IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received September 7, 2018, accepted September 28, 2018, date of publication October 23, 2018,
date of current version November 30, 2018.

Digital Object Identifier 10.1109/ACCESS.2018.2877587

Classification for Dermoscopy Images Using Convolutional Neural Networks Based on Region Average Pooling

JIAWEN YANG[1,2], FENGYING XIE[1,2], HAIDI FAN[1,2], ZHIGUO JIANG[1,2], AND JIE LIU[3]

[1]Image Processing Center, School of Astronautics, Beihang University, Beijing 100191, China
[2]Beijing Advanced Innovation Center for Biomedical Engineering, Beihang University, Beijing 100191, China
[3]3Department of Dermatology, Peking Union Medical College Hospital, Chinese Academy of Medical Sciences and Peking Union Medical College, Beijing 100730, China
Corresponding author: Fengying Xie (xfy_73@buaa.edu.cn)

This work was supported by the National Natural Science Foundation of China under Grant 61871011, Grant 61471016, and Grant 61771031.

ABSTRACT In this paper, a novel melanoma classification method based on convolutional neural networks is proposed for dermoscopy images. First a *region average pooling* (RAPooling) method is introduced which makes feature extraction can focus on the region of interest. Then an end-to-end classification framework combining with segmentation information is designed, which uses the segmented lesion region to guide the classification by RAPooling. Finally, a linear classifier RankOpt based on the area under the ROC curve is used to optimize and obtain the final classification result. The proposed method integrates segmentation information into the classification task, and in addition, by the optimization of RankOpt, a better classification performance for imbalanced dermoscopy image dataset is obtained. Experiments are conducted on ISBI 2017 *skin lesion analysis towards melanoma detection* challenge dataset, and comparisons with the other state-of-the-art methods demonstrate the effectiveness of our method.

INDEX TERMS Convolutional neural networks, dermoscopy images, melanoma detection, region average pooling.

I. INTRODUCTION

Melanoma, a type of skin cancer that mostly starts in pigment cells, is one of the deadliest forms of cancer. [27]. According to American Cancer Society [28], about 87110 new cases of melanoma are estimated to be diagnosed and about 9730 fatalities are estimated in United States in 2017. The most important way to increase the survival rate is to detect melanoma in its early stages and treat it properly [5].

The development of dermoscopy technique can significantly contribute to improving the diagnostic accuracy of melanoma, and thus improving the survival rate of patients. Dermoscopy [3] is a noninvasive skin imaging technique, which uses polarized light to make the contact area translucent, and can reveal the subsurface skin structure. However, manual interpretation of the dermoscopy image is usually time-consuming, experiential, and subjective. Therefore, computer-aided diagnosis (CAD) has been developed to provide fast, quantitative, and objective evaluation for dermatologists. According to [26], when skilled dermatologists use CAD systems to evaluate skin lesions, the diagnostic accuracy of melanoma can be increased from 75% to 92%.

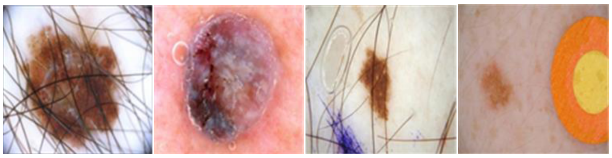


FIGURE 1. Examples of challenging dermoscopy image.

However, automated melanoma classification for dermoscopy images is quite a challenging task. First, there is no unified manifestation of melanoma. Usually, melanomas have huge variations in terms of color, texture and shape, which makes it difficult to extract robust features. Second, many dermoscopy images have hairs, veins, bubbles, as shown in Fig. 1. These noisy artifacts heavily interfere with the recognition of melanoma. Finally, the problems of data insufficiency and class imbalance in public dermoscopy image datasets also greatly limit the performance of algorithms.

Melanoma classification has been studied for many years, which can be found in the literature as early as 1987 [4].

Traditional methods for melanoma classification usually include two stages: feature extraction and classifier design. For example, in [22], Kusumoputro and Ariyantor extracted shape and color features from dermoscopy images, with which an artificial neural network was trained to separate malignant melanomas from benign lesions. In [5], Celebi *et al.* extracted 437 color and texture features, from which 18 optimal features were selected to train a support vector machine (SVM) classifier. In [33], Xie *et al.* extracted color, texture and border features to train a neural network ensemble model to classify skin lesions into benign nevi or melanomas. Considering dermoscopy images may not always capture entire lesions, Situ *et al.* [30] and Barata *et al.* [2] extracted local features in a patch and used bag-of-features (BoF) models to classify lesions. Because the extracted features are low-level and hand-crafted, these traditional classification methods are usually not robust for complex skin lesions.

Recently, deep-learning methods [14], especially convolutional neural networks (CNNs), have shown outstanding performance and powerful generalization ability in many medical image analysis tasks, including but not limited to segmentation [16], [25], classification [1], [20] and detection [10], [11]. Deep learning is able to learn multiple levels of representation from raw image data, and the extracted features are more high-level and more robust. In [9], Deng *et al.* developed a two-branch CNN to extract global and local features to obtain the lesion border. In [35], Yuan *et al.* designed a novel loss function based on Jaccard distance for their CNN model to improve its performance on dermoscopy image segmentation. Codella *et al.* [6] utilized a pre-trained CNN as the feature extractor to obtain high-level features, and then provided them to a SVM classifier after sparse coding. In [34], Yu *et al.* constructed a fully convolutional residual network (FCRN) for skin lesion segmentation, and then cropped the lesion image patches with which a new deep residual network for classification was trained.

Generally, there are three main differences between deep-learning methods and traditional machine-learning methods for melanoma classification. First, deep-learning methods do not need to design hand-crated features, which have the capability of learning hierarchical features from raw dermoscopy images. Second, deep-learning models are usually trained end to end, and directly predict the type of skin lesion without segmentation, although segmentation is a key step before classification in the traditional classification framework. Finally, deep-learning models usually have a very large amount of parameters, which means they have a higher requirement for the amount of training data.

Inspired by the latest advance in deep learning research and melanoma classification, we propose a novel framework based on CNN to automatically discriminate melanomas from non-melanomas.

The main contributions of our work can be summarized as follows:

1) We propose a weighted global average pooling operation, namely region average pooling, which can help classifier put the focus on the region of interest.
2) We design a CNN based classifier for the dermoscopy image which has two branches: segmentation branch and classification branch. The segmentation branch can obtain the location information of skin lesions, and through the proposed region average pooling, the lesion location information can be provided to the classification branch to facilitate the lesion classification.
3) We utilize an AUC-based classifier RankOpt as the post-processing of the CNN model, which improves the robustness to class imbalance.

The remainder of this paper is organized as follows. In Section II, we introduce the details of the proposed region average pooling method, the designed CNN framework, and the AUC-based classifier RankOpt. Experiments and discussion are presented in Section III and IV, respectively. Finally, we conclude our work in Section V.

II. METHOD

A. REGION AVERAGE POOLING

It is well-known that feature maps in CNNs can capture rich spatial information, especially for those models using global average pooling [23], such as GoogLeNet [32] and ResNet [17]. The work by Zhou *et al.* [36] has shown that CNN models using global average pooling for object classification task can retain its remarkable localization ability, and the convolutional layer can behave as object detector, although no object location information is provided for supervised learning. In [37], Zhou *et al.* proposed a general technique called class activation mapping (CAM). The CAM can visualize layer activations and highlight the discriminative image region used by the CNN to identify the category.

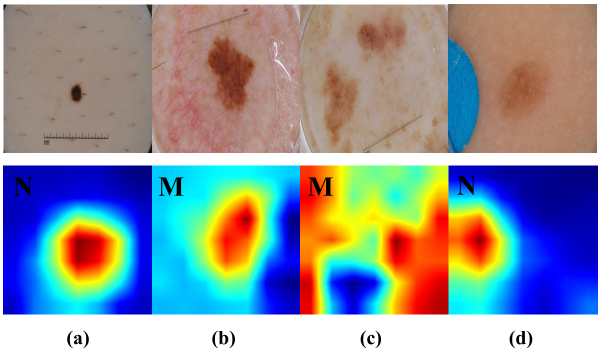


FIGURE 2. Examples of dermoscopy image and their CAMs corresponding to their true classes by using the regular ResNet50. (b) (c) are melanomas, and (a) (d) are non-melanomas. The classification results of (a) (b) (d) are correct, and (c) is wrong.

We trained a regular ResNet50 model [17] with the global average pooling for melanoma classification on ISBI 2017 dataset [7]. Fig. 2 shows four dermoscopy images and their CAMs corresponding to their true classes. Images in Fig. 2(a) and 2(b) are correctly classified and it is obvious that the classification model has correctly focused on the lesion regions. However, in Fig. 2(c), the model fails to locate the lesion region, which leads to misclassification. In Fig. 2(d), although the classification result is correct, the model focuses on the blue disc instead of the lesion region. In the training data, there are several images with color discs, and all of them are non-melanoma images, which might lead the model to learn that the color disc represents non-melanoma. In other words, the classification result in Fig. 2(d) is actually unreliable.

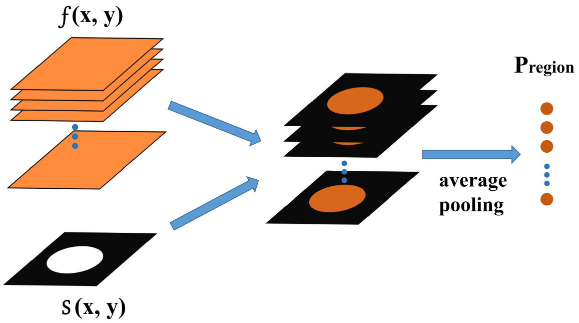


FIGURE 3. **Illustration of the proposed region average pooling.**

From the analysis above, we can see that the melanoma classification network is implicitly learning skin lesion segmentation. Therefore, by introducing segmentation information, we propose a new pooling operation named *region average pooling* to improve the classification performance. Being different from the global average pooling [17], [32] which calculates the mean of all the activation of a feature map in the last convolutional layer, the proposed region average pooling is limited in a region of interest, as shown in Fig. 3, and it is defined as:

$$P^i_{region} = \frac{1}{WH}\sum_{x=1}^{W}\sum_{y=1}^{H} f_i(x,y)\cdot S(x,y), \tag{1}$$

where W and H denote the width and height of feature maps, respectively, and $f_i(x,y)$ represents the activation of the i^{th} feature map at spatial location (x,y), and $S(x,y)$ represents the weight map which measures the importance at different spatial locations. Because we focus on the features in the lesion region, $S(x,y)$ is actually the score map of lesion regions obtained from the segmentation task in this paper.

B. DESIGNED CLASSIFICATION FRAMEWORK BASED ON THE REGION AVERAGE POOLING

Based on the region average pooling, our proposed CNN framework for melanoma classification is shown in Fig. 4, which includes a ResNet50 [17] structure, and followed by two branches: segmentation branch and classification branch.

In ResNet50, the residual structure is designed for improving the learning ability of the network. A typical residual structure is illustrated in Fig. 4(a), in which the rectified linear unit (ReLU) and batch normalization (BN) layers are hidden for simplicity, and only convolutional layers are presented. The ReLU layer [21] gives the network the ability to fit a nonlinear mapping, and can also alleviate the gradient vanishing problem to some extent. The BN layer [19] normalizes the data distribution in convolutional layers in order to reduce the internal covariate shift. In the residual structure, both BN and ReLU layers are used after each convolutional layer. The ResNet50 structure includes a convolution layer, a max pooling layer and then four residual blocks. To reduce overfitting, dropout is added in four residual blocks. In our framework, the dermoscopy image is fed into the ResNet50 structure to extract high-level features, and then these features are input to the classification branch as well as the segmentation branch.

In the segmentation branch, feature maps with two channels are obtained by the 1 × 1 convolution layer. Because of striding and max pooling in ResNet50 structure, the spatial resolution of the feature maps has been reduced to $1/32$. Then upsampling is performed through a transposed convolution with a stride of 32 to recover the spatial size of the feature maps to compute pixel-wise cross entropy loss.

Before upsampling, the two-channel feature maps are also input to a softmax function to obtain the score map of lesion regions, which is actually the weight map in the region average pooling in (1). Let h_1 and h_0 denote the two channels of the feature maps, respectively, and $h_1(x,y)$ represents the score of being lesion at spatial location (x, y), and $h_0(x,y)$ represents the score of being non-lesion. Then the normalized score map $S(x,y)$ of being lesion can be given by:

$$S(x,y) = \frac{\exp(h_1(x,y))}{\exp(h_1(x,y)) + \exp(h_0(x,y))}. \tag{2}$$

The score map $S(x,y)$ describes the location information of skin lesions. Ideally, it is close to 1 in the lesion region and 0 in the background region.

In the classification branch, $S(x,y)$ is input to the region average pooling to weight the features, and more discriminative features can be extracted. After the pooling, a two-way fully connected layer is used to predict the scores of melanoma and non-melanoma, as shown in Fig. 4.

The proposed classification framework looks like a classic multi-task model, which improves the discrimination of features, and ultimately improves the accuracies of the two tasks by introducing the segmentation and classification information at the same time. But there are two main differences between our model and the classic multi-task model. Because of the region average pooling, the segmentation task and the classification task in our model are not independent of each other, which makes it difficult to jointly train the two tasks from scratch. The other difference is that the prime goal of our model is to improve the classification performance, and segmentation is used to serve classification, therefore we will put more focus on the classification task in joint training.

Based on the analysis above, we train the designed framework in four steps:

1) **Step1. Fine-tuning:** The size of the public dermoscopy dataset is usually small, which makes a

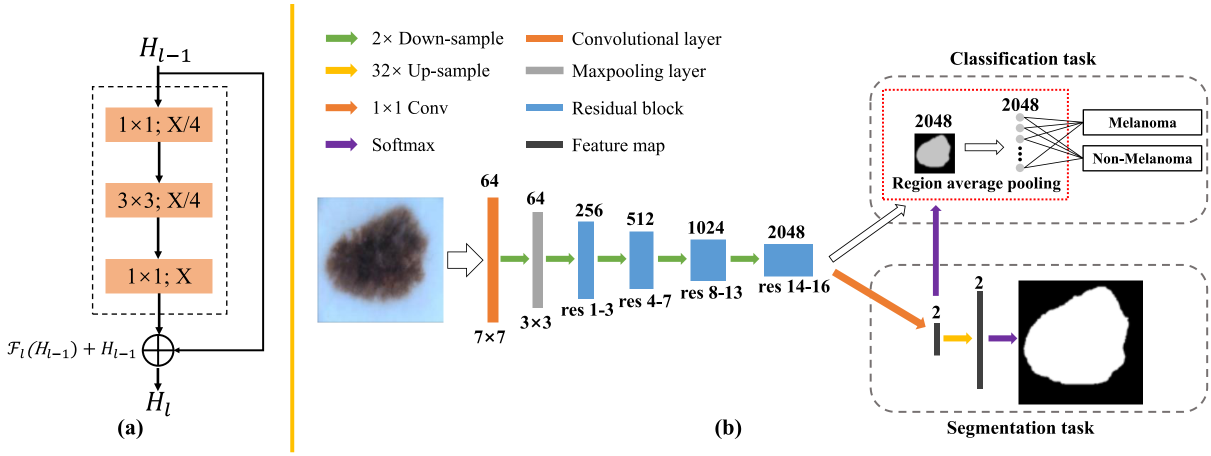


FIGURE 4. The flowchart of the proposed framework. (a) The diagram of a residual structure. The ReLU and BN layers are hidden for simplicity. (b) Architecture of our designed classification network with a segmentation branch based on the region average pooling.

deep neural network initialized with random weights have long training time, difficult convergence and low robustness. Therefore, we utilize a ResNet50 pre-trained on ImageNet [8] to fine-tune our model. The weights of layers are all initialized with the weights in the pre-trained network. As for the top fully connected layer, the output classes are changed to binary (melanoma and non-melanoma), and the weights are initialized randomly.

2) **Step2. Training classification task:** To make the network easier to converge, we only train the classification task with the global average pooling at beginning. In this way, the network can be able to retain its remarkable localization ability in convolutional layers.
3) **Step3. Training segmentation task:**When the loss value on validation data is very low and the validation accuracy does not tend to increase, we begin to train the segmentation task. In order to reduce the influence of segmentation training on the classification performance, we only train the weights of the convolutional layer in the segmentation branch, and other weights are frozen.
4) **Step4. Joint training:** In joint training, the global average pooling is replaced with the region average pooling which combines the classification task with the segmentation task. The loss in joint training L_{joint} is formulated as:

$$L_{joint} = L_{class} + \lambda L_{seg}, \tag{3}$$

where L_{class} and L_{seg} represent the cross entropy losses in the classification branch and the segmentation branch, respectively, and λ is a hyper-parameter to control the balance between the two losses.

C. CLASSIFICATION USING RANKOPT

Due to the difference in morbidity, there are often serious problems of class imbalance in the dermoscopy image dataset, which may cause the classifier to perform sub-optimally, and the trained classifier is inclined to classify samples into non-melanomas to improve the overall accuracy [24]. In our proposed framework, we use the softmax classifier and the cross entropy loss function to train the network. Actually, most of CNNs are trained using the accuracy based loss function, for example the mean square error loss [13] and the cross entropy loss [17]. This kind of loss function can only use part of samples during each iteration in training, and thus is suitable for deep learning with a big and balanced dataset. However, on an imbalanced dataset, because the ratio of size of the majority class to the minority class is not taken into account, the classifier which is trained by this kind of loss usually has poor performance.

For a binary classification task, the AUC stands for the area under the ROC curve, which considers the ordering relations between positive samples and negative samples, and is not sensitive to sample distribution. In this paper, we take the proposed CNN as a feature extractor, and input the extracted features (the output of the region average pooling in Fig. 4) to a linear classifier RankOpt [18] based on AUC to optimize, and obtain the final classification result.

The linear classifier RankOpt adopts the AUC statistic as its objective function and optimizes it directly using gradient descent. Taking melanoma images as positive samples and non-melanoma images as negative samples. P and Q represent the numbers of positive and negative samples, respectively, and P < Q for most dermoscopy image datasets. And for a linear classifier with the weight vector $\boldsymbol{\beta}$, its AUC statistic is defined as:

$$AUC(\boldsymbol{\beta}) = \frac{1}{PQ}\sum_{i=1}^{P}\sum_{j=1}^{Q} g(\boldsymbol{\beta} \cdot (x_i^+ - x_j^-)), \tag{4}$$

$$g(x) = \begin{cases} 0, & x < 0 \\ 0.5, & x = 0 \\ 1, & x > 0 \end{cases} \tag{5}$$

where x_i^+ and x_j^- denote the feature vectors of the i^{th} positive sample and the j^{th} negative sample, respectively.

Since $g(x)$ in (4) is undifferentiable, it is replaced by the sigmoid function $s(x) = 1/(1+e^{(-x)})$. Then, the rank statistic $R(\boldsymbol{\beta})$ is defined as:

$$R(\boldsymbol{\beta}) = \frac{1}{PQ}\sum_{i=1}^{P}\sum_{j=1}^{Q} s(\boldsymbol{\beta}\cdot(x_i^+ - x_j^-)). \tag{6}$$

Because $\lim_{|x|\to\infty} s(x) = g(x)$, the $R(\boldsymbol{\beta})$ will be a good approximation to the AUC statistic in (4) when $\|\boldsymbol{\beta}\|$ is large. According to (4) and (6), the AUC value actually depends on the direction of $\boldsymbol{\beta}$ and not on its magnitude. Therefore, we can constrain $\boldsymbol{\beta}$ to a hypersphere $\|\boldsymbol{\beta}\| = B$ (B is fixed as a large number) and optimize its iteration. In this way, the optimal weight vector $\boldsymbol{\beta}_{opt}$ of the linear classifier can be defined as:

$$\boldsymbol{\beta}_{opt} = \arg\max_{\boldsymbol{\beta}} R(\boldsymbol{\beta}) \quad s.t.\ \|\boldsymbol{\beta}\| = B. \tag{7}$$

The partial derivative of (6) for the k^{th} weight of $\boldsymbol{\beta}$ is given by:

$$\frac{\partial R(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}_k} = \frac{1}{PQ}\sum_{i=1}^{P}\sum_{j=1}^{Q} s(\boldsymbol{\beta}\cdot(x_i^+ - x_j^-)) \cdot(1 - s(\boldsymbol{\beta}\cdot(x_i^+ - x_j^-)))\cdot(x_{ik}^+ - x_{jk}^-). \tag{8}$$

Then the $\boldsymbol{\beta}_{opt}$ can be obtained by gradient descent optimization.

Note that the computational complexity of $R(\boldsymbol{\beta})$ is $O(n^2)$ in the number of samples. In [18], Herschtal and Raskutti showed more details about RankOpt and reduced its computational complexity to $O(n)$. However, it is still time-consuming and memory-consuming to integrate the RankOpt classifier into our proposed CNN framework for end-to-end training. Therefore, we use the proposed CNN in section II-B as a feature extractor to extract high-level features, with which to train the RankOpt classifier. And the final classification result is obtained.

III. EXPERIMENT RESULTS AND ANALYSIS

Experiments were implemented with PyTorch library on a computer equipped with a NVIDIA GTX1080 GPU with 8GB of memory. The proposed network was trained by stochastic gradient descent (SGD) optimization method with momentum 0.9 and weight decay 0.0005. We set batch size as 10, the initial learning rate as 0.001 and reduced it by a factor of 10 every 2000 iterations. The weights were initialized from the ResNet50 pre-trained on ImageNet dataset.

Experimental images are from ISBI 2017 Skin Lesion Analysis Towards Melanoma Detection Challenge dataset [7], which is provided by the International Skin Imaging Collaboration (ISIC). This dataset consists of 2000 dermoscopy images (374 melanomas) as training data, 150 images (30 melanomas) as validation data, and 600 images (117 melanomas) as test data. The sizes of these images range from 542 × 718 to 2848 × 4288. And the segmentation ground truth is also available from the challenge, which is based on the manual delineation by clinical experts.

Considering the small size of the pubic dataset, we artificially transformed the original images to increase the size and diversity of training data, including rotation, flipping and cropping. For each iteration in training, input images were resized to 256 × 256 and then randomly cropped to 224 × 224. At the same time, the segmentation ground truth masks were also resized and cropped.

To evaluate the performance of our proposed method quantitatively, four evaluation metrics are used for comparison, including sensitivity (Sen), specificity (Spe) and accuracy (Acc). They are defined as follows:

$$Sen = \frac{TP}{TP + FN}, \tag{9}$$

$$Spe = \frac{TN}{TN + FP}, \tag{10}$$

$$Acc = \frac{TP + TN}{TP + FP + TN + FN}, \tag{11}$$

where TP, FP, TN and FN represent the numbers of true positives, false positives, true negatives and false negatives, respectively. Besides, the AUC metric is also calculated for further comparison.

In these metrics, the sensitivity represents the percentage of melanomas that are correctly classified, which is more clinically relevant. While in ISBI 2017 Challenge, participants were ranked according to the AUC, although other metrics were also computed for scientific completeness. Therefore, we also chose the AUC as the primary metric for evaluation in this paper.

A. PARAMETER DETERMINATION

In the joint training step, the joint loss consists of the classification loss and the segmentation loss, in which a hyper-parameter λ is used to balance the importance between the two losses.

TABLE 1. Statistical results of JA and AUC using different λ on validation data.

λ	0.01	0.1	1	5	7.5	10	12.5	15	20
JA	0.075	0.401	0.560	0.635	0.633	0.628	0.638	0.632	0.626
AUC	0.775	0.818	0.813	0.860	**0.863**	0.859	0.851	0.807	0.806

We used grid search over a discrete range of λ={0.01, 0.1, 1, 5, 7.5, 10, 12.5, 15, 20} to determine the optimal λ. We calculated Jaccard index (JA) and AUC metrics for different λ on validation data, and Table 1 shows the results. JA is the metric of segmentation accuracy [35], and a high JA indicates a good segmentation result. It can be seen that, with λ becomes bigger, the joint loss puts more focus on segmentation, and the segmentation accuracy gradually increases. When $\lambda \geq 5$, the segmentation performance becomes stable. AUC measures the classification accuracy. It can be seen that the classification accuracy increases as the segmentation accuracy increases. When the segmentation performance becomes stable, continually increasing λ will neglect the importance of classification, and the classification accuracy decreases. Considering the AUC is the highest when $\lambda = 7.5$, it is determined as the optimal λ.

B. EFFECTIVENESS OF THE REGION AVERAGE POOLING IN CLASSIFICATION

In this paper, a region average pooling operation is proposed to integrate the segmentation information into the classification task to obtain better classification performance. In order to verify the effectiveness of the proposed region average pooling in classification, we compared the performances of different deep learning networks on ISBI 2017 test data when combined with the region average pooling, including VGG [29], GoogLeNet [12], [32] and ResNet50 [17], [34]. Table 2 shows the results of the three widely used networks. It can be seen that the AUC metrics of the three networks are all improved when combined with the region average pooling.

TABLE 2. Statistical results of sensitivity, specificity, accuracy and AUC using different networks combined with/without the region average pooling on ISBI 2017 test data.

	Sen	Spe	Acc	AUC
VGG	0.325	0.932	0.813	0.767
VGG+RAPooling	0.325	0.936	0.817	**0.801**
GoogLeNet	0.444	0.909	0.818	0.821
GoogLeNet+RAPooling	0.436	0.917	0.823	**0.825**
ResNet50	0.334	0.959	0.837	0.814
ResNe50+RAPooling	0.444	0.923	0.833	**0.838**

From Fig. 2, it can be known that the regular ResNet50 with global average pooling sometimes cannot put the focus on the lesion region. Fig. 5 shows the same dermoscopy images presented in Fig. 2 and their CAMs obtained by the ResNet50 with the region average pooling. It can be seen that our proposed network can put its focus on the lesion region very well. The region average pooling based classification framework combines classification with segmentation, and through joint training, the classification task is guided by the object region obtained from the segmentation branch, and thus better classification performance is obtained. In Table 2, the ResNet50 with the region average pooling (ResNet50+RAPooling) obtained the highest AUC, which was used in our final classification framework.

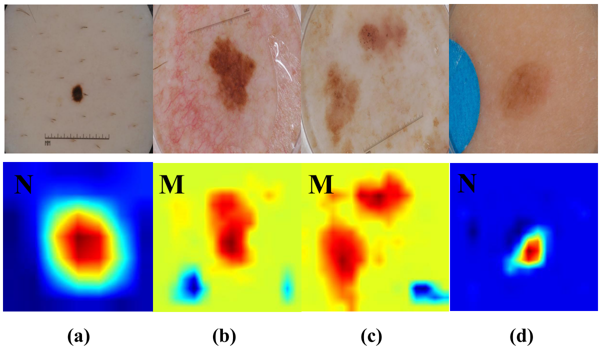


FIGURE 5. Examples of dermoscopy image and their CAMs corresponding to their true classes by using our proposed ResNet50+RAPooling framework. The classification results of (a) (b) (c) and (d) are all correct.

C. EFFECTIVENESS OF CLASSIFICATION USING RANKOPT

Usually, the melanoma samples are much less than the non-melanoma samples in a dermoscopy image dataset due to the difference in morbidity, and in order to improve the overall accuracy, the trained classifier is inclined to predict samples into non-melanomas. In this paper, we used the ResNet50+RAPooling model as the feature extractor, and used the linear classifier RankOpt [18] to optimize and obtain the final classification result, which alleviated the class imbalance problem to an extent.

TABLE 3. Statistical results of sensitivity, specificity, accuracy and AUC using ResNet50+RAPooling and ResNet50+RAPooling+RankOpt on ISBI 2017 test data.

	Sen	Spe	Acc	AUC
ResNet50+RAPooling	0.444	0.923	0.833	0.838
ResNet50+RAPooling+RankOpt	0.607	0.884	0.833	**0.842**

Table 3 shows the results of ResNet50+RAPooling and our final framework (ResNet50+RAPooling+RankOpt). It can be seen that the AUC metric is improved. Besides, the ResNet50+RAPooling only yields sensitivity of 44.4% because of the imbalanced samples in training data (374 melanomas vs. 1626 non-melanomas). While our final framework used the RankOpt classifier to optimize the result of ResNet50+RAPooling, and the sensitivity is greatly improved by 16.3% with a still high specificity of 88.4%. Therefore, by using RankOpt, a better balance between sensitivity and specificity can be obtained, which is clinically significant.

D. COMPARISON WITH OTHER CLASSIFICATION METHODS

We compared our final classification method with the top 8 methods in ISBI 2017 challenge, and four other methods including Harangi's [15], Sultana *et al.*'s [31], Wiselin's and Xie *et al.*'s [33] methods. Harangi's and Sultana's methods are based on deep learning. Wiselin's method is based on SVM, which is the only traditional machine learning method in ISBI 2017 challenge, and Xie's method is our early work. Table 4 shows the classification results of these methods on ISBI 2017 test data, where the first 8 methods are the top 8 of the leaderboard. Xie's method is based on neural network ensemble, and it used voting method to obtain the final prediction from neural network individuals, therefore its AUC metric cannot be given in Table 4.

Wiselin's and Xie's methods are two traditional methods, which extracted low-level features, and their classification performances are obviously lower than the 11 deep-learning methods.

In Table 4, the first five methods are the top five in the challenge, and they used hundreds or thousands additional images, but our method did not used. It can be seen that our method has better AUC than the fifth method TD. And although the fourth method Monty has a higher AUC than our method, it has the lowest sensitivity of 10.3% among all the

TABLE 4. Statistical results of sensitivity, specificity, accuracy and AUC using our proposed method and other methods on ISBI 2017 test data.

	Sen	Spe	Acc	AUC
RECOD Titans	0.547	0.95	0.872	0.874
Lei Bi	0.427	0.963	0.858	0.870
Kazuhisa Matsunaga	0.735	0.851	0.828	0.868
Monty python	0.103	0.998	0.823	0.856
TD	0.350	0.965	0.845	0.836
Xueleiyang	0.436	0.925	0.830	0.830
rafael sousa	0.521	0.901	0.827	0.805
finalv_L2C1_trir	0.376	0.957	0.843	0.804
Harangi [15]	0.402	0.719	0.852	0.851
Sultana [31]	0.529	0.905	0.832	0.789
Wiselin	0.470	0.511	0.503	0.495
Xie [33]	0.427	0.675	0.627	
ResNet50+RAPooling+RankOpt	0.607	0.884	0.830	0.842

13 methods. Sensitivity is more important than other metrics in the clinic, and low sensitivity will easily cause missed detection, and it is dangerous for the patient with malignant melanoma. Therefore, with a higher sensitivity, our method is actually better than the Monty.

The other five compared methods, including Xueleiyang, rafael and finalv_L2C1_trir, Harangi and Sultana, did not use additional data. It can be seen that our method has better performance on AUC and sensitivity, and only Harangi's method achieved a higher AUC than us, but the increase is only 0.9%, while its sensitivity and specificity are lower at least 16.5% than our method. Besides, Harangi's method ensembled the GoogLeNet, AlexNet, ResNet and VGGNet with a carefully weighted fusion mechanism, which was very complicated.

Therefore, although no additional data was used, our method is better than two methods which used additional data. And at the same time, our method is actually better than other compared methods which used no additional data.

IV. DISCUSS

For deep-learning methods in melanoma classification, it is common to use a single end to end network which directly takes input the whole dermoscopy image without considering the lesion region [15], or use two networks which segment skin lesion patches before inputting these patches into the classification network [34]. Although the latter takes the skin lesion region into consideration, the segmentation and the classification are actually two separate stages. And it will be difficult to extract discriminative features for classification if the former segmentation network outputs a wrong lesion mask. In this paper, the designed network contains two branches, the classification branch and the segmentation branch. They share the high-level features in the single network, and by the proposed region average pooling, the lesion location information in the segmentation branch can be directly provided to the classification branch. And more discriminative features are extracted in this way, which improves the classification performance.

In medical image analysis, class imbalance caused by different morbidity is an issue that cannot be ignored. Although the dermoscopy images in ISBI challenge are manually selected, the melanoma samples are still significantly less than non-melanomas, which can cause quite low sensitivity metric. Here we utilize the AUC-based RankOpt algorithm to further optimize the classifier, which can alleviate the class imbalance problem to an extent. And high sensitivity at high specificity can be achieved, which is more clinically significant.

In our proposed framework, the classification branch has used the location information of skin lesions to extract more discriminative lesion features, and the lesion location information is provided by the segmentation branch. Theoretically, more accurate location information can achieve more classification performance gains. However, obtaining the lesion location information is conducted on the feature maps which spatial sizes are 1/32 of the input image, and it is actually difficult to obtain very accurate lesion borders on these feature maps. Besides, the classification task and the segmentation task share most of weights in the framework, and in joint training, putting too much focus on segmentation will cause the classification performance degradation, therefore it needs to carefully control the balance between the two tasks.

V. CONCLUSION

Automated melanoma classification for dermoscopy images is quite a challenging task. In this paper, a novel framework based on CNN to automatically discriminate melanoma images from non-melanoma images is proposed. Regular CNNs with the global average pooling learn implicitly skin lesion segmentation to extract lesion features for melanoma classification. However, without direct supervision from the location information of lesions, the CNN sometimes cannot put the focus on lesion regions very well, which causes unreliable prediction. We propose a region average pooling method, which limits feature extraction in the lesion region. Based on the region average pooling, a CNN framework is proposed, which combines the classification task with the segmentation task through joint training, and can directly provide the lesion location information for classification, and thus a good classification result is obtained. Due to the difference in morbidity, there are often serious problems of class imbalance in the dermoscopy image dataset, which may cause the classifier to perform sub-optimally, and the trained classifier is inclined to classify samples into non-melanomas to improve the overall accuracy. We propose to use a linear classifier RankOpt based on the AUC to optimize the classification, and better performance is obtained. Experiments were conducted on ISBI 2017 dataset, and the results verified the effectiveness of our proposed method. Further investigations include exploring more efficient network structures, and combining with the region average pooling to extract more discriminative features for melanoma classification.

REFERENCES

[1] M. Anthimopoulos, S. Christodoulidis, L. Ebner, A. Christe, and S. Mougiakakou, "Lung pattern classification for interstitial lung diseases using a deep convolutional neural network," *IEEE Trans. Med. Imag.*, vol. 35, no. 5, pp. 1207–1216, May 2016.

[2] C. Barata, M. Ruela, M. Francisco, T. Mendonça, and J. S. Marques, "Two systems for the detection of melanomas in dermoscopy images using texture and color features," *IEEE Syst. J.*, vol. 8, no. 3, pp. 965–979, Sep. 2014.

[3] M. Binder *et al.*, "Epiluminescence microscopy of small pigmented skin lesions: Short-term formal training improves the diagnostic performance of dermatologists," *J. Amer. Acad. Dermatology*, vol. 36, no. 2, pp. 197–202, 1997.

[4] N. Cascinelli, M. Ferrario, T. Tonelli, and E. Leo, "A possible new tool for clinical diagnosis of melanoma: The computer," *J. Amer. Acad. Dermatology*, vol. 16, no. 2, pp. 361–367, 1987.

[5] M. E. Celebi *et al.*, "A methodological approach to the classification of dermoscopy images," *Comput. Med. Imag. Graph.*, vol. 31, no. 6, pp. 362–373, 2007.

[6] 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 *Proc. Int. Workshop Mach. Learn. Med. Imag.* Heidelberg, Germany: Springer, 2015, pp. 118–126.

[7] N. C. F. Codella *et al.* (2017). "Skin lesion analysis toward melanoma detection: A challenge at the 2017 international symposium on biomedical imaging (ISBI), hosted by the international skin imaging collaboration (ISIC)." [Online]. Available: https://arxiv.org/abs/1710.05006

[8] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2009, pp. 248–255.

[9] Z. Deng, H. Fan, F. Xie, Y. Cui, and J. Liu, "Segmentation of dermoscopy images based on fully convolutional neural network," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2017, pp. 1732–1736.

[10] Q. Dou, H. Chen, L. Yu, J. Qin, and P.-A. Heng, "Multilevel contextual 3-D CNNs for false positive reduction in pulmonary nodule detection," *IEEE Trans. Biomed. Eng.*, vol. 64, no. 7, pp. 1558–1567, Jul. 2017.

[11] Q. Dou *et al.*, "Automatic detection of cerebral microbleeds from MR images via 3D convolutional neural networks," *IEEE Trans. Med. Imag.*, vol. 35, no. 5, pp. 1182–1195, May 2016.

[12] A. Esteva *et al.*, "Dermatologist-level classification of skin cancer with deep neural networks," *Nature*, vol. 542, no. 7639, pp. 115–118, 2017.

[13] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 580–587.

[14] H. Greenspan, B. V. Ginneken, and R. M. Summers, "Guest editorial deep learning in medical imaging: Overview and future promise of an exciting new technique," *IEEE Trans. Med. Imag.*, vol. 35, no. 5, pp. 1153–1159, Mar. 2016.

[15] B. Harangi, "Skin lesion classification with ensembles of deep convolutional neural networks," *J. Biomed. Informat.*, vol. 86, pp. 25–32, Oct. 2018.

[16] M. Havaei *et al.*, "Brain tumor segmentation with deep neural networks," *Med. Image Anal.*, vol. 35, pp. 18–31, Jan. 2017.

[17] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 770–778.

[18] A. Herschtal and B. Raskutti, "Optimising area under the ROC curve using gradient descent," in *Proc. 21st Int. Conf. Mach. Learn.*, Jul. 2004, p. 49.

[19] S. Ioffe and C. Szegedy. (2015). "Batch normalization: Accelerating deep network training by reducing internal covariate shift." [Online]. Available: https://arxiv.org/abs/1502.03167

[20] S. Kiranyaz, T. Ince, and M. Gabbouj, "Real-time patient-specific ECG classification by 1-D convolutional neural networks," *IEEE Trans. Biomed. Eng.*, vol. 63, no. 3, pp. 664–675, Mar. 2016.

[21] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2012, pp. 1097–1105.

[22] B. Kusumoputro and A. Ariyanto, "Neural network diagnosis of malignant skin cancers using principal component analysis as a preprocessor," in *Proc. IEEE Int. Joint Conf. Neural Netw. World Congr. Comput. Intell.*, vol. 1, May 1998, pp. 310–315.

[23] M. Lin, Q. Chen, and S. Yan. (2013). "Network in network." [Online]. Available: https://arxiv.org/abs/1312.4400

[24] X.-Y. Liu, J. Wu, and Z.-H. Zhou, "Exploratory undersampling for class-imbalance learning," *IEEE Trans. Syst., Man, Cybern. B, Cybern.*, vol. 39, no. 2, pp. 539–550, Apr. 2009.

[25] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput.-Assist. Intervent.* Springer, 2015, pp. 234–241.

[26] T. Schindewolf, W. Stolz, R. Albert, W. Abmayr, and H. Harms, "Classification of melanocytic lesions with color and texture analysis using digital image processing," *Anal. Quant. Cytol. Histol.*, vol. 15, no. 1, pp. 1–11, 1993.

[27] P. Schmid-Saugeon, J. Guillod, and J.-P. Thiran, "Towards a computer-aided diagnosis system for pigmented skin lesions," *Comput. Med. Imag. Graph.*, vol. 27, no. 1, pp. 65–78, 2003.

[28] R. L. Siegel, K. D. Miller, and A. Jemal, "Cancer statistics, 2017," *CA, Cancer J. Clin.*, vol. 67, no. 1, pp. 7–30, 2017.

[29] K. Simonyan and A. Zisserman. (2014). "Very deep convolutional networks for large-scale image recognition." [Online]. Available: https://arxiv.org/abs/1409.1556

[30] N. Situ, X. Yuan, J. Chen, and G. Zouridakis, "Malignant melanoma detection by bag-of-features classification," in *Proc. 30th Annu. Int. Conf. Eng. Med. Biol. Soc. (EMBS)*, Aug. 2008, pp. 3110–3113.

[31] N. N. Sultana, B. Mandal, and N. B. Puhan, "Deep residual network with regularised Fisher framework for detection of melanoma," *IET Comput. Vis.*, to be published, doi: 10.1049/iet-cvi.2018.5238.

[32] C. Szegedy *et al.*, "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 1–9.

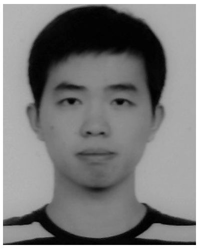
[33] F. Xie, H. Fan, Y. Li, Z. Jiang, R. Meng, and A. Bovik, "Melanoma classification on dermoscopy images using a neural network ensemble model," *IEEE Trans. Med. Imag.*, vol. 36, no. 3, pp. 849–858, Mar. 2017.

[34] L. Yu, H. Chen, Q. Dou, J. Qin, and P.-A. Heng, "Automated melanoma recognition in dermoscopy images via very deep residual networks," *IEEE Trans. Med. Imag.*, vol. 36, no. 4, pp. 994–1004, Apr. 2017.

[35] Y. Yuan, M. Chao, and Y.-C. Lo, "Automatic skin lesion segmentation using deep fully convolutional networks with Jaccard distance," *IEEE Trans. Med. Imag.*, vol. 36, no. 9, pp. 1876–1886, Sep. 2017.

[36] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba. (2014). "Object detectors emerge in deep scene CNNs." [Online]. Available: https://arxiv.org/abs/1412.6856

[37] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 2921–2929.



JIAWEN YANG received the B.S. degree from the School of Astronautics, Beihang University, Beijing, in 2017, where he is currently pursuing the degree. His research interests include biomedical image segmentation and classification.

FENGYING XIE received the Ph.D. degree in pattern recognition and intelligent system from Beihang University, Beijing, China, in 2009. She was a Visiting Scholar with the Laboratory for Image and Video Engineering, University of Texas at Austin, from 2010 to 2011. She is currently a Professor with the School of Astronautics, Beihang University. Her research interests include biomedical image processing, remote sensing image understanding and application, image quality assessment, and object recognition.

HAIDI FAN received the B.Eng. degree from the College of Information and Electrical Engineering, China Agricultural University, Beijing, China, in 2015, and the M.S. degree from the School of Astronautics, Beihang University, Beijing, China, in 2018. His research interests include biomedical image processing, segmentation, and classification.

ZHIGUO JIANG received the B.E., M.S., and Ph.D. degrees from Beihang University, Beijing, China, in 1987, 1990, and 2005, respectively. He was appointed as a Professor in image processing and pattern recognition in 2005. His research interests include remote sensing image analysis, medical imaging and analysis, target classification, detection, and recognition. He currently serves as a Standing Member for the Executive Council of the China Society of Image and Graphics.

JIE LIU received the B.S. and M.S. degrees from Peking University, Beijing, China, in 1998 and 2000, respectively, and the Ph.D. degree from the Peking Union Medical College, Beijing, in 2006. She held a post-doctoral position at Clipp, University of Bourgogne, Dijon, France, from 2010 to 2011. She is currently the Vice Director of the Deparment of Dermatology, Peking Union Medical College Hospital, Beijing. Her research interests include skin imaging and cutaneous lymphoma.