Melanoma Recognition via Visual Attention

Yiqi Yan(✉), Jeremy Kawahara, and Ghassan Hamarneh

Medical Image Analysis Lab, Simon Fraser University, Burnaby, BC, Canada
yiqiy@sfu.ca

Abstract. We propose an attention-based method for melanoma recognition. The attention modules, which are learned together with other network parameters, estimate attention maps that highlight image regions of interest that are relevant to lesion classification. These attention maps provide a more interpretable output as opposed to only outputting a class label. Additionally, we propose to utilize prior information by regularizing attention maps with regions of interest (ROIs) (e.g., lesion segmentation or dermoscopic features). Whenever such prior information is available, both the classification performance and the attention maps can be further refined. To our knowledge, we are the first to introduce an end-to-end trainable attention module with regularization for melanoma recognition. We provide both quantitative and qualitative results on public datasets to demonstrate the effectiveness of our method. The code is available at https://github.com/SaoYan/IPMI2019-AttnMel.

1 Introduction

Melanoma is one of the deadliest skin cancers in the world. The American Cancer Society reported that over 70% of skin cancer related deaths in the U.S. are associated with melanoma [19]. Fortunately, early diagnosis can facilitate proper treatment. However, accurate diagnosis of melanoma is non-trivial and requires expert human knowledge. Many automatic algorithms were proposed to classify melanoma from dermoscopy images. Particularly, deep learning based methods have been used in top-performing approaches [3,7].

Many deep learning methods turned to network or feature ensembles. Harangi et al. [8] trained an ensemble of AlexNet, VGGNet, and GoogLeNet, fusing their final features for a shared softmax classification layer. Codella et al. [2] trained an SVM using both deep convolutional features and sparse coding, which they later extended to an ensemble of 8 different features [4]. Similarly, Yu et al. [25,26] aggregated deep network features and fisher vector encoding. Training ensemble methods is time-consuming and is sensitive to how different models or feature extractors are tuned.

Other works trained a segmentation network to guide the classification. Yu et al. [24] designed a two-stage method. In the first step, a segmentation network was trained, which was used to detect and crop the lesion from the

A. C. S. Chung et al. (Eds.): IPMI 2019, LNCS 11492, pp. 793–804, 2019.
https://doi.org/10.1007/978-3-030-20351-1_62

entire image. Then a classification network was trained using the cropped images. Yang et al. and Chen et al. exploited the lesion segmentation in a parallel manner by applying a multi-task model that simultaneously tackled the problems of segmentation and classification [1,23]. When pixel-level annotations are not available, the training of these models becomes infeasible.

Although deep learning methods are widely used for skin lesion analysis, only a few efforts have been made to interpret which part of the image the model "concentrates" on. Van Molle et al. [21] visualized CNN features by rescaling the feature map to the input size and overlapping it with the input image. They attempted to gain insights into which image regions contribute to the results. They observed that the features seem to focus on specific characteristics, such as skin color, lesion border, hair, and artifacts, but there were no specific conclusions on how these features correlate with classification. A similar feature visualization was performed by Kawahara et al. [12]. Wu et al. [22] sought image biomarkers through *prediction difference analysis*. Specifically, a certain image region was corrupted each time, and the importance of that region was represented by the difference between the prediction scores based on the original and the corrupted images. Prediction difference analysis is a post-processing method designed to explain a fully trained network, while our model is trained end-to-end with learnable attention maps.

In this paper, we leverage attention mechanisms for melanoma recognition. A similar idea was presented by Ge et al. [6], who computed a class activation map (CAM) [27] as a saliency map to assign spatial weights to bilinear pooling features. CAM is a post-hoc analysis technique that requires extra computation based on a fully trained classification network. Similar to the works of Jetley et al. and Schlemper et al. [10,18], we propose an end-to-end solution via a trainable attention module. Our model extends the linear attention module proposed by Jetley et al. to more complex non-linear computations. Additionally, we propose to regularize the attention maps in order to train the model to focus on the expected regions of interest (ROIs). Our model not only yields state-of-the-art classification performance, but also produces attention maps indicating relevant image regions for classification. Our contributions are as follows:

- We incorporate end-to-end trainable attention modules for melanoma recognition. The attention maps automatically highlight image regions that are relevant to classification, which produces additional interpretable information as opposed to a mere class label. We perform a series of ablation studies to examine the effectiveness of attention.
- We introduce a method to efficiently utilize prior information via regularizing the attention maps with regions of interest (ROIs) (e.g., lesion segmentation, dermoscopic features). With prior information, the learned attention maps are refined and the classification performance is improved.
- The proposed regularization method can also be used to validate the effectiveness of ROI priors. For example, we show that regularizing using image background impedes the performance. This confirms that the model is properly

deeming the background less relevant to classification compared to the areas of skin lesion and dermoscopic features.

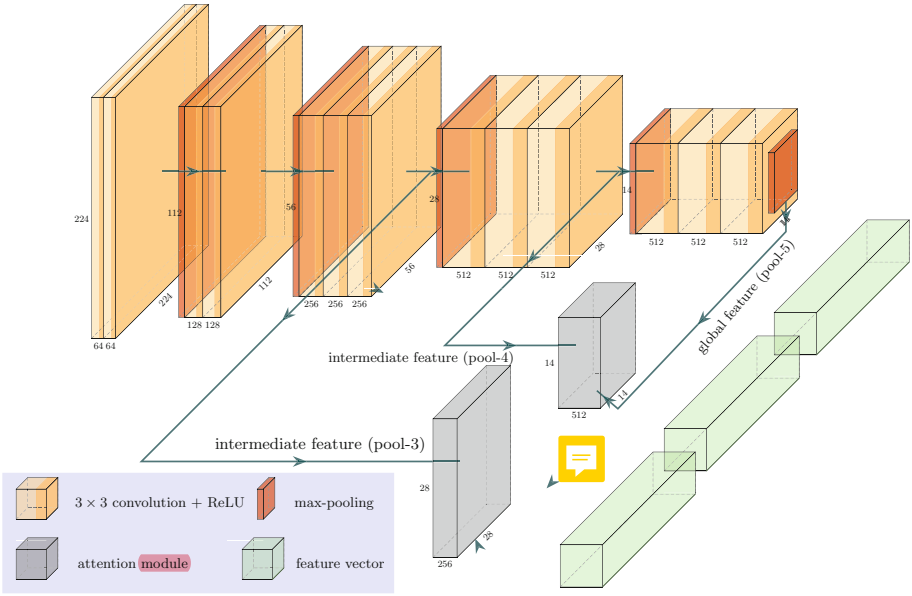

Fig. 1. The overall network architecture. The backbone network is VGG-16 (the yellow and red blocks) without any dense layers. Two attention modules (described in Fig. 2) are applied (the gray blocks). The three feature vectors (green blocks) are computed via global average pooling and are concatenated together to form the final feature vector, which serves as the input to the classification layer. The classification layer is not shown here. (Color figure online)

2 Proposed Method

2.1 Network Architecture

The human vision system focuses on objects in its field-of-view that are relevant to the task at hand. For example, when diagnosing skin cancer, dermatologists may focus more on the lesion rather than irrelevant areas such as background or hair. To imitate this visual exploration pattern, we use an attention module to estimate a spatial (pixel-wise) attention map. The proposed network architecture is illustrated in Fig. 1, with the attention modules shown as gray blocks. The inner details of the attention module are shown in Fig. 2.

We adopt VGG-16 [20], with all dense layers removed, as the backbone network of our model. We exploit intermediate feature maps (pool-3 and pool-4 in VGG-16) to infer attention maps. When computing the attention maps, the

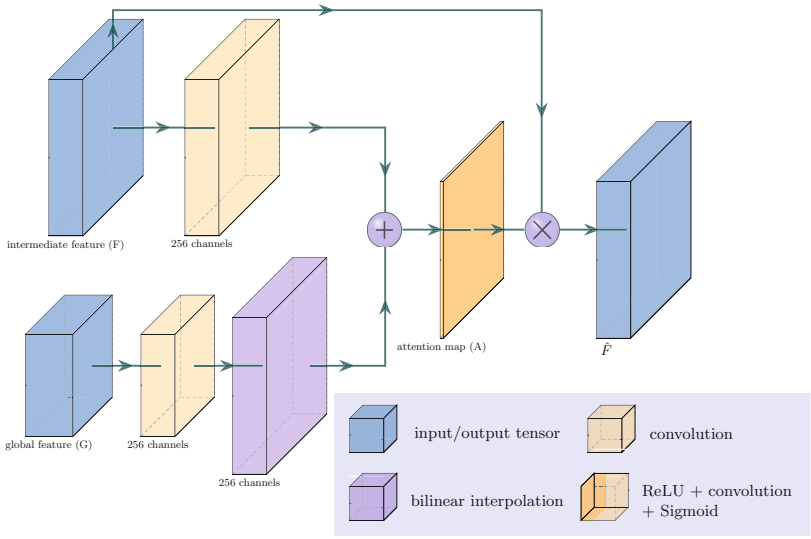

Fig. 2. Inner architecture of the attention module (i.e., the gray blocks in Fig. 1). When the spatial size of global and intermediate features are different, feature upsampling is done via bilinear interpolation. The sum operation is element-wise, and the multiplication is "pixel-wise" following Eq. 3

output of pool-5 serves as a form of "global guidance" (denoted as $\boldsymbol{\mathcal{G}}$) because the last-stage feature contains the most compressed and abstracted information over the entire image. Let $\boldsymbol{\mathcal{F}} = (\boldsymbol{f}_1, \boldsymbol{f}_2, \ldots, \boldsymbol{f}_n)$ denote the intermediate feature, where $\boldsymbol{f}_i$ is the feature vector at the i-th spatial location. $\boldsymbol{\mathcal{F}}$ and $\boldsymbol{\mathcal{G}}$ are fed through an attention module (Fig. 2), yielding a one-channel response $\boldsymbol{\mathcal{R}}$,

$$\boldsymbol{\mathcal{R}} = \boldsymbol{W} \circledast \mathrm{ReLU}\big(\boldsymbol{W}_f \circledast \boldsymbol{\mathcal{F}} + up\big(\boldsymbol{W}_g \circledast \boldsymbol{\mathcal{G}}\big)\big), \tag{1}$$

where $\circledast$ represents a convolutional operation, $\boldsymbol{W}_f$ and $\boldsymbol{W}_g$ are convolutional kernels with 256 filters, and the convolutional kernel $\boldsymbol{W}$ outputs a single channel. $up(\cdot)$ is bilinear interpolation that aligns the spatial size.

The attention map $\boldsymbol{\mathcal{A}}$ is then calculated as the normalization of $\boldsymbol{\mathcal{R}}$,

$$\boldsymbol{\mathcal{A}} = Sigmoid\big(\boldsymbol{\mathcal{R}}\big). \tag{2}$$

Each scalar element $a_i \in \boldsymbol{\mathcal{A}}$ represents the degree of attention to the corresponding spatial feature vector in $\boldsymbol{\mathcal{F}}$. The feature map with attention ($\hat{\boldsymbol{\mathcal{F}}}$) is then computed by "pixel-wise" multiplication. That is, each feature vector $\boldsymbol{f}_i$ is multiplied by the attention element,

$$\hat{\boldsymbol{f}}_i = a_i \cdot \boldsymbol{f}_i. \tag{3}$$

Now that we have the attention version of pool-3 and pool-4 features ($\hat{\boldsymbol{\mathcal{F}}}^{(3)}$, $\hat{\boldsymbol{\mathcal{F}}}^{(4)}$), we obtain the final feature vector by concatenating the global average

pooling of $\hat{\boldsymbol{\mathcal{F}}}^{(3)}$, $\hat{\boldsymbol{\mathcal{F}}}^{(4)}$, and $\boldsymbol{\mathcal{G}}$ (green blocks in Fig. 1). A softmax classification layer is then formed based on this final feature. The whole network is trained end-to-end.

2.2 Regularization via Regions of Interest

Given binary maps of some specific ROIs, we incorporate these maps as prior information to guide the attention maps. To this end, we introduce a regularization term where these ROIs serve as a reference. Inspired by [11], we minimize a negative Sørensen-Dice-F1 loss,

$$\mathcal{L}_D(\boldsymbol{\mathcal{A}}, \bar{\boldsymbol{\mathcal{A}}}) = 1 - D(\boldsymbol{\mathcal{A}}, \bar{\boldsymbol{\mathcal{A}}}) = 1 - \frac{2 \cdot \sum_{i=1}^{n} (a_i \cdot \bar{a}_i)}{\sum_{i=1}^{n} (a_i + \bar{a}_i)}, \tag{4}$$

where $\bar{\boldsymbol{\mathcal{A}}}$ is a reference binary map of ROIs. We do not compute $\mathcal{L}_D$ per image to avoid division-by-zero when there exists $\bar{\boldsymbol{\mathcal{A}}}$ with no positive pixel labels. Instead, we treat one batch of data as a high dimensional tensor and calculate $\mathcal{L}_D$ using these two tensors. The overall loss with regularization becomes

$$\mathcal{L} = \mathcal{L}_{focal} + \lambda_1 \mathcal{L}_D(\boldsymbol{\mathcal{A}}^{(3)}, \bar{\boldsymbol{\mathcal{A}}}^{(3)}) + \lambda_2 \mathcal{L}_D(\boldsymbol{\mathcal{A}}^{(4)}, \bar{\boldsymbol{\mathcal{A}}}^{(4)}), \tag{5}$$

where $\mathcal{L}_{focal}$ is the focal loss [13], which is a modified cross-entropy loss designed to deal with imbalanced training data; $\boldsymbol{\mathcal{A}}^{(3)}$, $\boldsymbol{\mathcal{A}}^{(4)}$ are the attention maps corresponding to pool-3 and pool-4 with $\bar{\boldsymbol{\mathcal{A}}}^{(3)}$, $\bar{\boldsymbol{\mathcal{A}}}^{(4)}$ being their reference maps respectively. The original reference maps, which have the same size as the input image, are downsampled to the size of pool-3 and pool-4 before computing the loss. We fix $\lambda_1 = 0.001$, $\lambda_2 = 0.01$. λ_2 has a larger value as the features in the deeper layers should be more discriminative.

3 Experiments

3.1 Implementation Details

Data Preparation and Preprocessing. Our experiments are performed on ISIC 2016 [7] and ISIC 2017 [3]. ISIC 2016 contains two classes: benign and malignant (melanoma). While in ISIC 2017 there are three classes: melanoma, nevus, and seborrheic keratosis. Participants were tasked with two independent binary classification tasks: melanoma vs others, and seborrheic keratosis vs others. We focus on melanoma recognition, which is the harder task. For a fair comparison, we use the exact same training, validation, and test sets as were provided in the challenge. We preprocess the data by center-cropping the image to a squared size with the length of each side equal to $0.8 \times \min(Height, Width)$, and then resizing to 256×256.

Dealing with Imbalanced Data. The ISIC dataset is highly imbalanced. For example, there are 304 benign and 75 malignant samples in the training set of

ISIC 2016. Classifiers are prone to bias towards the more frequent label. We perform data oversampling in our experiments. Besides, we use focal loss [13] as the main classification loss term in Eq. 5, as it can automatically down-weight easy samples in the training set.

Table 1. Quantitative results on ISIC 2016 test set. The first ranking in terms of AP or AUC is highlighted in **bold**, and the second ranking is indicated in *italics*. **The proposed method (*AttnMel-CNN*) achieves state-of-the-art without using an ensemble of models or ground truth segmentations. Notations**: *AP*: average precision; *AUC*: the area under the ROC curve; *Lesion*: requires lesion segmentation or not; *Interp*: interpretable or not; *Ensemble*: ensemble method or not.

		AP	AUC	Lesion	Interp	Ensemble
#1	Yu et al. [24]	0.637	0.804	✓	✗	✗
#2	Codella et al. [4]	0.596	0.808	✗	✗	✓
#3	Yu et al. [25,26]	*0.685*	**0.852**	✗	✗	✓
#4	VGG-16	0.602	0.806	✗	✗	✗
#5	VGG-16-GAP	0.635	0.815	✗	✓	✗
#6	Mel-CNN	0.664	*0.844*	✗	✗	✗
#7	**AttnMel-CNN**	**0.693**	**0.852**	✗	✓	✗

Network Training. We implement our method using PyTorch [17]. The backbone network is initialized with VGG-16 pre-trained on ImageNet, and the attention modules are initialized using He's initialization [9]. The whole network is trained end-to-end for 50 epochs using stochastic gradient descent with momentum. The initial learning rate is 0.01 and is decayed by 0.1 every 10 epochs. We apply run-time data augmentation (random cropping, rotation, and flipping) via PyTorch transform modules.

Model Evaluation. The performance is evaluated over the test set based on the average precision (AP) and the area under the ROC curve (AUC)[1], as they were the official metrics used in the ISIC 2016 and 2017 challenge [3,7], respectively. We always pick the best epoch according to the area under the ROC curve (AUC) on the validation set, and report the final result on the test set.

3.2 Ablation Study

First, we train our model *without* regularization, i.e., only $\mathcal{L}_{focal}$ is used for training. We denote this model as *AttnMel-CNN*. We compare *AttnMel-CNN*

[1] We use APIs *average_precision_score* and *roc_auc_score* from scikit-learn toolbox (https://scikit-learn.org).

Table 2. Quantitative results on the ISIC 2017 test set. The highest rankings in terms of AP or AUC are highlighted in **bold**, and the second ranking is indicated in *italics*. **The proposed method with attention maps achieves comparable performance without external data, model ensembles, or any ground truth ROIs (*AttnMel-CNN*).** When ROIs are available, the performance is further improved. **Notation:** *AP*: average precision; *AUC*: the area under the ROC curve; *Lesion*: use lesion segmentation or not; *Dermo*: use dermoscopic features or not; *Interp*: interpretable or not; *Ensemble*: ensemble method or not; *External*: use external training data or not.

		AP	AUC	Lesion	Dermo	Interp	Ensemble	External
#1	ISIC 2017 Winner 1 [15]	–	0.868	✗	✗	✗	✓	✓
#2	ISIC 2017 Winner 2 [5]	–	0.856	✓	✓	✗	✗	✓
#3	ISIC 2017 Winner 3 [16]	–	*0.874*	✗	✗	✗	✓	✓
#4	Harangi et al. [8]	–	0.836	✗	✗	✗	✓	✗
#5	Mahbod et al. [14]	–	*0.873*	✗	✗	✗	✓	✓
#6	VGG-16	0.600	0.824	✗	✗	✗	✗	✗
#7	VGG-16-GAP	0.627	0.834	✗	✗	✓	✗	✗
#8	Mel-CNN	0.653	0.854	✗	✗	✗	✗	✗
#9	**AttnMel-CNN**	0.655	0.872	✗	✗	✓	✗	✗
#10	**AttnMel-CNN-Dermo**	*0.665*	0.864	✗	✓	✓	✗	✗
#11	**AttnMel-CNN-Lesion**	**0.672**	**0.883**	✓	✗	✓	✗	✗
#12	AttnMel-CNN-Bkg	0.647	0.849	✓	✗	✓	✗	✗

with three baselines (*VGG-16*, *VGG-16-GAP*, *Mel-CNN*) to verify the effectiveness of attention. Then we add regularization using different ROIs, yielding *AttnMel-CNN-Lesion* and *AttnMel-CNN-Dermo*. We also apply background (the inverse of lesion segmentation) as a "wrong" ROI to demonstrate how improper attention influence the performance. We discuss the details of each model in the following paragraphs.

Comparing with the Original VGG. The first baseline model is the original VGG network. We modify the last classification layer to have 2 output nodes, and the rest of the network parameters are initialized with ImageNet pre-training. We denote this baseline *VGG-16*. Note that even though our backbone network is based on the VGG network (Fig. 1), we remove the two dense layers and add our own attention modules. Since dense layers take nearly 90% of the parameters in *VGG-16*, our network is much more lightweight (around $100M$ fewer parameters). Referring to Table 1 (rows 4,7) and Table 2 (rows 6,9), *AttnMel-CNN* achieves better performance despite the large degree of parameter reduction.

Comparison with the Truncated VGG. The poor performance of the original *VGG-16* could be due to overfitting. For a fair comparison, we design another

baseline, termed *VGG-16-GAP*, by replacing the dense layers with global average pooling. Note that this is also equivalent to our model without attention. Referring to Tables 1 and 2, *VGG-16-GAP* slightly outperforms the original *VGG-16*, but is surpassed by the proposed *AttnMel-CNN*. This demonstrates that overfitting can be reduced by removing the dense layers, but that further improvements come from the proposed architecture, which explicitly leverages the intermediate features.

Does Attention Help? After confirming the usefulness of intermediate features, one may ask whether it helps to assign attention maps to these features. In order to validate the effectiveness of attention modules themselves, we compute global average pooling on pool-3 and pool-4 instead of their attention versions. We denote this baseline *Mel-CNN*. According to Tables 1 and 2, this baseline yields worse performance than *AttnMel-CNN*. This is an expected result because shallow features are not well compressed and abstracted, and attention maps help rule out irrelevant information within shallow features.

How does the Regularization Influence the Model? We re-train the network using the loss proposed in Eq. 5 with three different reference maps ($\bar{\boldsymbol{\mathcal{A}}}$): (1) *AttnMel-CNN-Lesion* uses the whole lesion segmentation map (available from ISIC 2017 Task 1); (2) *AttnMel-CNN-Dermo* uses the union of four dermoscopic features[2] (available from ISIC 2017 Task 2); and (3) *AttnMel-CNN-Bkg* uses image background (the inverse of whole lesion segmentation). Table 2 shows that encouraging attention to lesion or dermoscopic features yields better performance, while improper attention (*AttnMel-CNN-Bkg*) harms the performance.

3.3 Visual Interpretability

In order to show whether better attention correlates with higher performance, we evaluate the learned attention maps both qualitatively and quantitatively.

Qualitative Analysis. We visualize the learned attention maps of *AttnMel-CNN*, *AttnMel-CNN-Lesion* and *AttnMel-CNN-Dermo* on the ISIC 2017 test data by upsampling $\boldsymbol{\mathcal{A}}$ (Eq. 2) to align with the input image. The results are shown in Fig. 3. When comparing rows 2 and 3, we observe that the shallower layer (pool-3) tends to focus on more general and diffused areas, while the deeper layer (pool-4) is more concentrated, focusing on the lesion and avoiding irrelevant objects. Furthermore, rows 4–7 demonstrate that the models with additional regularization pay attention to more semantically meaningful regions, which accounts for the performance improvement illustrated in Table 2.

Quantitative Analysis. We quantify the "quality" of the learned attention map by computing its overlap with the ground truth lesion segmentation. First,

[2] We convert the superpixel labels to binary pixel labels in the same way as [11], and use the union across all the dermoscopic features.

Table 3. Jaccard index of (binarized) attention maps and class activation maps with respect to the ground truth lesion segmentations.

AttnMel-CNN		AttnMel-CNN-Dermo		AttnMel-CNN-Lesion		VGG-16-GAP
pool3	pool4	pool3	pool4	pool3	pool4	CAM
0.3105	0.3186	0.3621	0.4767	0.5533	0.6560	0.2825

we re-normalize each attention map to $[0, 1]$ and binarize it using a threshold of 0.5. Then we compute the Jaccard index with respect to the ground truth lesion segmentation. We also calculate the class activation map (CAM) [27] from *VGG-16-GAP* and follow the same procedure as above to compute the Jaccard index value. The results reported in Table 3 lead to several conclusions: (1) The proposed learnable attention module highlights the relevant image regions better than the post-processing-based attention (CAM). (2) The attention map of the deeper layer (pool-4) yields a higher Jaccard index value, demonstrating that the deeper layer learns more discriminative features than the shallower layer. (3) The regularization encourages the attention maps to concentrate more on relevant ROIs.

3.4 Comparison with Previous Methods

We summarize previous work in Table 1 rows 1–3 and Table 2 rows 1–5. Comparison with [1,23] is not feasible as separate results for melanoma classification are not reported. The advantages of our method are:

- Our method yields state-of-the-art performance for melanoma classification even without additional regularization (*AttnMel-CNN*), and produces further performance improvements when reference ROIs are available (*AttnMel-CNN-Lesion* and *AttnMel-CNN-Dermo*). Additionally, we achieve state-of-the-art performance without any external training data.
- Our method relies on a single model, avoiding complex model ensembles.
- Compared with other methods utilizing segmentation maps [1,5,23,24], our method is more robust and flexible in that: (1) One of our models (*AttnMel-CNN*), optimized using only the focal loss, performs well without any regions of interests, while network training in those competing works requires pixel-wise annotations. (2) The competing works can only utilize whole lesion segmentations, but our regularization method can efficiently use dermoscopic (*AttnMel-CNN-Dermo*). We note that in a fair number of images, no dermoscopic features occur, and our proposed model is improved through these "sparse" reference maps.

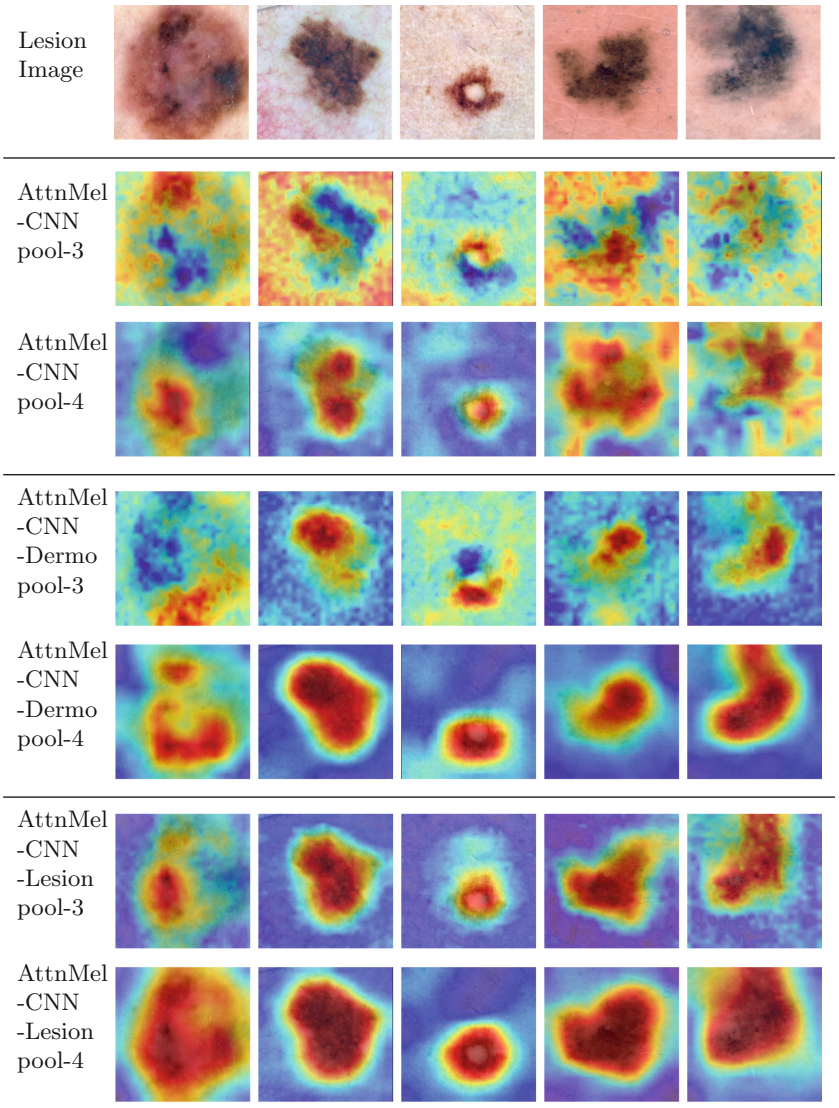

Fig. 3. Visualization of attention maps for different models. The deeper layer (pool-4) exhibits more concentrated attention to valid regions than the shallower layer (pool-3). The models with additional regularization (rows 4–7) produce more refined and semantically meaningful attention maps.

4 Conclusion and Discussion

In this paper, we proposed an attention-based network for melanoma recognition with a novel technique to regularize the attention maps with prior information. We achieve state-of-the-art performance for melanoma classification on

two public datasets without external training data or complex model ensembles. One limitation of this work is that we only apply the model to a binary classification task. Future work would explore visual attention in more general skin lesion classification problems.

Acknowledgement. Partial funding for this project is provided by the Natural Sciences and Engineering Research Council of Canada (NSERC). The authors are grateful to the NVIDIA Corporation for donating a Titan X GPU used in this research. We use PlotNeuralNet (https://github.com/HarisIqbal88/PlotNeuralNet) for drawing the network diagrams in this paper.

References

1. Chen, S., Wang, Z., Shi, J., Liu, B., Yu, N.: A multi-task framework with feature passing module for skin lesion classification and segmentation. In: IEEE International Symposium on Biomedical Imaging, pp. 1126–1129 (2018)
2. Codella, N., Cai, J., Abedini, M., Garnavi, R., Halpern, A., Smith, J.R.: Deep learning, sparse coding, and SVM for melanoma recognition in dermoscopy images. In: Zhou, L., Wang, L., Wang, Q., Shi, Y. (eds.) MLMI 2015. LNCS, vol. 9352, pp. 118–126. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24888-2_15
3. Codella, N.C., et al.: 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 International Symposium on Biomedical Imaging, pp. 168–172 (2018)
4. Codella, N.C., et al.: Deep learning ensembles for melanoma recognition in dermoscopy images. IBM J. Res. Dev. **61**(4), 1–15 (2017)
5. Díaz, I.G.: Incorporating the knowledge of dermatologists to convolutional neural networks for the diagnosis of skin lesions. arXiv preprint arXiv:1703.01976 (2017)
6. Ge, Z., Demyanov, S., Chakravorty, R., Bowling, A., Garnavi, R.: Skin disease recognition using deep saliency features and multimodal learning of dermoscopy and clinical images. In: Descoteaux, M., Maier-Hein, L., Franz, A., Jannin, P., Collins, D.L., Duchesne, S. (eds.) MICCAI 2017. LNCS, vol. 10435, pp. 250–258. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-66179-7_29
7. Gutman, D., et al.: Skin lesion analysis toward melanoma detection: a challenge at the international symposium on biomedical imaging (ISBI) 2016, hosted by the international skin imaging collaboration (ISIC). arXiv preprint arXiv:1605.01397 (2016)
8. Harangi, B., Baran, A., Hajdu, A.: Classification of skin lesions using an ensemble of deep neural networks. In: 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, pp. 2575–2578. IEEE (2018)
9. He, K., Zhang, X., Ren, S., Sun, J.: Delving deep into rectifiers: surpassing human-level performance on imagenet classification. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 1026–1034 (2015)
10. Jetley, S., Lord, N.A., Lee, N., Torr, P.H.: Learn to pay attention. In: International Conference on Learning Representation (2018)
11. Kawahara, J., Hamarneh, G.: Fully convolutional neural networks to detect clinical dermoscopic features. IEEE J. Biomed. Health Inform. **23**(2), 578–585 (2019)
12. Kawahara, J., Daneshvar, S., Argenziano, G., Hamarneh, G.: Seven-point checklist and skin lesion classification using multitask multimodal neural nets. IEEE J. Biomed. Health Inform. **23**(2), 538–546 (2019)

13. Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollar, P.: Focal loss for dense object detection. In: 2017 IEEE International Conference on Computer Vision (ICCV), pp. 2999–3007. IEEE (2017)
14. Mahbod, A., Schaefer, G., Ellinger, I., Ecker, R., Pitiot, A., Wang, C.: Fusing fine-tuned deep features for skin lesion classification. Comput. Med. Imaging Graph. **71**, 19–29 (2018)
15. Matsunaga, K., Hamada, A., Minagawa, A., Koga, H.: Image classification of melanoma, nevus and seborrheic keratosis by deep neural network ensemble. arXiv preprint arXiv:1703.03108 (2017)
16. Menegola, A., Tavares, J., Fornaciali, M., Li, L.T., Avila, S., Valle, E.: Recod titans at isic challenge 2017. arXiv preprint arXiv:1703.04819 (2017)
17. Paszke, A., et al.: Automatic differentiation in pytorch. In: NIPS Workshop on Autodiff (2017)
18. Schlemper, J., et al.: Attention-gated networks for improving ultrasound scan plane detection. In: Medical Imaging with Deep Learning Conference (2018)
19. Siegel, R.L., Miller, K.D., Jemal, A.: Cancer statistics. CA: Cancer J. Clin. **67**(1), 7–30 (2017)
20. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. In: International Conference on Learning Representation (2015)
21. Van Molle, P., De Strooper, M., Verbelen, T., Vankeirsbilck, B., Simoens, P., Dhoedt, B.: Visualizing convolutional neural networks to improve decision support for skin lesion classification. In: Stoyanov, D., et al. (eds.) MLCN/DLF/IMIMIC -2018. LNCS, vol. 11038, pp. 115–123. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-02628-8_13
22. Wu, J., Li, X., Chen, E.Z., Jiang, H., Dong, X., Rong, R.: What evidence does deep learning model use to classify skin lesions? arXiv preprint arXiv:1811.01051 (2018)
23. Yang, X., Li, H., Wang, L., Yeo, S.Y., Su, Y., Zeng, Z.: Skin lesion analysis by multi-target deep neural networks. In: 2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), pp. 1263–1266. IEEE (2018)
24. Yu, L., Chen, H., Dou, Q., Qin, J., Heng, P.A.: Automated melanoma recognition in dermoscopy images via very deep residual networks. IEEE Trans. Med. Imaging **36**(4), 994–1004 (2017)
25. Yu, Z., Jiang, X., Wang, T., Lei, B.: Aggregating deep convolutional features for melanoma recognition in dermoscopy images. In: Wang, Q., Shi, Y., Suk, H.-I., Suzuki, K. (eds.) MLMI 2017. LNCS, vol. 10541, pp. 238–246. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-67389-9_28
26. Yu, Z., et al.: Melanoma recognition in dermoscopy images via aggregated deep convolutional features. IEEE Trans. Biomed. Eng. **66**, 1006–1016 (2018)
27. Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., Torralba, A.: Learning deep features for discriminative localization. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2921–2929 (2016)