An Attention-based Convolutional Neural Network for Melanoma Recognition

Qi Chen¹, Lidan Wang¹,3,4,5, Xiuling Gan¹, Shukai Duan²,3,4,5*

¹Southwest University, College of Electronic & Information Engineering, Chongqing, 400715, China
²Southwest University, College of Artificial Intelligence, Chongqing, 400715, China
³Brain-inspired Computing & Intelligent Control of Chongqing Key Lab, Chongqing, 400715, China
⁴National & Local Joint Engineering Laboratory of Intelligent Transmission and Control Technology, Chongqing, 400715, China
⁵Chongqing Brain Science Collaborative Innovation Center, Chongqing, 400715, China

* Corresponding author’s e-mail: duansk@swu.edu.cn

Abstract: Early automatic and accurate melanoma recognition is an important method to reduce melanoma deaths. Existing methods are less sensitive to the position of the lesion areas. Network training may be affected by the uncorrelated noisy parts. In light of this circumstance, an end-to-end attention-based network AF-CNN for accurate melanoma recognition is proposed in this paper, which is mainly composed of pre-trained VGG19, attention blocks and a classification layer. Instead of treating each part of the input dermoscopy images equally, our AF-CNN model has strong discriminative ability to focus on the lesion areas. The AF-CNN was evaluated on the ISIC2017 dataset and concluded the proposed single model achieves the state-of-the-art result in melanoma recognition task.

1. Introduction

Melanoma is a very dangerous skin cancer, 70% of skin cancer deaths are caused by melanoma[1] and its morbidity and mortality are rising[2]. Fortunately, early diagnosis and treatment of melanoma is an effective way to improve its survival rate. Clinically, a dermatologist diagnoses by analyzing dermoscopy images, which has certain subjectivity and requires high professional skills resulting in limited diagnosis speed and accuracy. Automatic melanoma recognition algorithm can effectively assist doctors in diagnosis.

Many automatic skin cancer classification methods have been proposed. Among them, deep learning algorithms have obtained relatively good results. However, they are not sensitive to the lesion area in the dermoscopy images, so network training is affected by irrelevant areas. In order to enable the model to be able to focus on the diseased area like a doctor, an end-to-end attention-based AF-CNN is proposed for melanoma recognition in this paper. The contributions of the paper are as follows: (1) we propose an attention-based AF-CNN that can adaptively focus on the lesion area in the dermoscopy images which has strong discriminative ability. Our method adopts neither a complicated time-consuming ensemble method nor external data; (2) in the classification layer, we use multi-scale fusion of high-level and low-level attention features to enhance the features’ richness. Therefore, the network can make...
a more accurate classification according to the learned information; and (3) the proposed model achieves an AUC of 0.885 on the ISIC2017 test set, exceeding the current highest AUC of 0.875.

2. Proposed AF-CNN

2.1. Overall network structure
The overall network structure of AF-CNN is shown in Figure 1. In order to prevent overfitting and improve the generalization ability of our model, we utilize the pre-trained VGG19 network with all fully connected layers removed as the backbone network in our model. Attention blocks are respectively added after the third, fourth, and fifth pooling layers of the pre-trained VGG19 network. Finally, the outputs of the three attention blocks are concatenated together to form the final feature vector.

![Figure 1. The overall AF-CNN architecture.](image)

2.2. Attention Block
As some examples showed in Figure 4, a large part of the dermoscopy images are task-independent background areas. In order to allow the network to focus on the lesion areas related to classification task, three attention blocks are added to the backbone network VGG19, and this model is named after A-CNN. Inspired by Woo et al., the CBAM module is used as the attention block in our proposed model[3]. But we adopt the spatial-channel cascade method, that is, the feature is weighted in the spatial dimension first and then weighted in the channel dimension.

![Figure 2. The internal structure of attention block](image)

The internal structure of the attention block is shown in Figure 2. The input of attention block is the intermediate feature $F$ ($C \times H \times W$) of VGG19. Firstly, the spatial module in attention block generates a spatial attention map $M_S(1 \times H \times W)$, multiply $M_S$ with the intermediate feature $F$ to get the spatial-refined feature $F'$. Next, the channel module in attention block generates a channel attention map $M_C(C \times 1 \times 1)$, multiply $M_C$ with $F'$ to get the final refined feature $F''$. In summary, the calculation process in attention block can be described as:

$$F' = M_S(F) \odot F; F'' = M_C \odot F'$$

(1)
Among them, $\otimes$ means element-wise multiplication. Each value in the spatial (channel) attention map represents the degree of attention paid to the corresponding feature vector in the feature. After attention block, the refined feature $F''$ is as the same size as the intermediate feature $F$.

Figure 3 (a) and (b) show the internal details of the spatial and channel module, respectively. The spatial module generates spatial attention map by aggregating information in the channel direction, including average-pool and max-pool. And the two pooled features are concatenated together, and then a convolution operation is performed to obtain the spatial attention map. Similar to the spatial module, the channel module performs average-pooling and max-pooling operations along the spatial direction. After passing through a multi-layer perceptron (MLP), the two one-dimensional features are merged by element-wise summation to produce the channel attention map.

![Spatial Attention Module](image1)

![Channel Attention Module](image2)

**Figure 3.** (a) spatial module and (b) channel module

### 2.3. Fusing lower attention features

The features extracted from different convolutional layers in a convolutional neural network are different. Low-level features have higher resolution and more local detail information, but have less semantic information. And high-level features have richer semantic information.

In our proposed method, we fuse high-level features with low-level features, effectively using the complementarity of different level features to improve last feature vector’s richness and discriminative ability. Particularly, we fuse the weighted features here. The outputs of the three attention blocks in the network shown in Figure 1 are all pass through the global average pooling layer (GAP)[4]. And then concatenate the pooled three feature vectors to generate the final feature vector $g$. Feature vector $g$ is the input of the final classification layer. The feature fusion process can be described as:

$$g = \text{concat}[\text{GAP}(F''_1) + \text{GAP}(F''_2) + \text{GAP}(F''_3)]$$

### 3. Experiments and results

#### 3.1. Dataset

The proposed model was evaluated on the International Skin Imaging Collaboration 2017 (ISIC2017) skin lesion classification dataset[5], which has 2000 training images, 150 validation images and 600 test images and contains 3 categories: melanoma, nevus, and seborrheic keratosis. And Figure 4 shows some samples of dermoscopy images. As melanoma is extremely harmful, we only identify melanoma.
3.2. Implementation details
We use a system with a graphics card of type GTX 1080Ti and a deep learning framework PyTorch[6] to implement our model. An online data augmentation method was used to expand the training dataset including random rotation, horizon and vertical flips. We adopt SGD algorithm with momentum 0.9 to train our model for total 50 epochs. The initial learning rate is 0.01 and the learning rate is adjusted according to epoch via multiplying by 0.1 every 10 epochs. Besides, in order to solve the problem of imbalanced data we use data up-sampling and use focal loss[7] as the loss function.

3.3. Results
3.3.1. The effectiveness of attention mechanism
We verify the effectiveness of the attention mechanism from both quantitative and qualitative perspectives. Firstly, we quantitatively illustrate the effectiveness of the attention mechanism. Experiments are conducted to demonstrate the effectiveness of A-CNN model compared with two baseline models, VGG19 and VGG19-GAP. The architectures of the two baseline models are shown in Table 1. Add attention blocks in VGG19-GAP model to form the A-CNN model (feature fusion is not included here). From the first three rows of Table 2, the A-CNN model performs better than the two baseline models.

Then, we illustrate the effectiveness of attention mechanism from a qualitative perspective. From the first three rows of Table 2, the A-CNN model performs better than the two basic models. In order to show that the performance improvement is related to the better attention, we visualize class activation maps (CAMs)[8] obtained by VGG19-GAP and A-CNN. The visualization diagrams of the CAMs obtained from the VGG19-GAP model and the A-CNN model are respectively located in the middle row and the bottom row in Figure 5. As shown in Figure 5, the CAMs obtained from A-CNN model highlights the more discriminative regions in dermoscopy images, which means that the A-CNN model can focus on the lesion area. Therefore, this may explain why A-CNN model has better recognition results than VGG19-GAP.
Table 1. The architecture of two baseline models. ×2, ×4 indicate the number of convolution layers

| VGG19       | VGG19-GAP      |
|-------------|----------------|
| input(224×224) | input(224×224) |
| (conv3-64)×2 | (conv3-64)×2   |
| max-pool     | max-pool       |
| (conv3-128)×2| (conv3-128)×2  |
| max-pool     | max-pool       |
| (conv3-256)×4| (conv3-256)×4  |
| max-pool     | max-pool       |
| (conv3-512)×4| (conv3-512)×4  |
| max-pool     | max-pool       |
| (conv3-512)×4| (conv3-512)×4  |
| max-pool     | max-pool       |
| FC-4096      | GAP            |
| FC-4096      | FC-2           |
| FC-2         |                |

3.3.2. The effectiveness of fusion of attention features

We compare AF-CNN (shown in Figure.1) with A-CNN to prove the effectiveness of fusing lower attention features. From the last two rows of Table 2, AF-CNN achieves a higher AUC value than A-CNN, which strongly proves that the attention feature fusion method can effectively improve the performance.

Table 2. Quantitative evaluation results on ISIC 2017 test set (Ensemble: Whether an integration approach is used. External Data: Whether additional training data is used.)

| Methods       | Ensemble | External Data | AUC  |
|---------------|----------|---------------|------|
| VGG19         | N        | N             | 0.847|
| VGG19-GAP     | N        | N             | 0.859|
| A-CNN         | N        | N             | 0.872|
| AF-CNN        | N        | N             | 0.885|

3.3.3. Comparison with previous methods

As shown in Table 3, the comparisons between our proposed AF-CNN with previous methods are presented. The previous methods include the top three[9-11] in ISIC2017 classification competition, Mahbod et al.[12] and Zhang et al.[13]. It can be seen from the second column of Table 3 that [10, 11] and [12] all adopt an integrated approach to improve the performance. The third column in the Table 3 shows that all methods except [12] used additional training data. Our proposed method neither adopts a complicated and time-consuming integration method, nor uses additional training data, and obtains the state-of-the-art performance.

Table 3. Compare with previous methods on ISIC2017 testing dataset

| Methods                  | Ensemble | External data | AUC  |
|--------------------------|----------|---------------|------|
| ISIC2017 Winner1[10]     | Y        | Y             | 0.868|
| ISIC2017 Winner2 [9]     | N        | Y             | 0.856|
| ISIC2017 Winner3 [11]    | Y        | Y             | 0.874|
| Mahbod et al.[12]        | Y        | N             | 0.873|
| Zhang et al.[13]         | N        | Y             | 0.875|
| **Ours**                 | N        | N             | **0.885**|

3.4. Effect of the degree of attention and the cascading mode in attention block

3.4.1. Degree of attention

As previously described in section 2, the proposed A-CNN model only has attention modules after the 3rd, 4th, and 5th pooling layers of the VGG19 backbone network. We conducted experiments to
compare multiple ways of adding attention modules. As shown in the table 4, compared with the results in the third row, the results of the first two methods are worse. We infer that this is because some information has been lost due to too much attention.

| the location of attention blocks | AUC  |
|----------------------------------|------|
| Pool1-Pool5                      | 0.862|
| Pool2-Pool5                      | 0.866|
| Pool3-Pool5                      | 0.872|

3.4.2. The cascading method in attention block

In the CBAM module proposed by Woo et al.[3], the channel-spatial cascade method has achieved good results. In our experiments, we compared two cascading methods (channel-spatial and spatial-channel) in the AF-CNN model, and the results in table 5 showed that the latter obtained better results.

| Cascading method                  | AUC  |
|----------------------------------|------|
| AF-CNN(channel-spatial)          | 0.856|
| AF-CNN(spatial-channel)         | 0.885|

4. Conclusion

In this paper, we propose an end-to-end AF-CNN model for accurate melanoma recognition. The proposed model combines attention mechanism with convolutional neural network, which encourages the model to focus on the lesion areas and improves the discriminative ability of convolutional neural network. Besides, we take advantage of the complementarity of features extracted from different layers to fuse lower attention features to form the final feature vector. The proposed AF-CNN model is evaluated on the ISIC2017 dataset. For the melanoma recognition task, the proposed model achieves the state-of-the-art performance with an AUC value of 0.885.

Acknowledgment

The work was jointly supported by National Key R&D Program of China (Grant Nos. 2018YFB1306600), National Natural Science Foundation of China (Grant Nos.61571372, 61672436, 61601376), Fundamental Science and Advanced Technology Research Foundation of Chongqing (cstc2017ycjBX0050, cstc2016jcyjA0547), Fundamental Research Funds for the Central Universities (Grant Nos. XDJK2016A001, XDJK2017A005).

References

[1] Siegel, R. L., Miller, K. D., Fedewa, S. A., et al., (2017) Colorectal cancer statistics, 2017. CA: a cancer journal for clinicians, 67(3): pp. 177-193.
[2] Apalla, Z., Nashan, D., Weller, R. B., et al., (2017) Skin cancer: epidemiology, disease burden, pathophysiology, diagnosis, and therapeutic approaches. Dermatology and therapy, 7(1): pp. 5-19.
[3] Woo, S., Park, J., Lee, J.-Y., et al. (Sep 2018), Cbam: Convolutional block attention module. In: Proceedings of the European Conference on Computer Vision (ECCV). Munich, Germany. pp. 3-19.
[4] Lin, M., Chen, Q., and Yan, S., (2013) Network in network. arXiv preprint arXiv:1312.4400.
[5] Codella, N. C., Gutman, D., Celebi, M. E., et al. (Apr 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: IEEE 15th International Symposium on Biomedical Imaging. Washington, DC, USA. pp. 168-172.
[6] Paszke, A., Gross, S., Chintala, S., et al., (2017) Automatic differentiation in pytorch.
[7] Lin, T.-Y., Goyal, P., Girshick, R., et al. (Oct 2017), Focal loss for dense object detection. In: IEEE International Conference on Computer Vision. Venice, Italy. pp. 2980-2988.
[8] Zhou, B., Khosla, A., Lapedriza, A., et al. (Jun 2016), Learning deep features for discriminative
localization. In: IEEE Conference on Computer Vision and Pattern Recognition Las Vegas, Nevada, USA. pp. 2921-2929.

[9] Diaz, I. G., (2017) Incorporating the knowledge of dermatologists to convolutional neural networks for the diagnosis of skin lesions. arXiv preprint arXiv:1703.01976.

[10] Matsunaga, K., Hamada, A., Minagawa, A., et al., (2017) Image classification of melanoma, nevus and seborrheic keratosis by deep neural network ensemble. arXiv preprint arXiv:1703.03108.

[11] Menegola, A., Tavares, J., Fornaciali, M., et al., (2017) RECOD titans at ISIC challenge 2017. arXiv preprint arXiv:1703.04819.

[12] Mahbod, A., Schaefer, G., Effinger, I., et al., (2019) Fusing fine-tuned deep features for skin lesion classification. Computerized Medical Imaging and Graphics, 71: pp. 19-29.

[13] Zhang, J. P., Xie, Y. T., Xia, Y., et al., (2019) Attention Residual Learning for Skin Lesion Classification. IEEE Transactions on Medical Imaging, 38(9): pp. 2092-2103.