Aggregating Deep Convolutional Features for Melanoma Recognition in Dermoscopy Images

Zhen Yu[1], Xudong Jiang[2], Tianfu Wang[1], and Baiying Lei[1(✉)]

[1] National-Regional Key Technology Engineering Laboratory for Medical Ultrasound, School of Biomedical Engineering, Health Science Center, Shenzhen University, Shenzhen, China
leiby@szu.edu.cn

[2] School of Electric and Electronic Engineering, Nanyang Technological University, Singapore, Singapore

Abstract. We present a novel framework for automated melanoma recognition in dermoscorpy images, which is a quite challenging task due to the high intra-class and low inter-class variations between melanoma and non-melanoma (benign). The proposed framework shares merits of deep learning method and local descriptors encoding strategy. Specifically, the deep representations of a dermoscopy image are first extracted using a very deep residual neural network pre-trained on ImageNet. Then these local deep descriptors are aggregated by fisher vector (FV) encoding to build a holistic image representation. Finally, the encoded representations are classified using SVM. In contrast to previous studies with complex preprocessing and feature engineering or directly using existing deep learning architectures with fine-tuning on the skin datasets, our solution is simpler, more compact and capable of producing more discriminative features. Extensive experiments performed on ISBI 2016 Skin lesion challenge dataset corroborate the effectiveness of the proposed method, outperforming state-of-the-art approaches in all evaluation metrics.

Keywords: Dermoscopy image · Melanoma recognition · Residual network · Fisher vector

1 Introduction

Melanoma skin cancer is one of the most rapidly increasing and deadliest cancers in the world, which accounts for 79% of skin cancer deaths [1]. Early diagnosis is of great importance for treating this disease as it can be cured easily at early stages. To improve the diagnosis of this disease, dermoscopy has been introduced to assist dermatologists in clinical examination since it is a non-invasive skin imaging technique that provides clinicians high quality visual perception of skin lesion. Clinically, several heuristic approaches have been developed to enhance clinicians' ability to distinguish melanomas from benign nevi [3]. However, the correct diagnosis of a skin lesion is not trivial even for professionals. Furthermore, dermoscopic diagnosis made by human visual

Q. Wang et al. (Eds.): MLMI 2017, LNCS 10541, pp. 238–246, 2017.
DOI: 10.1007/978-3-319-67389-9_28

inspection is often subjective. Hence, unsatisfactory accuracy and poor reproducibility are still intractable issues for diagnosing this disease.

To tackle these issues, numerous automatic algorithms were proposed for der-moscopic image analysis. Interested readers can refer to [9] for a comprehensive summary of related work over the past decades. Most of the existing studies have mainly focused on feature engineering and classification on images either implicitly or explicitly, assuming a lesion object in well condition. However, dermoscopy images may not always capture entire lesions, or lesion object occupies only a small part of an image, as shown in Fig. 1. Several studies proposed to adopt the bag-of-features (BoF) model (i.e. fisher vector, FV) with local features to handle complex situations [1, 9]. Although feature encoding in BoF model has been widely used in various classification tasks, hand-crafted features based diagnostic performance is still unsatisfactory due to the high intra-class and low inter-class variations between melanoma and non-melanoma.

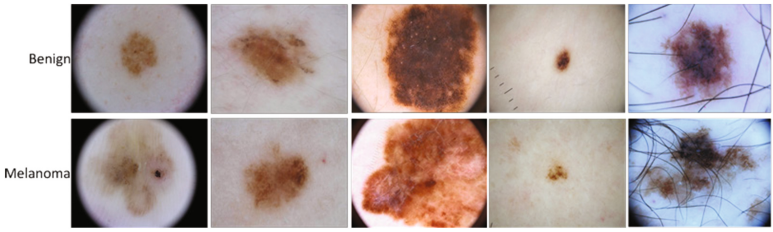

Fig. 1. Example dermoscopy image of skin lesions. There are low inter-class and high intra-class variations between the melanoma and non-melanoma (benign). The diagnosis of melanoma is non-trivial even for experienced clinicians.

Different from approaches that rely on the hand-crafted features, study in [12] demonstrated that transferred convolutional features can be utilized as generic visual representations, and convolutional neural network (CNN) architectures pre-trained on large ImageNet dataset also delivered promising results for other image recognition tasks even without retraining. To the end, transferred CNN features have also been applied in dermoscopy image classification in recent years [3, 4, 8]. By default, deep convolutional features are extracted from fully connected layers of a CNN model. Nevertheless, high-level CNN features are sensitive or vulnerable to geometric variations because they suffer from the paucity of descriptions of local patterns [2, 15]. For images with dramatic variations in viewpoint and scale, it would be a great challenge to perform classification directly using CNN features, not mention that in medical applications, these features are usually trained on limited training data. Several studies [5, 14, 15] were devoted to combining deep features with local descriptors encoding methods to beef up their discrimination capability and robustness. Although impressive improvement is achieved in some benchmarks, these approaches are highly computational intensive due to the adoption of sliding-windows to generate deep descriptors from local regions in original images or multi-scale pyramid pooling strategy to construct FV representations.

Motivated by [2, 17], in this work, we propose a novel, compact, and efficient framework based on very deep CNN and feature encoding strategy (FV encoding) for melanoma recognition in dermoscopy image (see Fig. 2). A very deep residual neural net-work (i.e. 50 layers) [7] pre-trained on ImageNet, is first applied to each input image, and then local deep descriptors are extracted from the dense activation maps of the last convolutional layer. These features are further encoded by FV into more invariant and discriminative representations. In addition, large scale of input images and special per image normalization are utilized to gain additional performance improvement. Experimental results demonstrate the effectiveness of our proposed approach.

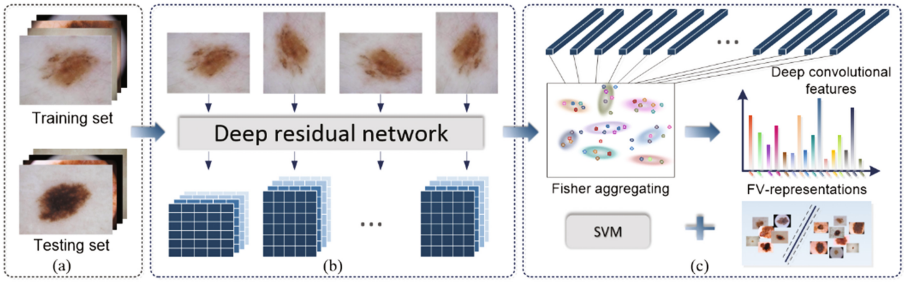

Fig. 2. Flowchart of proposed framework for melanoma recognition. (a) Dataset of dermoscopic lesion images. (b) Data augmentation and feature extraction. Local activations of intermediate layer are extracted as deep feature vectors. (c) FV encoding and classification.

2 Methodology

2.1 Image Preprocessing and Data Augmentation

Image preprocessing affects the performance of deep representations greatly as it takes the image characteristics into consideration. There is a huge variation in resolution of the skin lesion images dataset provided by ISBI 2016 challenge [6]. Resizing and cropping these images directly into required input size of CNN models may introduce object distortion and substantial information loss. Accordingly, in this study, we take relatively large images as inputs. For the skin lesion dataset, we resize each input image along the shortest side to a uniform scale (denoted as S for simplicity) while maintaining the aspect ratio. We also investigate the recognition performance with various values of S.

Typically, before processing by CNN, images are normalized by subtracting the mean pixel value calculated over the entire training dataset (denoted as all-img-mean). As a result, the RGB values are centered at zero. However, the lighting, skin tone and viewpoint of the skin lesion images vary greatly across the dataset, subtracting a uniform mean value does not well normalize individual image. Recent study [8] has illustrated this effect as well. To address this issue, we normalize each skin image by subtracting channel-wise average intensity values calculated over the individual image

(denoted as per-img-mean). For the data augmentation, we rotate each resized image by four fixed angles (0°, 90°, 180°, and 270°), and then pixel translation (with shift between −10 and 10 pixels) is randomly added over the rotated images. The deep features of these augmented images are aggregated into a single FV representation.

2.2 Extraction of Local Convolutional Features

Given a pre-trained network, an input skin lesion image $\mathcal{X}_i$ is first processed by the above-mentioned augmentation procedure, and thus we obtain four augmented images $\mathbb{X}_i = \{\mathcal{X}_{i1}, \mathcal{X}_{i2}, \mathcal{X}_{i3}, \mathcal{X}_{i4}\}$ for each skin image. These images are passed through the CNN model in a forward pass. In the l-th convolutional layer $\mathcal{L}_l$, we obtain $w_{ia}^l \times h_{ia}^l \times d^l$ spatial feature maps $\mathcal{M}_{ia}^l (a = 1, \ldots, 4)$, where w_{ia}^l and h_{ia}^l denote the width and height, respectively, d^l is the depth or channels of the current feature map. For brevity, we denote $\mathcal{N}_{ia}^l = w_{ia}^l \times h_{ia}^l$. It is worth noting that, for input images with different sizes, the size of the resulting feature maps can be different. Similar to [17], for activations at each location $c = (c_x, c_y)$, $1 \leq c_x \leq w_{ia}^l$ and $1 \leq c_y \leq h_{ia}^l$ in the feature map $\mathcal{M}_{ia}^l$, we obtain d^l-dimensional vector $f_{ia,c}^l \in \mathbb{R}^{d^l}$ which is considered as feature vector (local deep feature) in our study. Therefore, $\mathbb{N}_{ia}^l$ local deep feature vectors are obtained for each augmented image $\mathcal{X}_{ia}$. For i-th original skin lesion image $\mathcal{X}_i$, at convolutional layer $\mathcal{L}_l$ of the network, we obtain a set of deep features:

$$\mathbb{F}_i^l = \left\{ f_{i1,(1,1)}^l, \ldots f_{i4,(w_{i4}^l, h_{i4}^l)}^l \right\} \in \mathbb{R}^{\sum_{a=1}^{4} \mathcal{N}_{ia}^l \times d^l}. \tag{1}$$

These features are encoded by FV into single representation for the final classification. In our study, considering the transferability and discrimination of the deep features, the output of last convolutional layer of the ResNet is adopted as local features.

2.3 Fisher Vector Encoding Strategy

Each local deep convolutional feature f_n^l extracted from layer $\mathcal{L}_l$, refers to a small region (receptive field) in the input image, and reflects the local distinction of that region. This is similar to traditional local descriptors. Since each image contains a set of deep features, and thus we propose to aggregate these local deep representations into a holistic representation using FV encoding. The FV encoding derived from fisher kernel is effective for encoding local features and has demonstrated excellent performance in image recognition [11].

To implement FV encoding, the popular Gaussian mixture model (GMM) is adopted to model the probability distribution of deep features. For the purpose of constructing a GMM with K components, a collection of skin images are sampled from the training set. The local deep descriptors of these images are then extracted and utilized to learn the parameters $\lambda = \{\pi_k, \mu_k, \sum_k, k = 1, 2, \ldots, K\}$, which includes the prior probability $\pi_k \in \mathbb{R}_+$ subjected to the constraint $\sum_1^K \pi_k = 1$, mean vector $\mu_k \in \mathbb{R}^{d^l}$, and covariance matrix $\sum_k \in \mathbb{R}^{d^l \times d^l}$ constrained to be diagonal. For a set of

local deep features $\{\mathcal{F}_{i1}^{l}, \mathcal{F}_{i2}^{l}, \mathcal{F}_{i3}^{l}, \mathcal{F}_{i4}^{l}\}$ extracted from the augmented images regarding i-th skin lesion image, the first and second order differences of the GMM clusters are given by:

$$\omega_k = \frac{1}{N\sqrt{\pi_k}} \sum_{n=1}^{N} \gamma_{kn} \left(\frac{f_n^l - \mu_k}{\sum_{\boldsymbol{k}}^{1/2}} \right), \tag{2}$$

$$\upsilon_k = \frac{1}{N\sqrt{2\pi_k}} \sum_{n=1}^{N} \gamma_{kn} \left[\frac{\left(f_n^l - \mu_k\right)^2}{\sum_{\boldsymbol{k}}} - 1 \right], \; k = 1, 2 \ldots K, \tag{3}$$

where $N = \sum_{a=1}^{4} \mathcal{N}_{ia}^{l} \times d^{l}$ represents the number of local deep descriptors of a skin image; γ_{kn} denotes the soft-assignment of a certain feature vector f_n^l to cluster k. By concatenating ω_k and υ_k for all K components, we obtain the final FV representation Φ_i:

$$\Phi_i = \left[\omega_1^T, \upsilon_1^T, \ldots \omega_K^T, \upsilon_K^T \right]^T. \tag{4}$$

It is noteworthy that the dimensionality of deep feature vector is reduced by principle component analysis (PCA) before FV encoding because more Gaussian components are needed to capture the distribution of higher dimensional feature. For each FV representation, we further compute the improved FV by applying L2 and power normalization the same as that in [11].

2.4 Kernel-Based Classification

For the classification of the FV representations, we train a SVM classifier with Chi-squared (chi2) kernel. Although linear kernels are efficient for the classification, non-linear kernels tend to yield better performance and empirical studies have demonstrated the superiority of the chi2 kernel for image classification [11]. Prior to learning, the FV representations are further L2 normalized. During SVM training, the stochastic dual coordinate ascent algorithm is employed to minimize the regularized loss due to its efficiency and fast convergence rate.

3 Experimental Setting and Results

We validate our proposed method using ISBI 2016 challenge dataset of dermoscopic lesion images [6]. The dataset released in the challenge contains 1279 dermoscopic lesion images with corresponding class labels pre-partitioned into a training set of 900 images and a testing set of 379 images. There are two lesion categories in the dataset: melanoma and benign (non-melanoma). Approximately 20% of the dataset is melanoma (173 images in training set, 75 images in testing set). We obtain optimal hyper-parameters of SVM classifiers using cross-validation strategy on training data, then give final result on testing data. For performance metrics, we adopt the mean average precision (mAP), accuracy (Acc), area under receive operation curve (AUC),

sensitivity (Sen) and specificity (Spec). All experiments are conducted on a 128G RAM computer with CPU Inter Xeon E5-2680 @ 2.70 GHz, GPU NVIDIA Quadro K4000.

We start by investigating the influence of image preprocessing. We carry out the experiment with different rescaled images and perform two normalizations (per-img-mean and all-img-mean). As seen from Fig. 3(a), in the case of adopting the normalization strategy of per-img-mean, the classification performance improves gradually as the scale increases, and remains stable after the scale reaches 448. When the scale is larger than 448, however, no significant improvement is observed, which indicates there is no gain of information with higher computational burden. For normalization with all-img-mean, the classification performance first increases as S increases. It reaches a peak at $S = 384$ followed by a steep fall. This demonstrates the superiority of per-img-mean over all-img-mean in normalizing lesion images with large S. By balancing the memory consumption and efficiency, we fix S as 448 in the following experiments.

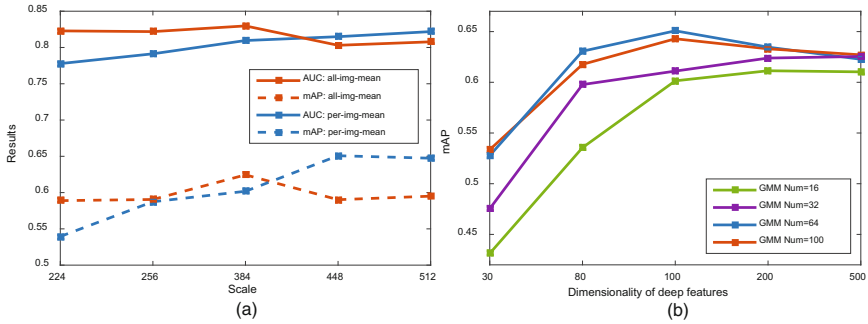

Fig. 3. Evaluation of proposed method on image preprocessing; (b) mAP of our method with varying number of Gaussians and dimensionality of deep features.

We conduct experiment to shed light on how parameters of FV encoding affect the classification performance. Apart from the size of GMM codebook, we also investigate the influence of PCA dimensionality of deep features. The result is illustrated in Fig. 3 (b). It can be observed that increasing the dimension of deep features yields significant improvements in performance of mAP initially. As the dimensionality becomes higher, mAP gradually drops in all GMM components setting, which indicates that the number of current Gaussians and the number of training samples are insufficient to model the distribution of higher dimensional features. In addition, we can see that the performance of larger GMM number setting (i.e. GMM Num = 100) outperforms the other fewer Gaussians in high dimensionality of 500, which suggests that increasing the number of GMM components can improve the performance. However, for larger number of GMM components, more training data is needed to estimate the GMM parameters. Furthermore, increasing Gaussian numbers under the case of high dimensional features leads to very high memory consumption and computational

complexity. Hence, it is crucial to examine settings of GMM components and feature dimensionality, given the limited training data and computational platform.

Apart from the proposed ResNet-50, we explore two other CNN models including 8-layers AlexNet [10], and 16-layers VGGNet (VGG-16) [13] for performance comparison. All the models are pre-trained on ImageNet. We keep the same setting in the process of classification whenever possible for fair comparison. Table 1 shows the experimental results. Also, the average running time of each network for processing single lesion image is provided. Finally, in Table 2, we compare our result with the top ranked method in the challenge and method reported in the recent published literature [4, 16], our method is denoted as LDF-FV. For the fusion case, deep features are extracted from middle and last convolutional layers of ResNet-50, respectively. The final scores are given by averaging scores of two different level descriptors based FV representations.

Table 1. Impact of network architectures on the classification results (%).

Network	Parameter	Layer	Sen	Spec	mAP	Acc	AUC	Time
AlexNet	61 M	Conv5	40.00	95.72	61.37	84.70	82.08	0.94 s
VGG-16	138 M	Conv5_3	45.33	94.08	57.66	84.43	81.18	2.72 s
ResNet-50	25.6 M	Conv5_9	**45.33**	**96.71**	**65.08**	**86.54**	**81.49**	1.33 s

Table 2. Comparison of the proposed approach with other methods (%).

Method	Network	Dimension	Sen	Spec	mAP	Acc	AUC
DSIFT-FV	Na	12800	33.33	95.39	55.63	83.11	78.01
CNN-SVM [6]	ResNet-50	2048	40.00	95.39	58.42	84.43	81.82
CNNaug-SVM [12]	ResNet-50	2048	41.33	94.41	59.93	83.19	81.73
Fine-tuned CNN [12]	ResNet-50	Na	48.00	94.08	63.36	84.96	81.58
CUMED [7]	FCRN (DRN)-50	Na	50.70	94.10	63.70	85.50	80.40
Codella [6]	Ensemble models	Na	**69.30**	83.20	64.50	80.50	83.80
LDF-FV (ours)	ResNet-50	12800	45.33	96.71	65.08	86.54	81.49
LDF-FV (fusion)	ResNet-50	12800	42.67	**97.70**	**68.49**	**86.81**	**85.20**

4 Conclusion

In this paper, we propose a novel framework for dermoscopy image classification. It utilizes the state-of-the-art local descriptors encoding method (FV) to encode local convolutional features extracted from very deep residual network into holistic representations, which is more discriminative than hand-crafted descriptors and CNN features. Systematical and extensive experiments are performed to investigate a range of key elements that could affect the performance of our method. Experiments on the publicly available ISBI 2016 challenge skin lesion dataset show the promising results. Also, we compare the proposed framework with a number of existing well-established

classification methods to further validate its superiority. Our future work will focus on studying the performance of our method on other applications and networks pre-trained on different datasets.

Acknowledgment. This work was supported partly by National Natural Science Foundation of China (Nos. 81571758, 61571304, 61402296, 61571304 and 61427806), National Key Research and Develop Program (No. 2016YFC0104703), Shenzhen Peacock Plan (NO. KQTD2016 053112051497), and the National Natural Science Foundation of Shenzhen University (No. 827000197).

References

1. Abder-Rahman, A.A., Deserno, T.M.: A systematic review of automated melanoma detection in dermatoscopic images and its ground truth data. In: Proceedings of SPIE Medical Imaging (2012)
2. Cimpoi, M., Maji, S., Vedaldi, A.: Deep filter banks for texture recognition and segmentation. In: CVPR (2015)
3. Codella, N., Cai, J., Abedini, M., Garnavi, R., Halpern, A., Smith, J.R.: Deep learning, sparse coding, and SVM for melanoma recognition in dermoscopy images. In: Zhou, L., Wang, L., Wang, Q., Shi, Y. (eds.) MLMI 2015. LNCS, vol. 9352, pp. 118–126. Springer, Cham (2015). doi:10.1007/978-3-319-24888-2_15
4. Codella, N., Nguyen, Q.-B., Pankanti, S., Gutman, D., Helba, B., Halpern, A., Smith, J.R.: Deep learning ensembles for melanoma recognition in dermoscopy images. arXiv preprint arXiv:1610.04662 (2016)
5. Gong, Y., Wang, L., Guo, R., Lazebnik, S.: Multi-scale orderless pooling of deep convolutional activation features. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014. LNCS, vol. 8695, pp. 392–407. Springer, Cham (2014). doi:10.1007/978-3-319-10584-0_26
6. Gutman, D., Codella, N.C., Celebi, E., Helba, B., Marchetti, M., Mishra, N., Halpern, A.: 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). arXiv preprint arXiv:1605.01397 (2016)
7. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR (2016)
8. Kawahara, J., Bentaieb, A., Hamarneh, G.: Deep features to classify skin lesions. In: ISBI (2016)
9. Konstantin, K., Rafael, G.: Computerized analysis of pigmented skin lesions: a review. Artif. Intell. Med. **2**(56), 69–90 (2012)
10. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: NIPS (2012)
11. Sánchez, J., Perronnin, F., Mensink, T., Verbeek, J.: Image classification with the Fisher vector: theory and practice. Int. J. Comput. Vision **3**(105), 222–245 (2013)
12. Sharif, R.A., Azizpour, H., Sullivan, J., Carlsson, S.: CNN features off-the-shelf: an astounding baseline for recognition. In: CVPR Workshop (2014)
13. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)

14. Uricchio, T., Bertini, M., Seidenari, L., Bimbo, A.D.: Fisher encoded convolutional bag-of-windows for efficient image retrieval and social image tagging. In: ICCV Workshop (2015)
15. Yoo, D., Park, S., Lee, J.-Y., So Kweon, I.: Multi-scale pyramid pooling for deep convolutional representation. In: CVPR Workshop (2015)
16. Yu, L., Chen, H., Dou, Q., Qin, J., Heng, P.A.: Automated melanoma recognition in dermoscopy images via very deep residual networks. IEEE Trans. Med. Imag. **4**(36), 994–1004 (2017)
17. Yue-Hei Ng, J., Yang, F., Davis, L.S.: Exploiting local features from deep networks for image retrieval. In: CVPR Workshop (2015)