosted by the international skin imaging collaboration (isic)," *eprint arXiv:1605.01397. 2016.*, 2016.

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

[19] Yangqing Jia, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long, Ross Girshick, Sergio Guadarrama, and Trevor Darrell, "Caffe: Convolutional architecture for fast feature embedding," in *ACM International Conference on Multimedia*, 2014, pp. 675–678.

[20] Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin, "Liblinear: A library for large linear classification," *Journal of machine learning research*, , no. 9, pp. 1871–1874, Aug 2008.