Deep metric attention learning for skin lesion classification in dermoscopy images

Xiaoyu He¹ · Yong Wang¹ · Shuang Zhao² · Chunli Yao³

Received: 17 April 2021 / Accepted: 31 October 2021 / Published online: 4 January 2022
© The Author(s) 2021

Abstract

Currently, convolutional neural networks (CNNs) have made remarkable achievements in skin lesion classification because of their end-to-end feature representation abilities. However, precise skin lesion classification is still challenging because of the following three issues: (1) insufficient training samples, (2) inter-class similarities and intra-class variations, and (3) lack of the ability to focus on discriminative skin lesion parts. To address these issues, we propose a deep metric attention learning CNN (DeMAL-CNN) for skin lesion classification. In DeMAL-CNN, a triplet-based network (TPN) is first designed based on deep metric learning, which consists of three weight-shared embedding extraction networks. TPN adopts a triplet of samples as input and uses the triplet loss to optimize the embeddings, which can not only increase the number of training samples, but also learn the embeddings robust to inter-class similarities and intra-class variations. In addition, a mixed attention mechanism considering both the spatial-wise and channel-wise attention information is designed and integrated into the construction of each embedding extraction network, which can further strengthen the skin lesion localization ability of DeMAL-CNN. After extracting the embeddings, three weight-shared classification layers are used to generate the final predictions. In the training procedure, we combine the triplet loss with the classification loss as a hybrid loss to train DeMAL-CNN. We compare DeMAL-CNN with the baseline method, attention methods, advanced challenge methods, and state-of-the-art skin lesion classification methods on the ISIC 2016 and ISIC 2017 datasets, and test its generalization ability on the PH2 dataset. The results demonstrate its effectiveness.

Keywords Deep metric learning · Attention mechanism · Skin lesion classification · Dermoscopy images

Introduction

Skin diseases are one of the most common disasters among people, which occur in all cultures and ages and affect 30–70% of the people's health [1]. Regarding skin cancer, the figures become more serious. Skin cancer is the most common cancer in America [2]. Current estimates are that one in five Americans will develop skin cancer in their lifetime

[3,4]. Dermoscopy is a non-invasive skin imaging technique, which has been widely used by dermatologists to diagnose skin lesions [5]. However, manual interpretation of dermoscopy images is usually time-consuming and subjective. Therefore, automatic classification of skin lesions based on dermoscopy images deserves in-depth research.

Traditional approaches often use human-designed features (such as color, geometry, or boundary) and machine learning classifiers for skin lesion classification [6–8]. Their main disadvantages are the heavy feature design process and weak generalization performance. Compared with traditional approaches, approaches based on convolutional neural networks (CNNs) can directly learn meaningful features from data and have achieved significantly improved performance [9–13]. For example, Esteva et al. [9] utilized the CNN architecture of Inception-V3 [14] to train a skin cancer classification model and achieved the performance on par with 21 board-certified dermatologists. Haenssle et al. [13] trained a melanoma recognition model using the Inception-

✉ Yong Wang
ywang@csu.edu.cn

✉ Shuang Zhao
shuangxy@csu.edu.cn

¹ School of Automation, Central South University, Changsha 410083, China

² Department of Dermatology, Xiangya Hospital, Central South University, Changsha 410008, China

³ Department of Dermatology, The Second Hospital, Jilin University, Changchun 130041, China

Fig. 1 Some dermoscopy image samples of nevus, seborrheic keratosis, and melanoma

V4 CNN architecture [15] and reported that the diagnostic performance of this model outperforms 58 professional dermatologists from 17 countries.

Although approaches based on CNNs have shown promising results in this field, precise skin lesion classification is still challenging because of the following three issues. First, the number of skin lesion images used for training is usually insufficient, which limits the success of CNNs. For example, the ImageNet benchmark dataset has tens of millions of images, while a skin lesion dataset usually only has thousands of images. Second, the classification of skin lesions is affected by inter-class similarities and intra-class variations [16,17]. Figure 1 presents three kinds of skin diseases: nevus, seborrheic keratosis, and melanoma. For each kind of skin disease, there are six dermoscopy images. From Fig. 1, dermoscopy images of different skin diseases could have visual similarities in color or geometry, while dermoscopy images of the same skin disease may have obvious visual differences. Third, the skin lesion area only takes up a fraction of the dermoscopy image, while large area of the dermoscopy image is occupied by regular tissue or other visual obstacles which are irrelevant to the disease recognition and may interfere with the recognition results. Considering these issues, the classification of skin lesions in dermoscopy images is usually more difficult than that of objects and scenes in natural images [18,19]. Therefore, it is necessary to develop more effective techniques to improve the performance of skin lesion classification.

The attention mechanism is a useful tool to guide a classification model focus more on key information. It has brought many successful applications in pattern recognition and machine translation. Recently, attention-based methods have also begun to appear in the classification of skin lesions. For example, Zhang et al. [10] proposed an attention residual learning CNN (ARL-CNN). They showed that ARL-CNN with a spatial-wise self-attention mechanism can attentively concentrate on discriminative skin lesion parts. As a result, the attention mechanism has already been verified to have the potential to address the above-mentioned third issue.

Metric learning is a learning paradigm that learns a metric function from training data by calculating the similarities or distances among samples [20]. Deep metric learning is the combination of deep learning and metric learning [21]. It aims at learning semantic embeddings of pairs or triplets of samples, and these embeddings are optimized to increase the distance between samples with different classes and decrease the distance between samples with the same class. In addition, the combinations of pairs or triplets are various; thus, the number of training samples can be largely increased. In this manner, deep metric learning has the potential to address the above-mentioned first and second issues.

Based on the above considerations, this paper proposes a deep metric attention learning CNN (called DeMAL-CNN) for skin lesion classification in dermoscopy images. In DeMAL-CNN, a triplet-based network (TPN) is first designed based on deep metric learning, whose input is a triplet of skin lesion images and outputs are their embeddings. Specifically, TPN consists of three weight-shared embedding extraction networks, each of which is stacked by a convolutional layer, a max pooling layer, a series of mixed attention residual learning (MARL) blocks, a global average pooling (GAP) layer, and a fully connected (FC) layer. After extracting the embeddings, three weight-shared classification layers are used to generate the final predictions based on these embeddings. The main contributions of this paper are summarized as follows:

- DeMAL-CNN integrates the advantages of the attention mechanism and deep metric learning for precise skin lesion classification. To the best of our knowledge, DeMAL-CNN is the first attempt to jointly use the attention mechanism and deep metric learning in automatic skin disease classification.
- DeMAL-CNN has the capability to address the aforementioned three issues in skin lesion classification. For the first issue, since TPN in DeMAL-CNN takes the triplet as input and the combinations of triplets are various, the number of training samples can be largely

increased. To address the second issue, a triplet loss is adopted to learn a metric space, where the output embeddings are optimized to pull samples with different classes farther away from each other and push samples with the same class closer to each other. In this way, the optimized embeddings are more robust to inter-class similarities and intra-class variations. Regarding the third issue, an effective mixed attention mechanism is designed and integrated into the residual blocks to construct our MARL blocks. This mixed attention mechanism considers the important information from both spatial-wise and channel-wise, and is able to enable the extracted embeddings concentrate more on discriminative skin lesion parts.

- Extensive experiments have been carried out on the ISIC 2016, ISIC 2017, and PH2 datasets to investigate the effectiveness of DeMAL-CNN. The results demonstrate that DeMAL-CNN achieves the state-of-the-art performance on the ISIC 2016 and ISIC 2017 datasets against all its competitors, and has good generalization performance on the PH2 dataset.

Related work

Deep learning-based skin lesion classification

Inspired by the remarkable achievements of deep learning methods in computer vision tasks, many researchers have applied deep learning methods to the classification of skin lesions. Some existing methods focus on improving the deep model's ability to locate skin lesions. Yu et al. [16] proposed a two-stage framework for automatic melanoma segmentation and classification. They found that the use of segmentation results can improve the performance of melanoma classification. Yang et al. [22] proposed a CNN framework with region average pooling for melanoma classification, and demonstrated the effectiveness of making feature extraction focus more on the region of interest (i.e., skin lesions). Xie et al. [23] proposed a mutual bootstrapping CNN (MB-CNN) to simultaneously segment and classify skin lesions, in which the segmentation network can provide the lesion masks for classification network to further boost the classification performance.

In addition, some methods have been proposed to alleviate the issue of insufficient training samples in skin lesion classification. Xie et al. [24] proposed a semi-supervised adversarial classification model (SSAC) to make full use of both labeled and unlabeled data for skin lesion classification. Qin et al. [25] proposed a skin lesion classification framework with a GAN-based data augmentation strategy, and showed that the synthetic dermoscopy images generated by GANs

can facilitate the performance of skin lesion classification with limited training samples.

Besides, some methods have been developed to deal with the issue of inter-class similarities and intra-class variations. Zhang et al. [26] proposed a synergic deep learning (SDL) method for skin lesion classification, which employs multiple CNNs to simultaneously learn discriminative feature representations from image pairs. Tang et al. [27] designed a global-part CNN (GP-CNN) for skin lesion classification, which consists of a G-CNN and a P-CNN. They trained G-CNN with downsampled images for global dermoscopy image feature extraction, and used multiscale image patches to train P-CNN to capture local fine-grained information.

Based on the above introduction, we find that existing methods usually only focus on addressing one issue in skin lesion classification to improve the performance, such as lack of the ability to concentrate on discriminative skin lesion parts, insufficient training samples, or inter-class similarities and intra-class variations. Therefore, to further improve the performance of skin lesion classification, a unified framework that can address the above three issues remains to be proposed.

Attention mechanism

The attention mechanism in deep learning is similar to the human visual attention mechanism, whose goal is to select information that is more critical to the current task from a lot of information. It has brought many breakthroughs to the fields of natural language processing [28–30] and computer vision [31,32]. Recently, many attention-based methods have been designed to strengthen the feature representation capabilities of CNNs in image classification tasks [33–35]. For example, Wang et al. [33] proposed a residual attention network (RAN) with a novel residual attention module. In this module, the attention weights are learned by the trainable convolutional layers, and then the attention feature map is obtained by implementing the multiplication between the attention weights and the convolutional feature map. Hu et al. [34] proposed a novel channel-wise attention network (SENet), which consists a series of squeeze and excitation blocks. In each SE block, the attention feature map is generated by performing the multiplication between the attention vector learned by two sequential FC layers and the input feature map.

In skin lesion classification, many attention-based approaches have been proposed. ARL-CNN [10] is one of the representatives. Different from RAN [33] and SENet [34] that learn attention weights via additional learnable layers, the attention weights in ARL-CNN can be generated by the training of its network, thus saving the computational costs

and having the advantage in avoiding overfitting on small training sets. Moreover, ARL-CNN has a strong capability to adaptively focus on discriminative skin lesion parts. However, one limitation in ARL-CNN is that it only considers the spatial-wise attention and ignores the important information in the channels. Therefore, it is necessary to make full use of the important information from both spatial-wise and channel-wise in skin lesion classification.

Deep metric learning

Deep metric learning is proposed by combining deep learning with metric learning [21], which employs the deep models to learn the similarity relationships among samples. The triplet network [36] is one of the representatives. As its name implies, a triplet network takes three samples as input, including an anchor sample, a positive sample (with the same class but not the same image as the anchor sample), and a negative sample (with a different class from the anchor sample). By calculating the distances among samples, triplet networks can learn to decrease the distance between the anchor and positive samples while increase the distance between the anchor and negative samples, thus having the potential to address the issue of inter-class similarities and intra-class variations. Moreover, since the combinations of triplets are various, the number of training samples can be largely increased. Due to these good properties, triplet networks have received widespread attention in image analysis tasks. For example, He et al. [37] proposed a triplet network with a hybrid loss, which combines triplet loss with center loss for 3D image retrieval. Lim et al. [38] designed a 3D image style detection model using a triplet network. Schroff et al. [39] proposed a novel face recognition model with an online triplet learning mechanism, which achieves outstanding recognition performance on both the Labeled Faces in the Wild dataset and YouTube Faces dataset.

However, these triplet-based approaches are designed for the recognition of objects in natural images or human faces. Compared with natural images or human faces, different skin lesions have very subtle differences in the image, so it is more difficult to identify. Therefore, to apply the triplet network to skin lesion classification, it is necessary to modify the structure of the triplet network to enhance its discriminative representation ability.

Proposed approach

Overview

Based on the analysis in Sect. [Related Work](#), we propose DeMAL-CNN for the classification of skin lesions, which integrates the advantages of both the attention mechanism

and deep metric learning and overcomes their disadvantages. In DeMAL-CNN, we design a mixed attention mechanism that can learn the attention information from both spatial-wise and channel-wise, and integrate it into the construction of a triplet-based network to enable the extracted embeddings concentrate more on discriminative skin lesion parts.

Figure 2 shows the architecture of DeMAL-CNN, which is composed of TPN and three weight-shared classification layers. Specifically, TPN works for the embedding extraction of a triplet of skin lesion images (i.e., a positive sample, an anchor sample, and a negative sample), and consists of three weight-shared embedding extraction networks. Each embedding extraction network is stacked by a convolutional layer, a max pooling layer, a series of MARL blocks, a GAP layer, and a FC layer. After the embedding extraction of the input triplet, on the one hand, these extracted embeddings are used to calculate the triplet loss; on the other hand, they are sent to three weight-shared classification layers to generate the final predictions. In the training procedure, we combine the triplet loss with the classification loss as a hybrid loss to train DeMAL-CNN. Next, we will first introduce the details of the proposed MARL block, then describe how to build the embedding extraction network based on the MARL blocks, and finally illustrate the hybrid loss function.

MARL block

Regarding current attention-based skin lesion classification approaches, the building blocks in their CNNs are constructed based on either the spatial-wise attention mechanism [10,33] or the channel-wise attention mechanism [34]. Different from these approaches, we propose a mixed attention mechanism and integrate it into the residual block to construct our MARL block. For the convenience of introducing the proposed MARL block in this paper, the residual block [40] is introduced first.

(1) *Residual block*: The residual block is the basic building block of ResNet [40] and is designed based on the residual learning mechanism. Figure 3a shows a residual block which is stacked by multiple convolutional, batch normalization (BN), and ReLU layers. For simplicity, the BN and ReLU layers are not presented in Fig. 3a. Assuming that x is the input of the residual block, the output (denoted as y) of the residual block is calculated according to the residual learning mechanism as follows:

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

where $F(\cdot)$ denotes the residual mapping formed by the stacked layers in the residual block, and W represents the weight of these stacked layers. In the forward propagation, the input signal can directly propagate from an arbitrary lower layer to an upper layer. As a result, the CNNs stacked by

Fig. 2 Architecture of DeMAL-CNN

Fig. 3 Two types of learning blocks used in CNNs. **a** Residual block. **b** The proposed MARL block

multiple residual blocks can address the vanishing gradient problem.

(2) *MARL block*: The MARL block is constructed based on the residual block and the proposed mixed attention mechanism, as shown in Fig. 3b. From the implementation perspective, the mixed attention mechanism is realized by two modules: a spatial-wise attention module and a channel-wise attention module.

First, we introduce the spatial-wise attention module. As shown in Fig. 3b, this module receives x and the feature map (denoted as $F(x)$) as input and generates the spatial-wise attention feature map (denoted as A_s) through the following two steps: (1) perform the spatial normalization function on $F(x)$ to generate an attention weight (denoted as $N_s(F(x))$), and (2) generate A_s through the element-wise multiplication between $N_s(F(x))$ and x . The above two steps can be defined as follows:

$$A_s = N_s(F(x)) \cdot x, \quad (2)$$

where $N_s(\cdot)$ represents the softmax operation that performs normalization in the spatial space. The reason why $N_s(F(x))$ can be used as the attention weight to refine x is due to the following two reasons: (1) $N_s(F(x))$ is obtained by x through the residual mapping followed by a softmax function, where the stacked layers in the residual block form a mapping from x to its attention weight; and (2) $N_s(F(x))$ and x are with the same spatial size and number of channels.

Second, we illustrate the channel-wise attention module, the design of which is derived from the squeeze and excitation block in SENet [34]. As shown in Fig. 3b, it is composed of a GAP layer, a FC layer, a ReLU layer, a FC layer, and a Sigmoid layer. Among these layers, the GAP layer first performs the squeeze operation which downsamples the spatial size of $F(x)$ to 1×1 . The aim of this squeeze operation is to generate a vector that describes the statistical information of $F(x)$ in the channel dimension. Next, the remaining layers work together to implement the excitation operation. The excitation operation is similar to the gating mechanism in recurrent neural networks [41]. Specifically, the two FC layers are used to explicitly model the correlations in the channel (i.e., the importance of each channel), where the first FC layer followed by a ReLU layer reduces the number of channels by 16 times, and the second FC layer followed by a Sigmoid layer restores the original number of channels. Suppose that the weights of the first and second FC layers are W_1 and W_2 , respectively, the output channel-wise attention weight (denoted as ω) by the excitation operation is defined as follows:

$$\omega = \sigma(W_2 \text{ReLU}(W_1 z)), \quad (3)$$

where z is the output of the squeeze operation and $\text{ReLU}(\cdot)$ and $\sigma(\cdot)$ represent the activation of ReLU and Sigmoid, respectively. Then, the final channel-wise attention feature map (denoted as A_c) is generated by performing the channel-wise multiplication between ω and $F(x)$:

$$A_c^i = \omega^i \cdot F(x)^i, i \in [1, C], \quad (4)$$

where ω^i , $F(x)^i$, and A_c^i represent the feature map components of ω , $F(x)$, and A_c in each channel, respectively, and C represents the number of channels of ω or $F(x)$.

Finally, the output of the MARL block is the element-wise addition of x , A_c , and A_s :

$$y = x + A_c + \alpha \cdot A_s. \quad (5)$$

where α is a learnable weighting factor. Note that α can automatically adjust the contribution of the channel-wise and spatial-wise attention feature maps during the model training process.

From the above descriptions, we can observe that our MARL block embeds both the residual learning mechanism and the mixed attention mechanism. On the one hand, the residual learning mechanism can avoid the problem of vanishing gradient; thus, the network can be stacked deeper. On the other hand, the mixed attention mechanism can help the network pay attention to the important information not only from the spatial-wise, but also from the channel-wise, thus further enhancing the skin lesion localization ability of CNNs in skin lesion classification tasks.

Embedding extraction network

To create deeper CNNs to obtain stronger feature representation capabilities, a common practice is to stack multiple basic building blocks in the network. In this paper, we take the proposed MARL block as the basic building block to construct our embedding extraction network. Specifically, we follow the macro structure of ResNet50 [40] and build the embedding extraction network by replacing the residual blocks with the MARL blocks. Table 1 presents the architectures of ResNet50 and our embedding extraction network. Both of them are composed of a 7×7 convolutional layer, a 3×3 max pooling layer, a series of residual or MARL blocks, a GAP layer, and an output FC layer. Note that the output dimension of ResNet50 is two (a binary classification task), while that of the embedding extraction network is 256 (with the embedding dimension of 256). To investigate the effect of the embedding dimension, Sect. [Dimension of the Embedding Layer](#) tests the classification performance of DeMAL-CNN with different embedding dimensions (e.g., 1024, 512, 256, 128, and 64).

Table 1 Architectures of ResNet50 and the proposed embedding extraction network

Layer name	Output size	ResNet50	Embedding extraction network
Conv1	112×112×64	Conv, 7×7, stride 2	Conv, 7×7, stride 2
Conv2_x	56×56×256	3×3 max pooling, stride 2 [Residual block] × 3	3×3 max pooling, stride 2 [MARL block] × 3
Conv3_x	28×28×512	Residual block, stride 2 [Residual block] × 3	MARL block, stride 2 [MARL block] × 3
Conv4_x	14×14×1024	Residual block, stride 2 [Residual block] × 5	MARL block, stride 2 [MARL block] × 5
Conv5_x	7×7×2048	Residual block, stride 2 [Residual block] × 2	MARL block, stride 2 [MARL block] × 2
GAP	1×1×2048	Global average pooling	Global average pooling
Output	2 or 256	Fully connected (output size 2)	Fully connected (output size 256)

Hybrid loss function

Traditional skin lesion classification approaches usually design neural networks to directly learn the mappings from input images to disease diagnoses, which cannot deal with the issues of insufficient training samples and inter-class similarities and intra-class variations very well. Different from these approaches, DeMAL-CNN first uses TPN to learn semantically meaningful embeddings of the input triplet, and then uses the learned embeddings to realize the final classification. To train DeMAL-CNN, we propose a hybrid loss function that combines the triplet loss with the classification loss:

$$L_{\text{hybrid}} = L_{\text{CE}} + \lambda L_{\text{TP}}. \quad (6)$$

where L_{CE} is the classification loss, L_{TP} is the triplet loss, and λ is a weighting factor that controls the contribution of L_{TP} . In this paper, λ is set to 1 in all our experiments.

The triplet loss measures the distances between the anchor sample and the positive sample, and between the anchor sample and the negative sample:

$$L_{\text{TP}} = \frac{1}{N} \sum_{i=1}^N [||\phi(x_a; \theta') - \phi(x_p; \theta')||_2^2 - ||\phi(x_a; \theta') - \phi(x_n; \theta')||_2^2 + \alpha]_+, \quad (7)$$

where x_a , x_p , and x_n represent the anchor, positive, and negative samples, respectively, $\phi(x_a; \theta')$, $\phi(x_p; \theta')$, and $\phi(x_n; \theta')$ are three learned triplet embeddings output from the embedding extraction network parameterized by θ' , $||\phi(x_a; \theta') - \phi(x_p; \theta')||_2^2$ and $||\phi(x_a; \theta') - \phi(x_n; \theta')||_2^2$ are the distance between the anchor sample and the positive sample and the distance between the anchor sample and the negative sample, respectively, α is a preset margin between these two distances (α is set to 1.0 in all our experiments), $[\cdot]_+$ is used to ensure that the loss not to be negative, and N is the number of valid triplets in a batch.

Regarding the selection of the valid triplets, we refer to the online triplet mining method in FaceNet [39] and remove

the easy triplets in all the triplets since they do not contribute to model training and may lower the average triplet loss of the batch. Specifically, the definition of the easy triplets is that the distance between x_a and x_n is not less than the sum of α and the distance between x_a and x_p .

In addition to the triplet loss, the classification loss of the cross entropy is added to improve the final classification performance and promote the convergence of the model training:

$$L_{\text{CE}} = -\frac{1}{3N} \sum_{i=1}^{3N} y_i \cdot \log(\phi(x_i; \theta)), \quad (8)$$

where x_i is a sample in the input triplet, y_i is its label, and $\phi(x_i; \theta)$ is the corresponding prediction output by DeMAL-CNN parameterized by θ . Note that θ represents the whole parameter set of DeMAL-CNN including both the embedding extraction network and the classification layer, while θ' in Eq. (7) only represents the parameter set of the embedding extraction network.

Experimental setup

Datasets

The performance of DeMAL-CNN was examined on two skin lesion classification challenge datasets (i.e., the ISIC 2016 and ISIC 2017 datasets) and a public benchmark dataset (i.e., the PH2 dataset). The first two datasets were collected and organized by the International Skin Imaging Collaboration (ISIC)¹. The ISIC 2016 dataset [42] was released by ISIC in 2016 for the challenge of classifying melanoma and non-melanoma skin lesions in dermoscopy images. This dataset contains 1279 dermoscopy images, of which 900 are used for training and 379 are used for testing. These images are 24-bit RGB images with the resolution varying from 679 × 556 to 4828 × 2848. The ISIC 2017 dataset [43] was released by ISIC in 2017 for the challenge of two independent binary

¹ <https://www.isic-archive.com>.

Table 2 Data distributions of the ISIC 2016, ISIC 2017, and PH2 datasets

Dataset	Disease	Training	Validation	Test
ISIC 2016	Melanoma	173	0	75
	Non-melanoma	727	0	304
ISIC 2017	Melanoma	374	30	117
	Seborrheic keratosis	254	42	90
	Nevus	1372	78	393
PH2 ^a	Melanoma	40		
	Nevus	160		

^aThe PH2 dataset does not officially separate the data for training, validation, and testing

classification tasks: melanoma vs. non-melanoma and seborrheic keratosis vs. non-seborrheic-keratosis. This dataset contains 2000, 150, and 600 dermoscopy images for training, validation, and test, respectively. These images are 24-bit RGB images with the resolution ranging from 679×453 to 6748×4499 . The PH2 dataset was collected at the Dermatology Service of Hospital Pedro Hispano², which contains 200 dermoscopy images (160 nevus and 40 melanomas). These images are 8-bit RGB images with a unified resolution of 768×560 . The data distributions of the ISIC 2016, ISIC 2017, and PH2 datasets are given in Table 2. It shows that these three datasets are unbalanced among different skin diseases, which makes the classification task more difficult. Moreover, these three datasets have been investigated by many participants in the challenges and other interested researchers. Therefore, it is convenient for us to compare the performance of different methods on these selected datasets.

Note that because the scale of the PH2 dataset is small (only 200 images in total), it is usually used for the evaluation of generalization performance of deep models. For example, Xie et al. [23] used the MB-CNN model trained on the ISIC 2017 dataset (with external 1320 training images from the ISIC archive) to directly test on the entire PH2 dataset to demonstrate its generalization performance. In this paper, we also used DeMAL-CNN trained on the ISIC 2017 dataset to evaluate the generalization performance on the PH2 dataset.

Algorithms for comparison

For the algorithms used for comparison on the ISIC 2016 and ISIC 2017 datasets, they can be simply classified into three categories: (1) the baseline and attention methods, (2) the top methods in the ISIC skin lesion classification challenges, and (3) the state-of-the-art skin lesion classification methods. Similar to [10], the original ResNet50 [40] was selected as the baseline. The attention methods include

RAN50 [33], SENet50 [34], and ARL-CNN50 [10], which are all attention-based variants of ResNet50. In addition, we selected the top six methods in the ISIC-2016 and ISIC-2017 skin lesion classification challenges, respectively. The performance of them was derived from [16] and [10]. For the selection of the top six methods in the challenges, we chose them according to the official leaderboards of the ISIC-2016 skin lesion classification challenge³ and the ISIC-2017 skin lesion classification challenge.⁴ These two leaderboards show the detailed rankings of the participants in the challenges. The state-of-the-art skin lesion classification methods used for comparison include SSAC [24], SDL [26], G-CNN [27], and GP-CNN [27]. We have introduced these four recent methods in Sect. Related Work, and the performance on the selected datasets can be available from their original publications.

The algorithms used for comparison on the PH2 dataset include the color constancy system (CCS) [44] and MB-CNN [23]. We selected these two algorithms because their performance has been tested on this dataset.

Implementation details

In the training phase, all training images of DeMAL-CNN were prepared with the form of the triplets, as described in Sect. Hybrid loss function and shown in Fig. 2. We first resized the images to the size of 224×224 and then employed the online data augmentation, including the random rotation ($[-40^\circ, +40^\circ]$), random horizontal flip, and random vertical flip, to prevent the overfitting in the model training. The mini-batch SGD algorithm was adopted as the optimizer, with the initial learning rate of 0.001, momentum of 0.9, weight decay of 0.0005, and the batch size of 16 for the ISIC 2016 dataset and 32 for the ISIC 2017 dataset, respectively. We loaded the pre-trained parameters on the ImageNet and reduced the learning rate by half every 30 epochs. The maximum number of training epochs was set to 100 and the validation set was used as the monitor to stop the training. Note that there is no validation set available in the ISIC 2016 dataset, so we randomly selected 30% of the training set for validation, and the remaining training set was used for training.

In the test phase, DeMAL-CNN took a test image as the input, and generated the embedding by the embedding extraction framework and the prediction result by the classification layer. The generated embeddings were used in our visualization experiments and the prediction results were used to calculate the classification performance. The source codes were released on Github⁵ for reproducing the experimental results reported in this paper.

³ <https://challenge.isic-archive.com/leaderboards/2016>.

⁴ <https://challenge.isic-archive.com/leaderboards/2017>.

⁵ <https://github.com/sugurs/DeMAL-CNN>.

² <https://www.fc.up.pt/addi/ph2%20database.html>.

Evaluation metrics

Five performance metrics were used to evaluate the model performance, including the accuracy (ACC), sensitivity, specificity, area under the receiver operating characteristic curve (AUC), and average precision (AP), which are defined below:

$$\text{ACC} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FN} + \text{TN} + \text{FP}}, \quad (9)$$

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}}, \quad (10)$$

$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}}, \quad (11)$$

$$\text{AUC} = \int_0^1 t(f) d(f), \quad (12)$$

$$\text{AP} = \int_0^1 p(r) d(r). \quad (13)$$

where TP, TN, FP and FN represent the number of true positive, true negative, false positive, and false negative samples, respectively, t and f are the true positive rate and false positive rate, respectively, p is the precision, and r is the recall. Among these performance metrics, ACC measures the overall classification accuracy of positive and negative samples, but it may not accurately characterize the actual performance of a model due to the imbalance of classes. Sensitivity and specificity can characterize the missed diagnosis rate and misdiagnosis rate of a model, respectively, but there is a trade-off between them. In some practical applications, if the sensitivity of a model increases, the specificity will be deteriorated to some extent [45,46]. AUC comprehensively considers the sensitivity and specificity of a model; thus, it is considered as the golden evaluation metric in the ISIC-2017 skin lesion classification challenge to rank all participants. In addition, AP can also measure the overall performance of the model, considering both the precision and recall of a model, so it is taken as the golden evaluation metric in the ISIC-2016 skin lesion classification challenge.

Results and discussions

Compared with the baseline and attention methods

First, we compared the performance of DeMAL-CNN with that of the baseline and attention methods on the ISIC 2016 dataset. The results are reported in Table 3. It indicates that both the three attention methods (RAN50, SENet50, and ARL-CNN50) and DeMAL-CNN perform significantly better than the baseline (ResNet50). Moreover, DeMAL-CNN achieves the highest AUC, ACC, sensitivity, and AP, and the second highest specificity among all the methods. Specif-

ically, compared with ResNet50, DeMAL-CNN improves the AUC, ACC, sensitivity, specificity, and AP by 5%, 2.2%, 18.7%, 3.3%, and 8.5%, respectively. Compared with ARL-CNN50, although DeMAL-CNN losses 0.4% in specificity, it improves the AUC, ACC, sensitivity, and AP by 3.4%, 1.2%, 12%, and 6%, respectively.

Then, we evaluated these methods on the ISIC 2017 dataset. The results in Table 4 show that DeMAL-CNN achieves the highest AUC, ACC, and sensitivity in both the melanoma classification and the seborrheic keratosis classification, though with slight performance degradation in terms of specificity. Specifically, in the melanoma classification, DeMAL-CNN achieves 1.6% AUC improvement, 1.8% ACC improvement, 2.5% sensitivity improvement, and 1% specificity improvement against ARL-CNN50, respectively. In the seborrheic keratosis classification, although DeMAL-CNN losses 2.0% in specificity against ARL-CNN50, it improves the AUC, ACC, and sensitivity by 1.6%, 1.9%, and 10.3%, respectively. In addition, in terms of the average AUC, DeMAL-CNN obtains 3.4% performance improvement against ResNet50, and 1.6% performance improvement against ARL-CNN50.

From the above experimental results, we can find that DeMAL-CNN performs better than its competitors in almost all performance indicators, except that the specificity is slightly lower than other methods. This phenomenon is caused by the trade-off between sensitivity and specificity. Since DeMAL-CNN achieves the highest sensitivity among all the methods in these experiments, it performs worse in specificity than other methods. AUC is an important performance indicator in medical applications, which considers both sensitivity and specificity [47]. Therefore, in terms of AUC, it can be concluded that the overall performance of DeMAL-CNN is better than that of the compared methods in these experiments.

Compared with the top algorithms in the challenges and state-of-the-art skin lesion classification methods

We also compared the performance of DeMAL-CNN with that of the top six algorithms in both the ISIC-2016 and ISIC-2017 skin lesion classification challenges and state-of-the-art skin lesion classification methods. The results are shown in Tables 5 and 6. In the ISIC-2017 skin lesion classification challenge, almost all methods (including the methods in [48–51,53]) use the external training images to enhance their classification performance. SDL [26] and SSAC [24] also adopt external training images. In addition, the ensemble learning strategy is adopted in [48] and [50,51,53] for an extra performance gain. Table 6 provides the number of the used external training images and whether using ensemble learning as a reference or not. Table 5 does not report such

Table 3 Results of DeMAL-CNN, the baseline (ResNet50), and the attention methods (RAN50, SENet50, and ARL-CNN50) on the ISIC 2016 dataset

Methods	ACC	Sensitivity	Specificity	AUC	AP
ResNet50 [40]	0.851	0.466	0.914	0.808	0.612
RAN50 [33]	0.860	0.480	0.917	0.811	0.628
SENet50 [34]	0.858	0.466	0.935	0.815	0.631
ARL-CNN50 [10]	0.861	0.533	0.951	0.824	0.637
DeMAL-CNN	0.873	0.653	0.947	0.858	0.697

For each evaluation metric, the highest performance is highlighted in boldface

Table 4 Results of DeMAL-CNN, the baseline (ResNet50), and the attention methods (RAN50, SENet50, and ARL-CNN50) on the ISIC 2017 dataset

Methods	Seborrheic keratosis classification				Melanoma classification				Average AUC
	ACC	Sensitivity	Specificity	AUC	ACC	Sensitivity	Specificity	AUC	
ResNet50 [40]	0.897	0.733	0.925	0.934	0.823	0.538	0.919	0.840	0.887
RAN50 [33]	0.910	0.722	0.943	0.944	0.838	0.564	0.909	0.848	0.896
SENet50 [34]	0.911	0.734	0.947	0.948	0.848	0.582	0.912	0.855	0.902
ARL-CNN50 [10]	0.908	0.778	0.931	0.951	0.837	0.590	0.896	0.859	0.905
DeMAL-CNN	0.927	0.881	0.911	0.967	0.855	0.615	0.906	0.875	0.921

For each evaluation metric, the highest performance is highlighted in boldface

Table 5 Results of DeMAL-CNN, the top six algorithms in the challenge, and the state-of-the-art skin lesion classification methods on the ISIC 2016 dataset

Methods	ACC	Sensitivity	Specificity	AUC	AP
#1 CUMED [16]	0.855	0.507	0.941	0.804	0.637
#2 GTDL	0.813	0.573	0.872	0.802	0.619
#3 BF-TB	0.834	0.320	0.961	0.826	0.598
#4 ThrunLab	0.786	0.667	0.816	0.796	0.563
#5 Jordan Yap	0.844	0.240	0.993	0.775	0.559
#6 Haebeom Lee	0.821	0.200	0.974	0.793	0.555
SDL [26]	0.863	–	–	0.822	0.681
G-CNN [27]	0.868	0.440	0.980	0.834	0.687
GP-CNN [27]	0.873	0.413	0.987	0.839	0.714
DeMAL-CNN	0.873	0.653	0.947	0.858	0.697

For each evaluation metric, the highest performance is highlighted in boldface. Note that AP is the golden evaluation metric to rank all the participants in this challenge

information because it cannot be available from [16] where we obtained the performance of the top six algorithms in the ISIC-2016 skin lesion classification challenge.

Table 5 shows that DeMAL-CNN achieves the highest AUC and ACC, and the second highest sensitivity and AP among all the methods. Compared with the top six challenge algorithms, DeMAL-CNN improves the best AUC (obtained by BF-TB) by 3.2% and the best AP (obtained by CUMED) by 6%. Compared with SDL that uses the synergic deep learning mechanism, DeMAL-CNN improves the AUC, ACC, and AP by 3.6%, 1%, and 1.6%, respectively. GP-CNN performs better than G-CNN because it integrates both the global and local information of dermoscopy images. Compared with GP-CNN, although DeMAL-CNN losses 4% in specificity and 1.7% in AP, respectively, it obtains 1.9% performance improvement in terms of AUC, and improves the

sensitivity by 24%. In the medical informatics community, sensitivity measures the missed diagnosis rate of a diagnostic model. Therefore, the low sensitivity of GP-CNN means a high missed diagnosis rate for melanoma detection, and the improved sensitivity of DeMAL-CNN means a decrease in missed diagnosis rate.

From Table 6, DeMAL-CNN achieves the highest AUC and the second highest sensitivity in the melanoma classification, and the highest AUC and ACC in the seborrheic keratosis classification. Moreover, DeMAL-CNN achieves the best average AUC among all the methods. Specifically, compared with the method in [48] which ranks the first in the ISIC-2017 skin lesion classification challenge, DeMAL-CNN improves the AUC, ACC, specificity by 0.7%, 2.7%, and 5.5% in the melanoma classification, respectively, and improves the AUC, ACC, and specificity by 1.4%, 12.4%,

Table 6 Results of DeMAL-CNN, the top six algorithms in the challenge, and the state-of-the-art skin lesion classification methods on the ISIC 2017 dataset

Methods	External data	With ensemble learning	Seborrheic keratosis classification				Melanoma classification				Average AUC
			ACC	Sensitivity	Specificity	AUC	ACC	Sensitivity	Specificity	AUC	
#1 [48]	1444	Yes	0.803	0.978	0.773	0.953	0.828	0.735	0.851	0.868	0.911
#2 [49]	900	No	0.875	0.178	0.998	0.965	0.823	0.103	0.998	0.856	0.910
#3 [50]	7544	Yes	0.895	0.356	0.990	0.943	0.872	0.547	0.950	0.874	0.909
#4 [51]	1600	Yes	0.918	0.589	0.976	0.921	0.858	0.427	0.963	0.870	0.896
#5 [52]	1341	Yes	0.913	0.556	0.976	0.935	0.845	0.350	0.965	0.836	0.886
#6 [53]	0	No	0.917	0.700	0.995	0.942	0.830	0.436	0.925	0.830	0.886
SDL [26]	1320	No	0.925	–	–	0.958	0.888	–	–	0.868	0.913
SSAC [24]	1320	No	0.912	0.889	0.916	0.959	0.835	0.556	0.903	0.873	0.916
G-CNN [27]	0	No	0.920	0.760	0.949	0.953	0.843	0.453	0.938	0.865	0.909
GP-CNN [27]	0	No	0.925	0.856	0.933	0.964	0.857	0.462	0.952	0.870	0.917
DeMAL-CNN	0	No	0.927	0.881	0.911	0.967	0.855	0.615	0.906	0.875	0.921

For each evaluation metric, the highest performance is highlighted in boldface. Note that the average AUC is the golden evaluation metric to rank all the participants in this challenge

and 13.8% in the seborrheic keratosis classification, respectively. Compared with SDL and SSAC that use 1320 external training images, DeMAL-CNN trained with only official training images still achieves higher AUC and sensitivity in the melanoma classification, and higher AUC and ACC in the seborrheic keratosis classification. Compared with GP-CNN, which also does not use external data and ensemble learning, DeMAL-CNN obtains 0.5% AUC improvement and 15.3% sensitivity improvement in the melanoma classification, and obtains 0.3% AUC improvement and 2.5% sensitivity improvement in the seborrheic keratosis classification. Note that the training of GP-CNN involves the training of G-CNN and P-CNN, and the training of P-CNN applies the CAM-guided probabilistic cropping strategy [27] to obtain 20 image patches for each image, which causes the increase of computational cost and training time. In this paper, DeMAL-CNN relies on the ability of the proposed mixed attention mechanism to make the network adaptively focus on discriminative skin lesion parts, thus saving computational cost and training time compared with GP-CNN.

Generalization performance

In the machine learning community, generalization performance measures how the model performs on predicting unseen data. For disease diagnosis, a model with good generalization performance has the potential to generate accurate diagnosis results in practical applications. To evaluate the generalization performance of DeMAL-CNN, we trained DeMAL-CNN on the melanoma classification task of the ISIC 2017 dataset, and directly tested it on the PH2 dataset. The results in Table 7 suggest that DeMAL-CNN improves AUC by 10% compared with CCS trained and tested on this dataset, demonstrating that DeMAL-CNN has a strong model generalization ability. Compared with MB-CNN trained on the ISIC 2017 dataset and additional 1320 dermoscopy images from the ISIC archive, DeMAL-CNN losses the AUC, ACC, and specificity by 1.3%, 1.5%, and 1.2%, respectively. As described in Sect. Related Work, MB-CNN is a multitask network, in which the segmentation network is trained to provide the lesion masks for the classification network. The lesion mask provided by the segmentation network is a strong prior about the location of the skin lesion in the dermoscopy image. Therefore, it is not surprising that MB-CNN achieves better generalization performance than DeMAL-CNN. We also fine-tuned DeMAL-CNN on the PH2 dataset according to the fine-tuning method in [23]. The results in Table 7 show that the fine-tuning operation improves the performance of both DeMAL-CNN and MB-CNN, and that the performance of the fine-tuned DeMAL-CNN is comparable to that of the fine-tuned MB-CNN. The above results suggest that DeMAL-CNN can obtain further performance improvement after it is fine-tuned.

Visualization of the learned embeddings and CAMs

DeMAL-CNN leverages both deep metric learning and mixed attention learning for skin lesion classification. The deep metric learning helps to deal with the inter-class similarities and intra-class variations and the mixed attention learning enables the model to concentrate on skin lesion parts. To validate this, we implemented two visualizations of the original ResNet50 and DeMAL-CNN: one is the visualization of embeddings and the other is the visualization of class activation mappings (CAMs).

To visualize the embeddings, we implemented t-SNE [54], which is a widely used data visualization method and can reduce the dimension of the feature vector to two or three. In this paper, the visualized embedding of ResNet50 is the output of its last GAP layer and that of DeMAL-CNN is extracted by the embedding extraction network. The dimensions of the visualized embedding in ResNet50 and DeMAL-CNN are 2048 and 256, respectively. Figure 4 depicts the visualization results of ResNet50 and DeMAL-CNN in the ISIC-2016 melanoma classification, ISIC-2017 melanoma classification, and ISIC-2017 seborrheic keratosis classification. The visualization results show the ability of ResNet50 and DeMAL-CNN to distinguish between positive and negative samples, where a yellow/purple dot in each graph is used to represent a positive/negative sample. Compared with ResNet50, DeMAL-CNN has a stronger discriminative representation ability to pull samples with different classes farther away from each other and push samples with the same class closer to each other.

The visualization of CAMs can highlight important regions in an image for predicting the concept. To generate CAMs of CNN models, we employed the gradient weighted CAM (Grad-CAM) [55], which finds the final convolutional layer in the network and then examines the gradient information flowing into that layer. Figure 5 shows the dermoscopy images and their corresponding CAMs of ResNet50 and DeMAL-CNN. It suggests that both ResNet50 and DeMAL-CNN can learn the attention regions that reflect the location of skin lesions, but they have different highlighted positions and concentrations. Compared with ResNet50, DeMAL-CNN has a more powerful attention ability to highlight the skin lesion area rather than the surrounding normal tissues, especially in the dermoscopy images presented in the 6th, 7th, and 8th columns. Therefore, DeMAL-CNN has better skin lesion classification performance than ResNet50.

Ablation study

DeMAL-CNN includes two main components: the TPN structure and the MARL blocks. To investigate their effectiveness, we conducted the ablation experiments on each of them. Tables 8 and 9 give the results of the ISIC 2016

Table 7 Results of DeMAL-CNN, CCS, and MB-CNN on the PH2 dataset

Methods	ACC	Sensitivity	Specificity	AUC
CCS [44]	—	0.925	0.763	0.843
MB-CNN [23]	0.885	0.825	0.900	0.956
DeMAL-CNN	0.868	0.763	0.887	0.928
MB-CNN [23] (fine-tuned)	0.940	0.950	0.938	0.977
DeMAL-CNN (fine-tuned)	0.935	0.950	0.931	0.974

For each evaluation metric, the highest performance is highlighted in boldface

Fig. 4 Visualizing the embeddings of ResNet50 (the first row) and DeMAL-CNN (the second row) **Fig. 5** Visualizing the CAMs of ResNet50 and DeMAL-CNN. The three images in each column are the dermoscopy image, the CAM of ResNet50, and the CAM of DeMAL-CNN from top to bottom

and ISIC 2017 datasets, respectively. In these experiments, DeMAL-CNN without the TPN structure represents the variant obtained by replacing the residual blocks in ResNet50 with the MARL blocks, and DeMAL-CNN without the MARL blocks represents the variant obtained by replacing the MARL blocks in DeMAL-CNN with the residual blocks.

(1) *Effectiveness of the TPN structure*: To valid the effectiveness of the TPN structure, we compared the results of DeMAL-CNN and DeMAL-CNN without the TPN structure. From Table 8, the usage of the TPN structure improves the AP by 3.2%. In addition, as shown in Table 9, the employment of the TPN structure improves the AUC by 1% and 0.9% in the melanoma classification and the seborrheic keratosis classification, respectively.

(2) *Effectiveness of the MARL blocks*: We also compared the results of DeMAL-CNN and DeMAL-CNN without the MARL blocks. From Table 8, the usage of the MARL blocks improves the AP by 1%. In addition, as shown Table 9, the employment of the MARL blocks improves the AUC by 0.7% and 0.4% in the melanoma classification and the seborrheic keratosis classification, respectively.

In summary, the above results demonstrate the effectiveness of both the TPN structure and the MARL blocks in DeMAL-CNN.

Dimension of the embedding layer

In the above experiments, the dimension of the embedding layer in DeMAL-CNN was fixed to 256. The dimension of the embedding layer can take different values. To evaluate the effect of the dimension of the embedding layer, we compared the performance of DeMAL-CNN with different embedding dimensions, including 1024, 512, 256, 128, and 64.

Figure 6 summarizes the AP for the ISIC-2016 melanoma classification and the AUC for the ISIC-2017 melanoma and seborrheic keratosis classifications. The results show that DeMAL-CNN with the embedding dimension of 1024 achieves the best classification performance in all the three classifications. As the embedding dimension becomes smaller, the performance of DeMAL-CNN becomes worse. For example, when the embedding dimension is reduced to 128 or 64, the performance of DeMAL-CNN is seriously degraded in all the three classifications, which can be attributed to the fact that the extracted low-dimensional embedding cannot represent the information of the input data well. The performance degradation of DeMAL-CNN with the embedding dimension of 256 is negligible compared with the best performance in all the three classifications. Note that DeMAL-CNN with the embedding dimension of 256 has a lower computational complexity and less training time, compared with DeMAL-CNN with a larger embedding dimension. Therefore, we chose the embedding dimension of 256 in DeMAL-CNN.

Attention ability to the artifacts

The classification of skin lesions may be affected by the presence of artifacts, such as hairs, rulers, or colored circles. Figure 7 shows some dermoscopy images surrounded or covered by various artifacts. To investigate the effectiveness of DeMAL-CNN on these artifacts, the corresponding CAMs of DeMAL-CNN were visualized in Fig. 7. The visualization results show that DeMAL-CNN can still accurately locate the skin lesion regions instead of hairs, rulers, colored circles, or other artifacts. Considering that the skin lesion classification in the clinical environment also often suffer from these artifacts, the attention ability of DeMAL-CNN will be highly adaptable.

Robustness analysis

To analyze the robustness of DeMAL-CNN, we conducted ten independent runs for DeMAL-CNN and ResNet50 on the ISIC-2016 melanoma classification and the ISIC-2017 melanoma and seborrheic keratosis classifications. The results of these three classifications are shown in Fig. 8. To test the statistical significance between DeMAL-CNN and ResNet50, the independent two-sample t -test at a 0.01 significance level was carried out for each classification task:

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2-2} \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \quad (14)$$

where $\bar{X}_1$ and $\bar{X}_2$ are the averages of two groups of samples, n_1 and n_2 are the numbers of two groups of samples, and S_1 and S_2 are the standard deviations of two groups of samples.

According to Eq. (14) and the ten independent experimental results shown in Fig. 8, the values of t_1 , t_2 , and t_3 of the three classifications are -208.21 , -78.72 , and -88.28 , respectively. Given that the significance level is 0.01, we get a rejection domain $W = \{t \leq -2.552\}$. Because all t_1 , t_2 , and t_3 belongs to the rejection domain, DeMAL-CNN performs significantly better than ResNet50 in all the three classifications.

Computational complexity

Our experiments were conducted on an NVIDIA GTX Titan XP GPU. The training of DeMAL-CNN took about 10 hours on the ISIC 2016 dataset and about 20 hours on the ISIC 2017 dataset. The training time of DeMAL-CNN is longer than that of ResNet50 because of the calculation of the triplet loss, while is shorter than the approach in [10] which adopts a random cropping method to increase the training samples by 60 times. For example, the training time of [10] is about 30 hours on the ISIC 2017 dataset. Besides, DeMAL-CNN

Table 8 Ablation study for the TPN structure and the MARL blocks on the ISIC 2016 dataset

Methods	ACC	Sensitivity	Specificity	AUC	AP
DeMAL-CNN without the TPN structure	0.871	0.560	0.931	0.828	0.665
DeMAL-CNN without the MARL blocks	0.863	0.573	0.941	0.853	0.687
DeMAL-CNN	0.873	0.653	0.947	0.858	0.697

For each evaluation metric, the highest performance is highlighted in boldface

Table 9 Ablation study for the TPN structure and the MARL blocks on the ISIC 2017 dataset

Methods	Seborrheic Keratosis Classification				Melanoma Classification				Average AUC
	ACC	Sensitivity	Specificity	AUC	ACC	Sensitivity	Specificity	AUC	
DeMAL-CNN without the TPN structure	0.907	0.844	0.918	0.958	0.852	0.614	0.885	0.865	0.912
DeMAL-CNN without the MARL blocks	0.923	0.822	0.951	0.963	0.838	0.599	0.896	0.868	0.916
DeMAL-CNN	0.927	0.881	0.911	0.967	0.855	0.615	0.906	0.875	0.921

For each evaluation metric, the highest performance is highlighted in boldface

Fig. 6 Performance of DeMAL-CNN with different embedding dimensions: **a** the results of the ISIC-2016 melanoma classification and **b** the results of the ISIC-2017 melanoma and seborrheic keratosis classifications **Fig. 7** Visualizing the dermoscopy images with the presence of artifacts and the corresponding CAMs obtained by DeMAL-CNN. The types of artifacts include: **a** hairs, **b** rulers, **c** colored circles, and **d** others

Fig. 8 10 independent test results of DeMAL-CNN and ResNet50: **a** the results of the ISIC-2016 melanoma classification, **b** the results of the ISIC-2017 melanoma classification, and **c** the results of the ISIC-2017 seborrheic keratosis classification

has a fast inference speed, which takes about 0.3 s per image on average. Therefore, DeMAL-CNN is highly adaptable to the clinical environments.

Conclusion

In this paper, we proposed DeMAL-CNN for skin lesion classification in dermoscopy images. In DeMAL-CNN, TPN was first designed based on deep metric learning, which consisted of three weight-shared embedding extraction networks. TPN adopted a triplet of samples as input and used the triplet loss to optimize the embeddings, which can not only increase the number of training samples, but also learn the embeddings robust to inter-class similarities and intra-class variations. In addition, a mixed attention mechanism considering both the spatial-wise and channel-wise attention information was designed and integrated into the construction of each embedding extraction network, which can further strengthen the skin lesion localization ability of DeMAL-CNN. We evaluated the performance of DeMAL-CNN on the ISIC 2016, ISIC 2017, and PH2 datasets. The results suggested that DeMAL-CNN achieved the state-of-the-art performance on the ISIC 2016 and ISIC 2017 datasets, and showed good generalization performance on the PH2 dataset. Moreover, the results of the ablation study verified the effectiveness of both the TPN structure and the MARL blocks in DeMAL-CNN. In the future, we plan to establish an end-to-end deep learning system to simultaneously segment and classify skin lesions, in which the performance of skin lesion classification can be

further enhanced through the information learned from the segmentation task.

Acknowledgements This work was supported in part by the National Natural Science Foundation of China under Grants 61976225 and 62041208, in part by the Natural Science Foundation of Changsha under Grant Kq2014294, and in part by the Foundational Research Funds for the Central Universities of Central South University under Grant 2020zzts132.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

- Hay Roderick J, Johns Nicole E, Williams Hywel C, Bolliger Ian W, Dellavalle Robert P, Margolis David J, Marks Robin, Naldi Luigi, Weinstock Martin A, Wulf Sarah K, The global burden of skin disease in, et al (2010) An analysis of the prevalence and impact of skin conditions. *J Invest Dermatol* 134(6):527–1534
- Guy Jr Gery P, Machlin Steven R, Ekwueme Donatus U, Robin YK (2015) Prevalence and costs of skin cancer treatment in the US, 2002– 2006 and 2007– 2011. *Am J Prevent Med* 48(2):183–187
- Stern RS (2010) Prevalence of a history of skin cancer in 2007: results of an incidence-based model. *Arch Dermatol* 146(3):279–282
- Robinson JK (2005) Sun exposure, sun protection, and vitamin D. *JAMA* 294(12):1541–1543
- Binder M, Puespoek-Schwarz M, Steiner A, Kittler H, Mueller M, Wolff K, Pehamberger H et al (1997) Epiluminescence microscopy of small pigmented skin lesions: short-term formal training improves the diagnostic performance of dermatologists. *J Am Acad Dermatol* 36(2):197–202
- Ganster H, Pinz P, Rohrer R, Wildling E, Binder M, Kittler H (2001) Automated melanoma recognition. *IEEE Trans Med Imaging* 20(3):233–239
- Barata C, Ruela M, Francisco M, Mendonça T, Marques JS (2013) Two systems for the detection of melanomas in dermoscopy images using texture and color features. *IEEE Syst J* 8(3):965–979
- Xie F, Fan H, Li Y, Jiang Z, Meng R, Bovik A (2016) Melanoma classification on dermoscopy images using a neural network ensemble model. *IEEE Trans Med Imaging* 36(3):849–858
- Esteva A, Kuprel B, Novoa RA, Ko J, Swetter SM, Blau HM, Thrun S (2017) Dermatologist-level classification of skin cancer with deep neural networks. *Nature* 542(7639):115–118
- Zhang J, Xie Y, Xia Y, Shen C (2019) Attention residual learning for skin lesion classification. *IEEE Trans Med Imaging* 38(9):2092–2103
- Gessert N, Sentker T, Madesta F, Schmitz R, Kniep H, Baltruschat I, Werner R, Schlaefer A (2019) Skin lesion classification using CNNs with patch-based attention and diagnosis-guided loss weighting. *IEEE Trans Biomed Eng* 67(2):495–503
- Li T, Zhang Y, Wang T (2021) SRPM-CNN: a combined model based on slide relative position matrix and CNN for time series classification. *Complex Intell Syst* 7(3):1619–1631
- Haenssle HA, Fink C, Schneiderbauer R, Toberer F, Buhl T, Blum A, Kalloo A, Hassen ABH, Thomas L, Enk A et al (2018) Man against machine: diagnostic performance of a deep learning convolutional neural network for dermoscopic melanoma recognition in comparison to 58 dermatologists. *Ann Oncol* 29(8):1836–1842
- Christian S, Vincent V, Sergey I, Jon S, Zbigniew W (2016) Rethinking the inception architecture for computer vision. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 2818–2826
- Christian S, Sergey I, Vincent V, Alexander AA (2017) Inception-v4, Inception-Resnet and the impact of residual connections on learning. In: Thirty-first AAAI conference on artificial intelligence
- Lequan Yu, Chen H, Dou Q, Qin J, Heng P-A (2016) Automated melanoma recognition in dermoscopy images via very deep residual networks. *IEEE Trans Med Imaging* 36(4):994–1004
- Yang S, Weidong C, Heng H, Yun Z, David Dagan F, Yue W, Fulham Michael J, Mei C (2015) Large margin local estimate with applications to medical image classification. *IEEE Trans Med Imaging* 34(6):1362–1377
- Hugo L, Dumitru E, Aaron C, James B, Yoshua B (2007) An empirical evaluation of deep architectures on problems with many factors of variation. In: Proceedings of the 24th international conference on machine learning, pp 473–480
- Han X, Kashif R, Roland V (2017) Fashion-Mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*
- Wang F, Sun J (2015) Survey on distance metric learning and dimensionality reduction in data mining. *Data Min Knowl Discov* 29(2):534–564
- Jiwen L, Junlin H, Zhou J (2017) Deep metric learning for visual understanding: an overview of recent advances. *IEEE Signal Process Mag* 34(6):76–84
- Yang J, Xie F, Fan H, Jiang Z, Liu J (2018) Classification for dermoscopy images using convolutional neural networks based on region average pooling. *IEEE Access* 6:65130–65138
- Xie Y, Zhang J, Xia Y, Shen C (2020) A mutual bootstrapping model for automated skin lesion segmentation and classification. *IEEE Trans Med Imaging* 39(7):2482–2493
- Xie Y, Zhang J, Xia Y (2019) Semi-supervised adversarial model for benign-malignant lung nodule classification on chest CT. *Med Image Anal* 57:237–248
- Qin Z, Liu Z, Zhu P, Xue Y (2020) A GAN-based image synthesis method for skin lesion classification. *Comput Methods Programs Biomed* 195:105568
- Zhang J, Xie Y, Qi W, Xia Y (2019) Medical image classification using synergic deep learning. *Med Image Anal* 54:10–19
- Tang P, Liang Q, Yan X, Xiang S, Zhang D (2020) GP-CNN-DTEL: global-part CNN model with data-transformed ensemble learning for skin lesion classification. *IEEE J Biomed Health Inform* 24(10):2870–2882
- Yequan W, Minlie H, Xiaoyan Z, Li Z (2016) Attention-based LSTM for aspect-level sentiment classification. In: Proceedings of the 2016 conference on empirical methods in natural language processing, pp 606–615
- Huimin C, Maosong S, Cunchao T, Yankai L, Zhiyuan L (2016) Neural sentiment classification with user and product attention. In: Proceedings of the 2016 conference on empirical methods in natural language processing, pp 1650–1659
- Xia H, Luo Y, Liu Y (2021) Attention neural collaboration filtering based on GRU for recommender systems. *Complex Intell Syst* 7(3):1367–1379
- Fang L, Wang C, Li S, Rabbani H, Chen X, Liu Z (2019) Attention to lesion: lesion-aware convolutional neural network for retinal optical coherence tomography image classification. *IEEE Trans Med Imaging* 38(8):1959–1970
- Gang X, Shifeng L, Yicao M (2020) A hybrid deep learning-based fruit classification using attention model and convolution autoencoder. *Complex Intell Syst* 1–11
- Fei W, Mengqing J, Chen Q, Shuo Y, Cheng L, Honggang Z, Xiaogang W, Xiaoou T (2017) Residual attention network for image classification. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 3156–3164
- Jie H, Li S, Gang S (2018) Squeeze-and-excitation networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 7132–7141
- Peng Y, He X, Zhao J (2017) Object-part attention model for fine-grained image classification. *IEEE Trans Image Process* 27(3):1487–1500
- Elad H, Nir A (2015) Deep metric learning using triplet network. In: International workshop on similarity-based pattern recognition, pp 84–92. Springer, New York
- Xinwei H, Yang Z, Zhichao Z, Song B, Xiang B (2018) Triplet-center loss for multi-view 3D object retrieval. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1945–1954

38. IsaaK L, Anne G, Leif K (2016) Identifying style of 3D shapes using deep metric learning. In: Computer graphics forum, vol 35, pp 207–215. Wiley Online Library, New York
39. Florian S, Dmitry K, James P (2015) FaceNet: a unified embedding for face recognition and clustering. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 815–823
40. Kaiming H, Xiangyu Z, Shaoqing R, Jian S (2016) Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 770–778
41. Kyunghyun C, Merriënboer Bart V, Caglar G, Dzmitry B, Fethi B, Holger S, Yoshua B (2014) Learning phrase representations using RNN encoder-decoder for statistical machine translation. arXiv preprint [arXiv:1406.1078](https://arxiv.org/abs/1406.1078)
42. David G, Codella Noel CF, Emre C, Brian H, Michael M, Nabin M, Allan H (2016) 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](https://arxiv.org/abs/1605.01397)
43. Codella Noel CF, David G, Emre Celebi M, Brian H, Marchetti Michael A, Dusza Stephen W, Aadi K, Konstantinos L, Nabin M, Harald K et al (2018) 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). In: 2018 IEEE 15th international symposium on biomedical imaging (ISBI 2018), pp 168–172. IEEE, New York
44. Barata C, Celebi ME, Marques JS (2014) Improving dermoscopy image classification using color constancy. IEEE J Biomed Health Inform 19(3):1146–1152
45. Jessica C, Gaia P, Weiss Noel S (2012) Tradeoffs between accuracy measures for electronic health care data algorithms. J Clin Epidemiol 65(3):343–349
46. Wenzhe S, Jan-Dirk S, Toshiyuki N (2020) On the tradeoff between sensitivity and specificity in bus bunching prediction. J Intell Transport Syst 1–17
47. Kumar R, Indrayan A (2011) Receiver operating characteristic (ROC) curve for medical researchers. Indian Pediatr 48(4):277–287
48. Kazuhisa M, Akira H, Akane M, Hiroshi K (2017) Image classification of melanoma, nevus and seborrheic keratosis by deep neural network ensemble. arXiv preprint [arXiv:1703.03108](https://arxiv.org/abs/1703.03108)
49. González DI (2017). Incorporating the knowledge of dermatologists to convolutional neural networks for the diagnosis of skin lesions. arXiv preprint [arXiv:1703.01976](https://arxiv.org/abs/1703.01976)
50. Afonso M, Julia T, Michel F, Lin LT, Sandra A, Eduardo V (2017) RECOD titans at ISIC challenge 2017. arXiv preprint [arXiv:1703.04819](https://arxiv.org/abs/1703.04819)
51. Lei B, Jinman K, Euijoon A, Dagan F (2017) Automatic skin lesion analysis using large-scale dermoscopy images and deep residual networks. arXiv preprint [arXiv:1703.04197](https://arxiv.org/abs/1703.04197)
52. Xulei Y, Zeng Z, Yong YS, Colin T, Hong Liang T, Yi S (2017) A novel multi-task deep learning model for skin lesion segmentation and classification. arXiv preprint [arXiv:1703.01025](https://arxiv.org/abs/1703.01025)
53. Terrance D, Dhanesh R (2017) Skin lesion classification using deep multi-scale convolutional neural networks. arXiv preprint [arXiv:1703.01402](https://arxiv.org/abs/1703.01402)
54. Van der Maaten L, Hinton G (2008) Visualizing data using t-SNE. J Mach Learn Res 9(2605):2579–2605
55. Selvaraju Ramprasaath R, Michael C, Abhishek D, Ramakrishna V, Devi P, Dhruv B (2017) Grad-CAM: visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE international conference on computer vision, pp 618–626

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.