

In this paper, we propose an attention residual learning convolutional neural network (ARL-CNN) model for the skin lesion classification. We jointly use the residual learning mechanism to train a DCNN with a small set of dermoscopy images and a novel attention learning mechanism to strengthen the discriminative representation ability of the DCNN via enabling it to focus more on semantically meaningful parts (i.e. lesions) in dermoscopy images. The proposed attention learning mechanism makes full use of the intrinsic and remarkable self-attention ability of classification-trained DCNNs and can work well under any DCNN frameworks without adding any extra attention layers, which is critical for small-sample learning problems like skin lesion classification. From an implementation perspective, both residual learning and attention learning can be embedded in each so-called ARL block. An ARL-CNN model with an arbitrary depth can be constructed by stacking multiple ARL blocks and be trained in an end-to-end manner. We evaluated the proposed ARL-CNN model on the ISIC-skin 2017 dataset [5] which is a largest publicly available skin dermoscopy image achieve, and achieved the state-of-the-art performance (i.e., an average AUC of 0.917).

The main contributions of this paper are thus summarized as follows: (1) we propose a novel ARL-CNN model for accurate skin lesion classification in dermoscopy images, which embeds simultaneously two learning mechanisms - residual learning and attention learning. Residual learning enables the network to become deep, and attention learning helps the network focus more on semantically important regions and thus improves its ability for discriminative representation; (2) we design an effective attention mechanism which takes advantages of the intrinsic self-attention ability of DCNNs, i.e., using the feature maps obtained by a high layer as the attention mask of a low layer, instead of learning the attention mask with extra layers; and (3) we achieve the state-of-the-art skin lesion classification performance on the ISIC-skin 2017 dataset by using a single 50-layer model, which is important for computer-aided diagnosis of skin cancer.

II. RELATED WORK

A. DCNN Models

In recent years, DCNNs have achieved the state-of-the-art performance in many computer vision applications, including image classification [23], [51], [54], [55], target detection [24], [25] and image segmentation [26], [27]. There are two basic operations in DCNNs - convolution and pooling. Convolutional layers apply a set of convolutional kernels to the input with the mechanism of sharing weights, which allows the DCNNs to be deeper with fewer parameters. Pooling layers use various pooling operations, such as the max-pooling, min-pooling, and average-pooling, which are indeed nonlinear down-sampling, to reduce the spatial dimension of the learning representation, number of parameters, and amount of computation.

Many successful DCNN models [48], [49], [50], [23], [51] have demonstrated that their performance depends heavily on the network depth. The increased availability of large

scale datasets, powerful computing devices, and computational tricks, such as the rectified linear units activation [37], dropout [40], batch normalization [38] and layer normalization [39], make it possible to design and train deep networks. However, due to degradation problems [41], [42], it is still hard to train a very deep network. The residual learning technique [23] successfully addressed this issue by using “shortcut connections”, and hence enables the training of a DCNN with as many as 1,000 layers.

B. Attention Mechanism Used in Classification

The attention mechanism is an effective technique that helps a model pay more attention to important information. It has made great progress in the cross fields of computer vision and natural language processing, such as image/video caption and visual question answering [14], [15], [16], [17], [18]. Recently, the attention mechanism has also been successfully used in DCNNs [19], [20], [17] to improve their feature representation ability in large scale image classification tasks. Wang *et al.* [19] proposed a residual attention network, which is constructed by stacking multiple attention modules in a residual network. Each attention module uses trainable layers with a bottom-up and top-down feedforward structure to learn soft weights, and then multiplies the weights with convolutional features. Hu *et al.* [20] recalibrated adaptively channel-wise feature responses by explicitly modelling the channel-wise interdependencies of convolutional features. They generated the attentive features using channel-wise multiplication between the attention weights learned by two additional fully-connected layers and the original feature map. Chen *et al.* [17] jointly used spatial and channel-wise attentions in convolutional networks. The spatial and channel-wise attention weights are generated by a neural network followed by a softmax layer, respectively. Although the attention mechanisms effectively improve the performance of deep learning models in large scale image classification tasks, the attention weights in these methods are learned by using additional learnable layers with a lot of extra parameters, which may cause not only computational costs but also overfitting on small training datasets.

C. Skin Lesion Classification

Many skin lesion classification solutions are based on hand-crafted features, including color, texture, shape, and combined descriptors of lesions [43], [44], [45], [46]. Barata *et al.* [45] proposed a global and local method to classify skin lesions by extracting a set of color and texture features and using them to train a classifier. Xie *et al.* [46] extracted the color, texture, and border descriptors of skin lesions and used an ensemble of neural networks to classify lesions.

To apply deep learning techniques directly to skin lesion classification, a straightforward solution is to collect more training data and fine-tune a powerful pre-trained DCNN [9], [10], [11], [12]. Esteva *et al.* [28] trained a DCNN using 129,450 clinical images for diagnosing the most common and deadliest skin cancers and achieved the performance that matches the performance of 21 board-certified dermatologists.

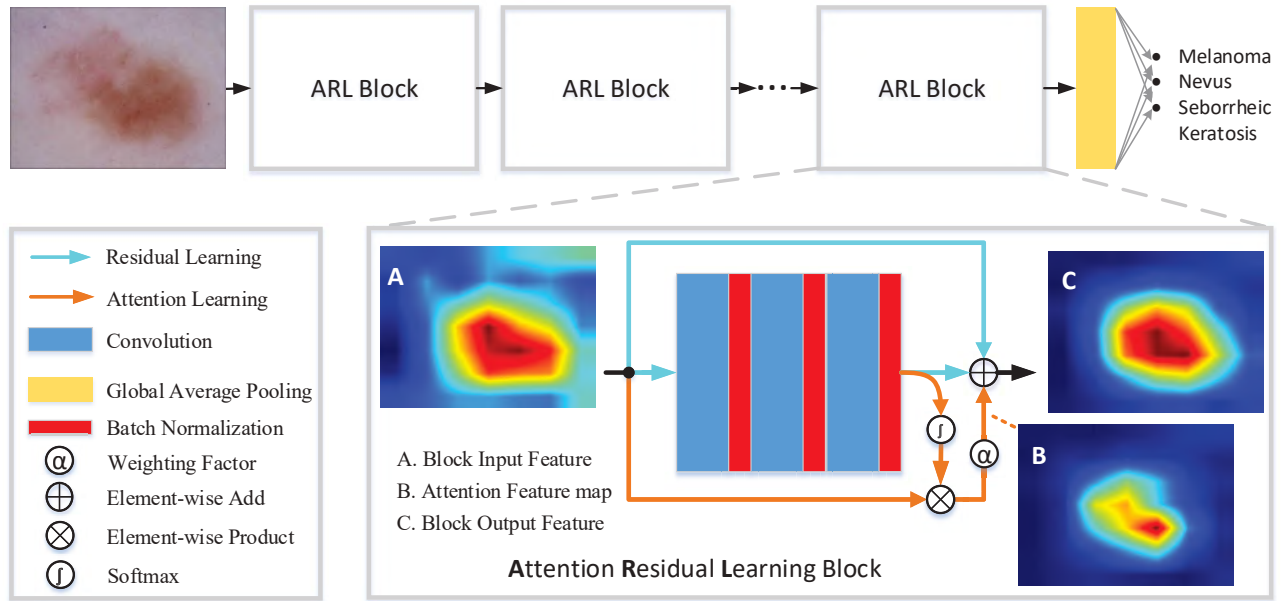


Fig. 2. Architecture of the proposed ARL-CNN model.

Ge *et al.* [29] jointly used the dermoscopy and clinical skin images to train DCNNs and demonstrated the effectiveness of cross-modality learning. Matsunaga *et al.* [31] proposed an ensemble of multiple pre-trained DCNNs with geometrically transformed images for the classification of melanoma, nevus and seborrheic keratosis. Menegola *et al.* [33] also improved the performance of skin lesion classification using pre-trained DCNNs and as much as possible training data.

To filter out the useless background, Yu *et al.* [30] leveraged very deep DCNNs for automated melanoma recognition in two steps - segmentation and classification, and found that lesion segmentation benefits the classification. Diaz *et al.* [32] also designed several convolutional networks that incorporate lesion segmentation and structure segmentation into the diagnosis of skin lesions. However, since the malignancy of skin lesion usually relates only to a part of the lesion, image features extracted in the entire lesion may lead to sub-optimal results. Meanwhile, these methods often requires the ground truth boundaries of lesions being marked in training data, which makes it even more difficult to obtain a large training dataset.

III. METHOD

The proposed ARL-CNN model is composed of multiple ARL blocks, a global average pooling (GAP) layer and a classification layer. In each ARL block, the residual learning mechanism is employed to address the degradation problem, and a novel attention mechanism is designed to strengthen the discriminative representation ability. The architecture of this model is shown in Fig. 2. We now delve into the details of this model.

A. ARL Block

1) *Residual Learning*: Let us consider a plain DCNN block which is composed of a few stacked convolution layers, as

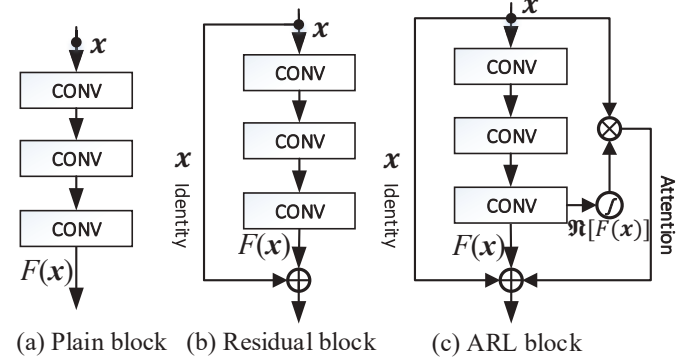


Fig. 3. Three types of learning blocks used in DCNNs.

shown in Fig. 3(a). The underlying mapping fitted by these layers is denoted as $H(x)$. These stacked layers are expected to directly approximate $H(x)$. Hence, a plain block is defined as follows

$$y = F(x, W) \quad (1)$$

where x and y represent the input and output of the block, respectively, and $F(\cdot)$ represents the underlying mapping function which is learned by these stacked layers with the parameter set W . In residual learning, we let these stacked layers learn a residual function $F(x) := H(x) - x$, instead of directly approximating $H(x)$. The definition of residual learning is

$$y = F(x, W) + x \quad (2)$$

where the function $F(\cdot)$ represents the residual mapping learned by these stacked layers. The formulation of residual learning can be implemented by the feedforward network with shortcut connections, as shown in Fig. 3(b).

TABLE I
ARCHITECTURES OF RESNET14, ARL-CNN14, RESNET50 AND ARL-CNN50 MODELS.

Layer name	Output size	ResNet14	ARL-CNN14	ResNet50	ARL-CNN50
Conv1	112x112x64	Conv, 7x7, stride 2	Conv, 7x7, stride 2	Conv, 7x7, stride 2	Conv, 7x7, stride 2
Conv2_x	56x56x256	3x3 max pool, stride 2	3x3 max pool, stride 2	3x3 max pool, stride 2	3x3 max pool, stride 2
		Residual block	ARL	[Residual block]x3	[ARL]x3
Conv3_x	28x28x512	Residual block, stride 2	ARL, stride 2	Residual block, stride 2	ARL, stride 2
				[Residual block]x3	[ARL]x3
Conv4_x	14x14x1024	Residual block, stride 2	ARL, stride 2	Residual block, stride 2	ARL, stride 2
				[Residual block]x5	[ARL]x5
Conv5_x	7x7x2048	Residual block, stride 2	ARL, stride 2	Residual block, stride 2	ARL, stride 2
				[Residual block]x2	[ARL]x2
GAP	1x1x2018	Global Average Pooling	Global Average Pooling	Global Average Pooling	Global Average Pooling
Output	2	Fully Connected	Fully Connected	Fully Connected	Fully Connected

2) *Attention Learning*: The traditional attention mechanisms designed for image classification are to learn the attention weights by using extra learnable layers, such as the convolutional layers used in [19] or the fully connected layers used in [20]. Different from these solutions, we propose a novel attention learning method to strengthen the discriminative representation of the network by generating the attention weights from the classification-trained network itself, without introducing extra learnable layers. Given that higher layers have a better ability for semantic abstraction than lower ones, we suppose that the attention ability of higher layers is stronger than that of lower layers. Hence, we propose to use the more abstract feature maps produced by higher layers as the attention mask of lower layers.

We denote a set of stacked layers in a DCNN as $\{L^{i+1}, \dots, L^{i+n}\}$, where $\{L^{i+1}\}$ is the low layer and $\{L^{i+n}\}$ is the corresponding high layer. The output feature map of $\{L^{i+n}\}$ is defined as \mathbf{O} with a dimension of $H \times W \times C$, where H , W and C represent the height, width and number of channels, respectively. The attention mask ϖ can be generated by applying a normalization function \mathfrak{N} to the feature map \mathbf{O}

$$\varpi = \mathfrak{N}(\mathbf{O}) \quad (3)$$

Three normalization functions are defined as follows

$$\mathfrak{N}^S(\mathbf{O}) = \{m | m_{i,j}^c = \frac{e^{\mathbf{O}_{i,j}^c}}{\sum_{i',j'} e^{\mathbf{O}_{i',j'}^c}}\} \quad (4)$$

$$\mathfrak{N}^C(\mathbf{O}) = \{m | m_{i,j}^c = \frac{e^{\mathbf{O}_{i,j}^c}}{\sum_{c'} e^{\mathbf{O}_{i,j}^{c'}}}\} \quad (5)$$

$$\mathfrak{N}^M(\mathbf{O}) = \{m | m_{i,j}^c = \sigma(\mathbf{O}_{i,j}^c)\} \quad (6)$$

where i, j represent the spatial position, c represents the channel index of \mathbf{O} , and $\sigma(\cdot)$ is the sigmoid function. The spatial attention $\mathfrak{N}^S(\cdot)$ uses a spatial softmax function and highlights the important regions in each channel. The channel attention $\mathfrak{N}^C(\cdot)$ uses a softmax function to perform the normalization in the channel space. The mixed attention $\mathfrak{N}^M(\cdot)$ uses the simple sigmoid normalization at each spatial and channel position.

Since we expect the model to focus on semantic regions of skin lesions, the spatial attention $\mathfrak{N}^S(\cdot)$ is adopted and the comparison given in the Section V.A also shows that the spatial attention performs better than other two attentions in the skin lesion classification task.

Then, we use the attention mask ϖ as the control gates of input neurons \mathbf{x} of $\{L^{i+1}\}$ which are similar to the gates used in the Highway Network [42]. The attention feature map is computed by multiplying \mathbf{x} with ϖ on an element-by-element basis.

$$\mathbf{A} = \varpi \cdot \mathbf{x} \quad (7)$$

This attention learning method strengthens the ability of DCNNs to focus on semantically meaningful regions without adding extra parameters, which is critical for small-sample learning tasks.

3) *ARL Block*: To train a deep DCNN with an improved ability for discriminative representation, we propose the ARL-CNN model, which embeds both residual learning and attention learning mechanisms. The architecture of an ARL block, a basic module of this model, is displayed in Fig. 3(c). It shows that we simultaneously use the identity mapping as designed in the residual block and generate an attention mask for the original input \mathbf{x} via applying the spatial normalization \mathfrak{N}^S to the feature map $F(\mathbf{x})$ obtained by the residual mapping. Then, we get the attention feature map through the following element-wise production

$$\mathbf{A} = \mathfrak{N}[F(\mathbf{x})] \cdot \mathbf{x} \quad (8)$$

The output of an ARL block is an element-wise addition of the identity map, residual feature map, and attention feature map, shown as follows

$$\mathbf{y} = \mathbf{x} + \underbrace{F(\mathbf{x})}_{\text{residual feature map}} + \alpha \cdot \underbrace{\mathfrak{N}[F(\mathbf{x})] \cdot \mathbf{x}}_{\text{attention feature map}} \quad (9)$$

where α is a learnable weighting factor called the scale gate that represents a trade-off between the attention feature map and other two maps. Since a well-trained DCNN has a stronger

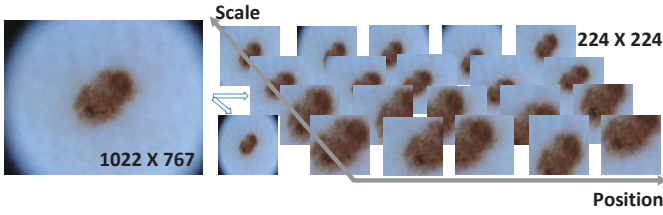


Fig. 4. An illustration of multi-scale patch extraction on dermoscopy images.

attention ability than a poorly trained DCNN, the scale gate is able to adaptively adjust the contribution of the attention feature map, avoiding the interference by a bad attention feature map obtained at the early stages of model training.

B. ARL-CNN

An ARL-CNN model with an arbitrary depth can be easily constructed by stacking ARL blocks. We introduce a lightweight version called ARL-CNN14 and a heavyweight version called ARL-CNN50 for skin lesion classification. ARL-CNN14 has 14 learnable layers, and ARL-CNN50 has 50 learnable layers.

TABLE I gives the architectures of the ResNet14, ARL-CNN14, ResNet50 and ARL-CNN50 models. Both ResNet and ARL-CNN are stacked from a $224 \times 224 \times 3$ input layer, a 7×7 convolutional layer, a max-pooling layer, a series of residual or ARL blocks, a GAP layer and a fully-connected (FC) classification layer. The difference is that the proposed ARL-CNN model replaces the residual blocks used in ResNet with ARL blocks. Each ARL block is stacked by a fixed mode of 1×1 , 3×3 and 1×1 convolutional layers followed by batch normalization layers. ARL-CNN14 can be trained from scratch for skin lesion classification, and ARL-CNN50 can be initialized by using a ResNet50, which has been trained on the ImageNet dataset, and fine-tuned with the skin lesion data. The pre-trained technique can not only improve the performance, but also reduce the training time.

IV. EXPERIMENTS AND RESULTS

A. Dataset

The proposed ARL-CNN model was evaluated on the International Skin Imaging Collaboration 2017 skin lesion classification (ISIC-skin 2017) dataset [5], which is the largest skin dermoscopy image dataset publicly available, consisting of 2000 training, 150 validation, and 600 test images screened for both privacy and quality assurance. Lesions in dermoscopy images are all paired with a gold standard (definitive) diagnosis, i.e. melanoma, nevus, and seborrheic keratosis. There are two binary classification sub-tasks - melanoma classification (melanoma vs. others) and seborrheic keratosis classification (seborrheic keratosis vs. others). We also collected 1320 additional dermoscopy images, including 466 melanoma, 822 nevus images, and 32 seborrheic keratosis images, from the ISIC Archive¹ to enlarge the training dataset.

¹<https://isic-archive.com/>

B. Implementation

Since the proposed ARL-CNN model takes 224×224 images as input, all dermoscopy images should be shrunk to this size before they can be fed into our model. However, skin lesions occupy only a small part of an image, and shrinking the image may lead the lesions to becoming too small to be classified. To address this issue, we randomly extracted 60 rectangular image patches from the central part of each image at different scales ($1/5$, $2/5$, $3/5$, and $4/5$ of original image size) on both official and extra training images, and then resized them to 224×224 using the bilinear interpolation, as shown in Fig. 4. Next, we employed online data augmentation, including random rotation ($[-10^\circ, +10^\circ]$), zoom (90%-110% of width and height), horizontal and vertical flips, to enlarge the training dataset.

The mini-batch SGD algorithm with a batch size of 32 was adopted as the optimizer. The learning rate was initialized to 0.01 for training ARL-CNN14 from scratch and 0.0001 for fine-tuning ARL-CNN50 with pre-trained parameters, and was reduced by half very 30 epochs. The initial weighting factor of the attention feature maps was set to 0.001 in each ARL block when fine-tuning the ARL-CNN50. The maximum epoch number was set to 100. We used the officially provided validation set to monitor the performance of our model and stopped the training process when the network fell into overfitting. In the test stage, we used the same patch extraction method to randomly crop nine patches from each test image, fed them to the trained network, and averaged the obtained scores as the predicted score of the image.

C. Evaluation Metrics

Quantitative evaluation - To quantitatively evaluate the proposed ARL-CNN model, we used the accuracy, sensitivity, specificity, and area under the receiver operating characteristic curve (AUC) as performance metrics, which are defined as

$$accuracy = \frac{TP + TN}{TP + FN + TN + FP} \quad (10)$$

$$sensitivity = \frac{TP}{TP + FN}, specificity = \frac{TN}{TN + FP} \quad (11)$$

$$AUC = \int_0^1 t_{pr}(f_{pr})df_{pr} = P(X1 > X0) \quad (12)$$

where TP , FN , TN and FP represent the number of true positive, false negative, true negative and false positive, respectively, t_{pr} is the true positive rate, f_{pr} is the false positive rate, and $X0$ and $X1$ are the confidence scores for a negative and positive instance, respectively. The AUC value describes the probability that a classifier ranks a randomly chosen positive instance higher than a randomly chosen negative one. The ISIC 2017 skin lesion classification challenge used the AUC value as a gold indicator, according to which all participants were ranked [5].

Qualitative evaluation - We adopted the class activation mapping (CAM) [21] to visualize the attention regions in obtained feature maps. We defined the output of a GAP layer as $f_c(i, j)$, where c represents the index of channel and i, j is

TABLE II
ARCHITECTURES OF RAN14, SENET14, RAN50 AND SENET50.

TABLE V
PERFORMANCE OF ARL-CNN50 IN TWO SUB-TASKS WHEN USING DIFFERENT ATTENTION NORMALIZATION METHODS.

Methods	Melanoma Classification				Seborrheic Keratosis Classification			
	AUC	ACC	Sensitivity	Specificity	AUC	ACC	Sensitivity	Specificity
Channel-wise softmax	0.861	0.847	0.607	0.905	0.952	0.852	0.867	0.849
Mixed Sigmoid	0.851	0.832	0.641	0.878	0.951	0.898	0.811	0.913
Spatial-wise softmax	0.875	0.850	0.658	0.896	0.958	0.868	0.878	0.867

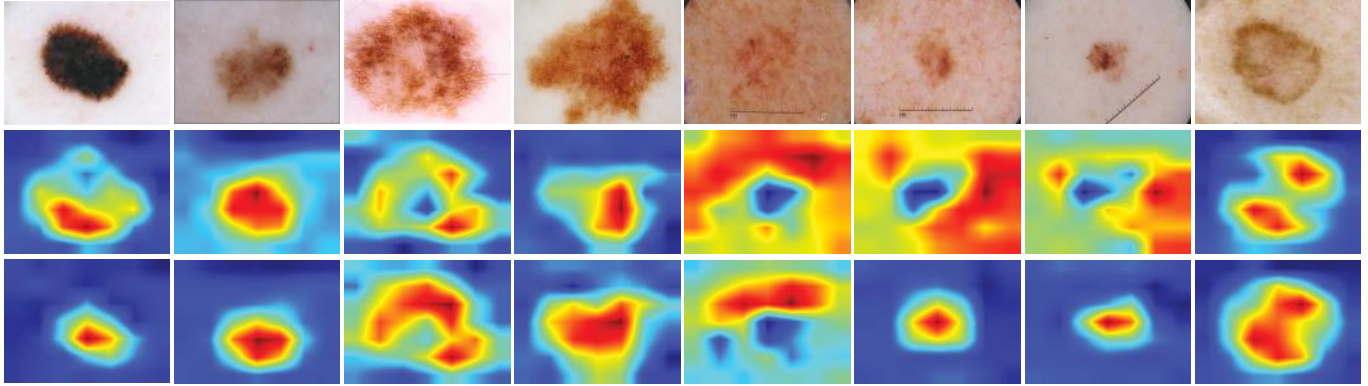


Fig. 5. Visualization of dermoscopy images (top row) and the corresponding CAMs obtained by ResNet50 (middle row) and our ARL-CNN50 (bottom row).

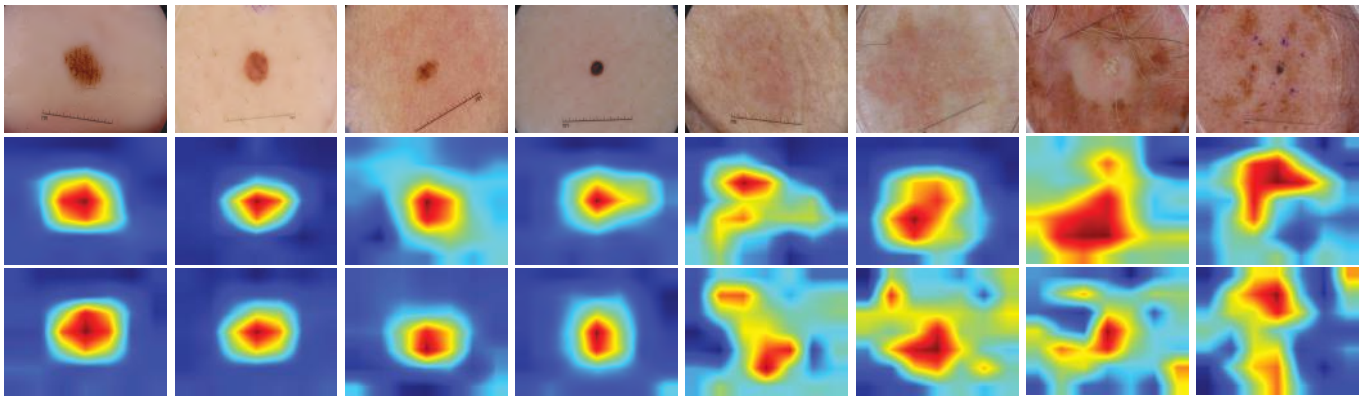


Fig. 6. Eight dermoscopy images (top row) and the corresponding CAMs obtained by ARL-CNN50 when applying it to melanoma classification (middle row) and seborrheic keratosis classification (bottom row), respectively.

representation. To validate this, we visualized the CAMs obtained by ResNet50 and ARL-CNN50 in Fig. 5. It shows that the attention regions learned by both models, i.e. the highlights in CAMs, have different positions and concentrations. Compared to ResNet50, our ARL-CNN50 model shows a stronger attention ability that highlights the discriminative lesion parts instead of background tissues in dermoscopy images, especially in the samples given in the 6th and 7th columns.

Although highlighting background regions, a powerful ResNet50 model could rely on its data fitting ability to easily remember those examples during the training process [47]. However, the generalization ability of such model is poor. We hope that a classification model could pay more attention to the regions, which have better discriminatory power, instead of focusing on irrelevant normal tissues. With a stronger attention

ability, the proposed ARL-CNN model makes more reliable and accurate classification based on discriminative regions. Hence, it may explain why the proposed ARL-CNN model has a better classification performance than the corresponding ResNet model.

V. DISCUSSION

A. Attention Mask Normalization

In the proposed ARL-CNN model, we use the spatial softmax, which normalizes a soft mask in the spatial domain, to generate the soft attention mask ϖ from the high-layer feature map O . Certainly, other mask normalization methods can also be used. In Table V, we compared the classification performance of ARL-CNN50 in two sub-tasks when using different attention soft mask normalization methods, including the channel-wise softmax, mixed sigmoid, and spatial-wise

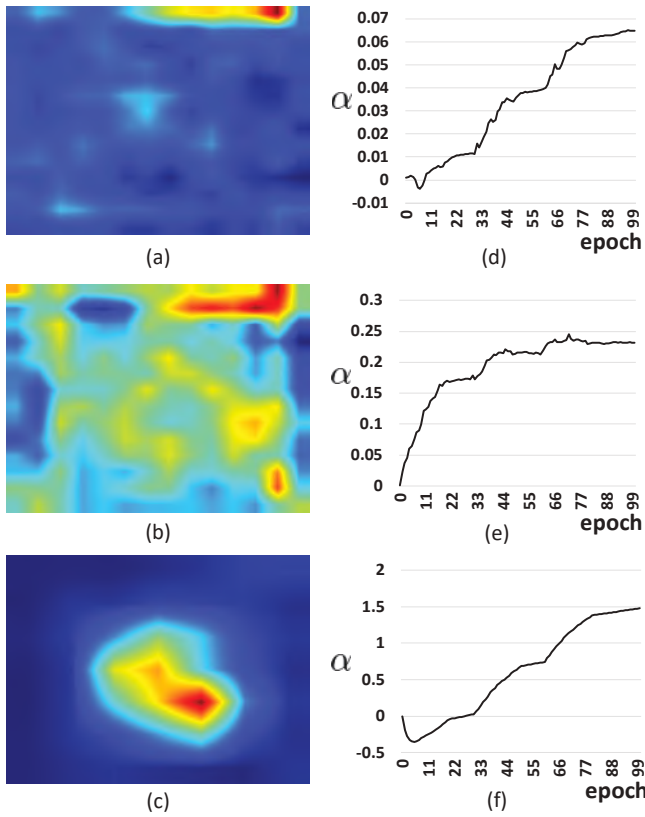


Fig. 7. Visualization of the feature maps in low, middle and high layers and the corresponding curves of weighting factors during the training process: (a) feature map in the 13th layer, (b) feature map in the 31st layer, (c) feature map in the 49th layer, (d) curve of weighting factors in the 13th layer, (e) curve of weighting factors in the 31st layer, and (f) curve of weighting factors in the 49th layer.

softmax. It shows that the spatial softmax method achieved the best performance in most metrics.

B. Attention Regions Learned from Different Tasks

In Fig. 6, we displayed eight dermoscopy images and the corresponding CAMs learned by ARL-CNN50 for melanoma classification and seborrheic keratosis classification, respectively. It reveals that, if the skin lesion is small (see left four images), the CAMs learned for both tasks show similar attention regions, which highlight the skin lesions; otherwise (see right four images), the CAMs learned for both tasks demonstrate different highlights. Therefore, the proposed ARL-CNN model is able to adaptively focus on task-related semantic regions when the skin lesion is large.

C. Weighting Factor of Attention Learning

As shown in Eq. (9), the output of each ARL block is a weighted sum of the identity map, residual feature map, and attention feature map. The weighting factor α , which is also a learnable parameter, represents the trade-off between the attention feature map and other two maps. We displayed the attention feature maps produced by the low (13-layer), middle (31-layer) and high layers (49-layer) of the proposed ARL-CNN50 model in the left column of Fig. 7. It shows that

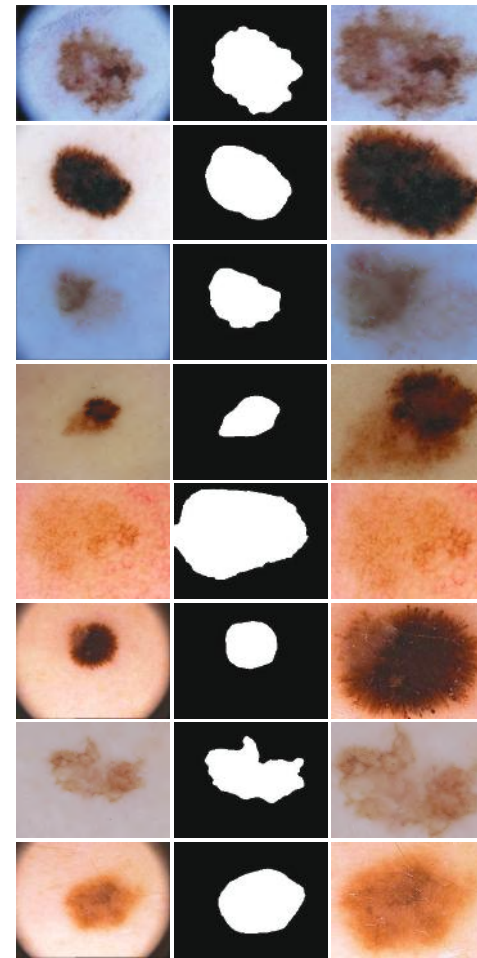


Fig. 8. Eight examples of (left column) dermoscopy images and the corresponding (middle column) segmentation masks and (right column) ROIs cropped according to segmentation masks.

TABLE VI
CLASSIFICATION PERFORMANCE OF RESNET50 AND ARL-CNN50 WITH / WITHOUT SEGMENTATION. (M: MELANOMA CLASSIFICATION, SK: SEBORRHEIC KERATOSIS CLASSIFICATION).

Methods	Segm	AUC of M	AUC of SK	Average AUC
ResNet50	×	0.857	0.948	0.903
ResNet50	✓	0.864	0.955	0.910
ARL-CNN50	×	0.875	0.958	0.917
ARL-CNN50	✓	0.876	0.96	0.918

a higher layer has a stronger attention ability than a lower layer. In the right column of Fig. 7, we plotted the variation of the learned value of α versus the epoch during the training process. Generally, the weighting factor α increases during training in all three cases. However, a high layer may have a large weighting factor α , and accordingly the attention feature map learned by a high layer makes more contributions to the classification process.

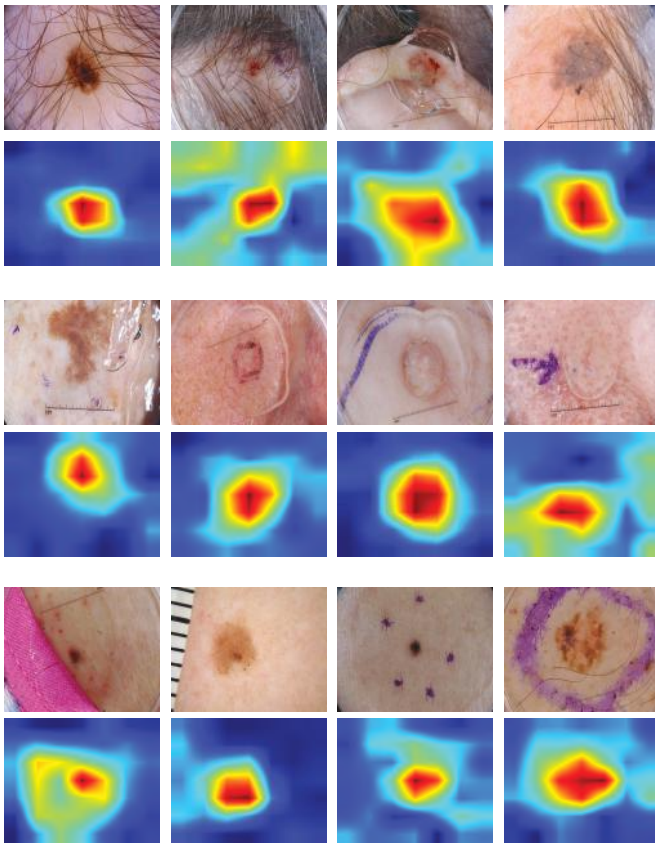


Fig. 9. Visualization of dermoscopy images (above) which contains some artifacts, such as hair and bubbles, and the corresponding CAMs (bellow) obtained by the proposed ARL-CNN50 model.

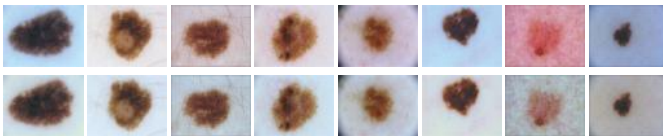


Fig. 10. Some typical examples of dermoscopy images normalized by using color constancy algorithm. Top row: original images; bottom row: normalized images.

D. Effect of Lesion Segmentation

In the above experiments, the lesions were not segmented before applying the images to the proposed ARL-CNN model and directly patches were extracted. To evaluate the effect of lesion segmentation, we also compared the baseline model ResNet50 and our proposed ARL-CNN50 model, both of which were trained with the segmented skin lesions - region of interest (ROI). To obtain an accurate segmentation mask of a lesion, we trained a state-of-the-art segmentation model deeplabV3+ [52] by using the ISIC 2017 Skin Lesion Segmentation dataset [5]. Some segmentation examples are shown in Fig. 8. We randomly extracted image patches in the ROI at different scales as introduced in section IV.B. In TABLE VI, we compared the AUC obtained by the ResNet50 model and our ARL-CNN50 model with and without segmentation in the melanoma classification and seborrheic keratosis classification. It shows that training the ResNet50 model based

on the segmented ROI can effectively resolve the background noise and generated more discriminative features for better classification, resulting in an improvement of the average AUC from 0.903 to 0.910. However, training the proposed ARL-CNN50 model based on the segmented ROI only slightly improved the average AUC from 0.917 to 0.918. The reason can be attributed to the effective attention learning mechanism which enables our ARL-CNN model to focus more on semantically meaningful parts and to achieve the comparable performance without segmentation. Considering that the pixel-wise segmentation has little effect on classification, but a rather high computational cost, we did not segment lesions before applying the images to the proposed model.

E. Attention ability to the artifacts

The classification of skin lesions may suffer from the presence of artifacts, including natural hairs and artificial air bubbles. Some skin lesions may be partly obscured or covered by these artifacts, as shown in Fig. 9. To evaluate the effect of our attention mechanism on these artifacts, we visualized the CAMs obtained by the proposed ARL-CNN50 model for these skin lesions in Fig. 9. It shows that the attention is still paid to the skin lesions, instead of hairs, bubbles, or other artifacts. The robustness of learned attention to these artifacts is significantly important for the accurate lesion classification.

F. Normalization of Luminance and Color

We exploited the gray-world color constancy algorithm [53] to normalize the luminance and color in dermoscopy images. As shown in Fig. 10, the top row is original dermoscopy images, and the second row is corresponding normalized images. We trained the proposed ARL-CNN50 model with these normalized images and achieved an average AUC of 0.918. It shows that our ARL-CNN model is able to achieve the comparable performance without normalizing the luminance and color, which shows a good generalization ability on diversified images.

G. Robutness Analysis

We evaluated the proposed ARL-CNN50 model and the baseline ResNet50 model 10 times independently ($m = n = 10$) and obtained 10 average AUC values for each model, which were listed in Table VII. We assumed that the average AUC values of ResNet50 and ARL-CNN50 are random variables X and Y , respectively, each following a Gaussian distribution, i.e. $X \sim N(\mu_1, \sigma_1^2)$, $Y \sim N(\mu_2, \sigma_2^2)$, and $\sigma_1^2 = \sigma_2^2$. We adopted the independent two-sample t -test to determine whether the ARL-CNN50 significantly improves the classification performance. The hypotheses to be tested are $H_0 : \mu_1 \geq \mu_2$ versus $H_1 : \mu_1 < \mu_2$. Given the significance level $\alpha = 0.01$, $t_{0.01}(10 + 10 - 2) = 2.552$, and we have a rejection domain $W = \{t \leq -2.552\}$. According to Table VII, we have

$$\begin{cases} \bar{x} = 90.26, & \bar{y} = 91.66 \\ s_1^2 = \frac{1}{m-1} \sum_{i=1}^m (x_i - \bar{x})^2 = \frac{0.0224}{9} \\ s_2^2 = \frac{1}{n-1} \sum_{j=1}^n (y_j - \bar{y})^2 = \frac{0.0064}{9} \end{cases} \quad (14)$$

- based Localization,” in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 618-626.
- [23] K. He, X. Zhang, S. Ren and J. Sun, ”Deep Residual Learning for Image Recognition, in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770-778.
- [24] K. He, G. Gkioxari, P. Dollr, R. Girshick. ”Mask R-CNN,” in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 2980-2988.
- [25] T. Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollar. ”Focal Loss for Dense Object Detection,” in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 2999-3007.
- [26] J. Long, E. Shelhamer, T. Darrell. ”Fully convolutional networks for semantic segmentation,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 3431-3440.
- [27] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille. ”DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 834-848, 2018.
- [28] A. Esteva, B. Kuprel, R. A. Novoa, J. Ko, S. M. Swetter, H. M. Blau, and S. Thrun, Corrigendum: Dermatologist-level classification of skin cancer with deep neural networks, *Nature*, vol. 542, no. 7639, pp. 115, 2017.
- [29] Z. Ge, S. Demyanov, R. Chakravorty, A. Bowling, R. Garnavi. ”Skin Disease Recognition Using Deep Saliency Features and Multimodal Learning of Dermoscopy and Clinical Images,” in *Proc. MICCAI*, 2017, pp. 250-258.
- [30] 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, 2017.
- [31] K. Matsunaga, A. Hamada, A. Minagawa, H. Koga. (Mar. 2017) ”Image Classification of Melanoma, Nevus and Seborrheic Keratosis by Deep Neural Network Ensemble,” [Online]. Available: <https://arxiv.org/abs/1703.03108>
- [32] I. G. Diaz. (Mar. 2017) ”Incorporating the Knowledge of Dermatologists to Convolutional Neural Networks for the Diagnosis of Skin Lesions,” [Online]. Available: <https://arxiv.org/abs/1703.01976>
- [33] A. Menegola, J. Tavares, M. Fornaciali, L. T. Li, S. Avila, E. Valle. (Mar. 2017) ”RECOD Titans at ISIC Challenge 2017,” (Online). Available: <https://arxiv.org/pdf/1703.04819>
- [34] L. Bi, J. Kim, E. Ahn, and D. D. Feng. (Mar. 2017) ”Automatic Skin Lesion Analysis using Large-scale Der-moscopy Images and Deep Residual Networks,” [Online]. Available: <https://arxiv.org/abs/1703.04197>
- [35] X. Yang, Z. Zeng, S. Y. Yeo, C. Tan, H. L. Tey, Y. Su. (Mar. 2017) ”A Novel Multi-task Deep Learning Model for Skin Lesion Segmentation and Classification,” [Online]. Available: <https://arxiv.org/abs/1703.01025>
- [36] T. DeVries, D. Ramachandram. (Mar. 2017) ”Skin Lesion Classification Using Deep Multi-scale Convolutional Neural Networks,” [Online]. Available: <https://arxiv.org/pdf/1703.01402>
- [37] V. Nair, and G. E. Hinton. ”Rectified linear units improve restricted boltzmann machines,” in *Proc. Int. Conf. Mach. Learn.*, 2010, pp. 807-814.
- [38] S. Ioffe, and C. Szegedy. ”Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift,” in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 448-456.
- [39] J. L. Ba, J. R. Kiros, and G. E. Hinton. (Jul. 2016) ”Layer Normalization”. [Online]. Available: <https://arxiv.org/abs/1607.06450>
- [40] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, Dropout: a simple way to prevent neural networks from overfitting, *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 1929-1958, 2014.
- [41] K. He and J. Sun. ”Convolutional neural networks at constrained time cost,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 5353-5360.
- [42] R. K. Srivastava, K. Greff, and J. Schmidhuber. ”Training Very Deep Networks,” in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 2377-2385.
- [43] H. Ganster, P. Pinz, R. Rohrer, E. Wildling, M. Binder, and H. Kittler. ”Automated melanoma recognition,” *IEEE Trans. Med. Imag.*, vol. 20, no. 3, pp. 233239, Mar. 2001.
- [44] M. E. Celebi et al., ”A methodological approach to the classification of dermoscopy images,” *Comput. Med. Imag. Graph.*, vol. 31, no. 6, pp. 362373, 2007.
- [45] C. Barata, M. Ruela, M. Francisco, T. Mendona, and J. S. Marques. ”Two Systems for the Detection of Melanomas in Dermoscopy Images Using Texture and Color Features,” *IEEE Sys. J.*, vol. 8, no. 3, Sep. 2014.
- [46] 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, Mar. 2017.
- [47] C. Zhang, S. Bengio, M. Hardt, B. Recht, O. Vinyals. ”Understanding deep learning requires rethinking generalization,” in *Proc. Int. Conf. Learn. Rep.*, 2017.
- [48] A. Krizhevsky, I. Sutskever, and G. Hinton. ”ImageNet Classification with Deep Convolutional Neural Networks,” in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 1106-1114.
- [49] K. Simonyan, and A. Zisserman. ”Very Deep Convolutional Networks for Large-Scale Image Recognition,” in *Proc. Int. Conf. Learn. Rep.*, 2015.
- [50] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. ”Going deeper with convolutions,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 1-9.
- [51] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger. ”Densely Connected Convolutional Networks,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 2261-2269.
- [52] L. C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam. ”Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation”, in *Proc. European Conference on Computer Vision (ECCV)*, 2018.
- [53] C. Barata, M. E. Celebi, and J. S. Marques. ”Improving Dermoscopy Image Classification Using Color Constancy”, in *IEEE Journal of Biomedical and Health Informatics*, vol. 19, no. 3, pp. 11461152, 2015.
- [54] Y. Xie, Y. Xia, J. Zhang, Y. Song, D. D. Feng, M. Fulham, W. Cai, ”Knowledge-based Collaborative Deep Learning for Benign-Malignant Lung Nodule Classification on Chest CT”, *IEEE Trans. Med. Imag.*, doi: 10.1109/TMI.2018.2876510, 2018.
- [55] J. Zhang, Y. Xie, Q. Wu, Y. Xia, ”Skin lesion classification in dermoscopy images using synergic deep learning”, in *Proc. MICCAI*, 2018, pp. 12-20.