Attention-based Skin Lesion Recognition

by

Yiqi Yan

B.Sc., Northwestern Polytechnical University, 2018

Thesis Submitted in Partial Fulfillment of the
Requirements for the Degree of
Master of Science

in the
School of Computing Science
Faculty of Applied Science

SIMON FRASER UNIVERSITY
Spring 2020

Approval

Name:	**Yiqi Yan**
Degree:	**Master of Science (Computing Science)**
Title:	**Attention-based Skin Lesion Recognition**
Examining Committee:	**Chair:** Yağız Aksoy Assistant Professor
	Ghassan Hamarneh Senior Supervisor Professor
	Manolis Savva Supervisor Assistant Professor
	Yasutaka Furukawa Examiner Associate Professor
Date Defended:	**April 7, 2020**

Abstract

Skin cancer is one of the most common types of cancers in the world and is a big concern for people's health. In recent years, automatic algorithms to recognize skin cancers from dermoscopy images have gained lots of popularity, especially deep-learning-based methods.

In this thesis, we propose an attention-based deep learning model for skin cancer 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). To our knowledge, we are the first to introduce an end-to-end trainable attention module with regularization for skin cancer recognition.

We provide both quantitative and qualitative results on public datasets to demonstrate the effectiveness of our method. Experiments show that: (1) the attention module is capable of ruling out irrelevant areas in the image; (2) when the proposed attention regularization terms are applied, both the classification performance and the attention maps can be further refined; (3) the attention regularization is quite robust and flexible in that it can take advantage of sparse or even imperfect ROI maps.

The code of this work is released at `https://github.com/SaoYan/IPMI2019-AttnMel`.

Keywords: skin cancer; deep learning; attention mechanism

Dedication

Dedicated to my beloved family and friends.

Acknowledgements

First of all, I'd like to thank my senior supervisor Prof. Ghassan Hamarneh for providing me with the opportunity of pursuing graduate study at SFU. His mentorship and passion made the study and research progress really enjoyable.

I'm also grateful to my labmate and collaborator Dr. Jeremy Kawahara for his insightful suggestions and ideas on my research. His rich experience in computer-aided skin cancer diagnosis facilitated this research to a large degree.

What's more, I'm very thankful to my parents for their endless and unconditional support, without which I'd never had the chance to study abroad an achieve my goals.

Gratitude should also be given to MITACS for offering me graduate fellowship in my first year at SFU, and all my friends for their accompany and support.

Table of Contents

List of Tables

List of Figures

Chapter 1

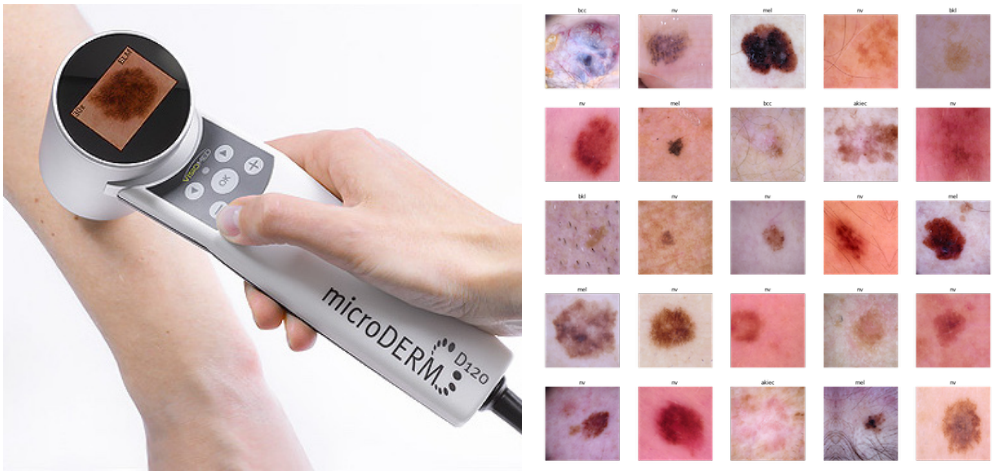
Introduction

1.1 Background

Skin diseases are quite common [37] and have been a big concern of people's health [15]. Early diagnosis of skin cancer is important for proper and timely treatment. Even though histopathology analysis has been the "gold standard" for recognizing skin lesions such as melanoma [41], biopsy is an invasive approach with higher cost and may even cause infections [43]. Non-invasive diagnosis [20, 11, 30], on the other hand, can reduce costs and avoid biopsy complications.

Among all the non-invasive methods, dermoscopy is one of the most common techniques. It provides enhanced imaging of deep levels of the skin by eliminating the skin's surface reflection (Fig. 1.1). Used by expert dermatologists, dermoscopy facilitates skin lesion diagnosis to a large degree. For example, medical research has witnessed an improvement of 49% ($p = 0.001$) in diagnostic accuracy for melanoma with dermoscopy technique, compared to standard photography [25].

In recent years, computer-aided skin cancer diagnosis has gained much popularity [13, 7, 6], including skin lesion recognition [21, 23], segmentation [33, 32, 18], and fermoscopic feature detection [24]. On the one hand, with the development of internet and inexpensive consumer dermatoscope attachments for smart phones [17], automated dermoscopic assessment algorithms can have a positive influence on patient care. For example, it has been shown that deep networks are capable of classifying skin cancer with an accuracy compa-



(a) (b)

Figure 1.1: (a) A typical dermoscopy imaging device (The image comes from `https://www.poliklinikabagatin.hr/`); (b) Dermoscopy image samples

rable to dermatologists [10]. On the other hand, computers can assist human experts by providing interpretable results [42, 46].

This thesis focuses on automated skin cancer classification based on dermoscopy images. An attention-based deep learning approach is proposed to not only achieve high accuracy, but also generate interpretable results to provide dermatologists with more insights into the diagnosis. Additionally, we propose to utilize prior information by regularizing attention maps with regions of interest (ROIs). 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 skin cancer recognition. We provide both quantitative and qualitative results on public datasets to demonstrate the effectiveness of our method.

1.2 Related Work

1.2.1 Network or Feature Ensembles

Ensemble-based methods either train multiple independent classifiers and combine their predictions (*network ensembles*, Fig. 1.2 a), or extract various kinds of features and pass the merged feature to one single classifier (*feature ensembles*, Fig. 1.2 b).

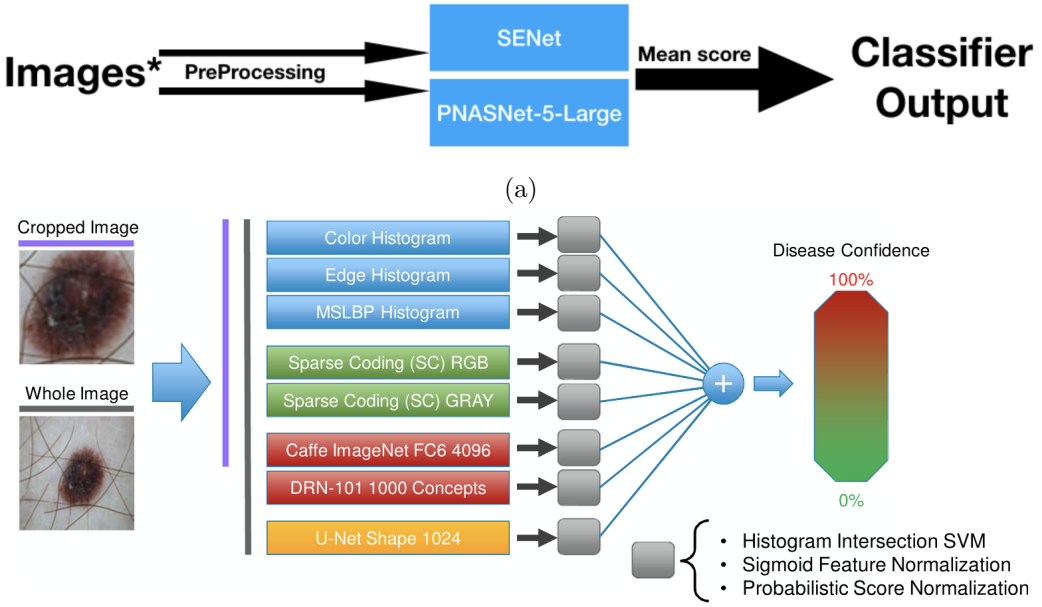


Figure 1.2: (a) Example of network ensemble (the figure is from Zhuang et al. [58]); (b) Example of feature ensemble (the figure is from Codella et al. [8])

Zhuang et al. [58] trained independent deep networks and used the mean confidence as the final output. Harangi et al. [14] used an ensemble of AlexNet, VGGNet, and GoogLeNet, fusing their final features for a shared softmax classification layer. Codella et al. [5] trained an SVM using both deep convolutional features and sparse coding, which they later extended to an ensemble of 8 different features [8]. Yu et al. [53, 54] aggregated deep network features and fisher vector encoding. Nozdryn-Plotnicki et al. [34] adopted as many as 18 deep models, trained them separately as feature extractors and built an XGBoost classifier on top of them.

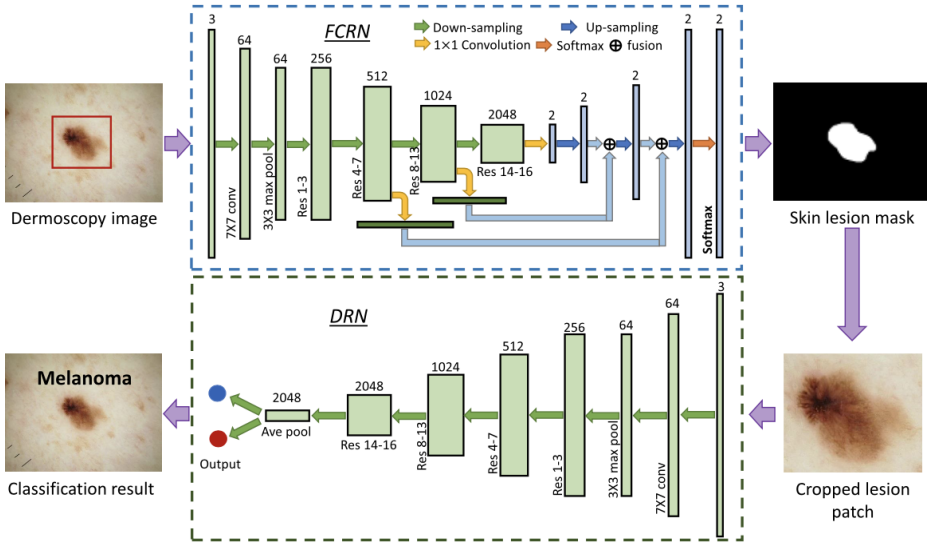
The main disadvantage of the ensemble-based method is the time-consuming training process. What's more, developers have to carefully strike the balance of all independent classifiers or feature extractors. Modification of one component may result in the tuning or even re-training of others. The work in this thesis adopts an end-to-end trained model, which is much easier to apply in the real situation.

1.2.2 Segmentation-guided Classification

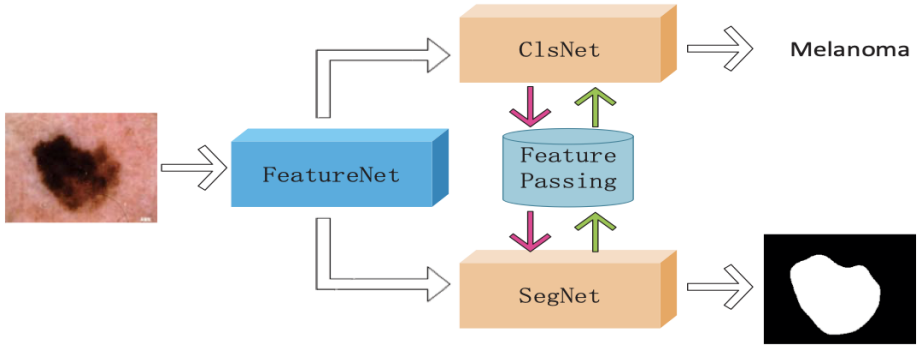
Several works trained a segmentation network to guide the classification (Fig. 1.3). Yu et al. [52] designed a two-stage method. In the first step, a segmentation network was trained,

which was used to detect and crop the lesion from the original image. Then a classification network was trained using the cropped image. Yang et al. [50] and Chen et al. [4] exploited the lesion segmentation in a parallel manner by applying a multi-task model that simultaneously tackled the problems of segmentation and classification.

These approaches require accurate and complete pixel-level annotations for each image in the training set, while in most cases only image-level labels or partial pixel labels (e.g., dermoscopic features, Fig. 2.4) are available. What's more, the two-stage method is not trained end-to-end. Each stage has to be tuned separately to achieve the optimal overall performance. This thesis aims at a more flexible model, which can take advantage of either complete or partial pixel-level labels but still works fine without them.



(a)



(b)

Figure 1.3: Segmentation-guided classification. (a) Sequential two-stage method [52] (b) Parallel method [4]

1.2.3 Visual Interpretability

Although deep learning methods have been widely used for skin lesion analysis, only a few efforts have been made on interpretability. Van Molle et al. [42] 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. [22]. Wu et al. [46] 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. Ge et al. [12] computed a class activation map (CAM) [57] as a saliency map to assign spatial weights to bilinear pooling features.

All these works involve some post hoc analysis that requires extra computation based on a fully trained classification network, while this thesis proposes a classification model with the learnable attention modules. No post-processing is required.

1.3 Attention Mechanism

Figure 1.4: Visualization of attention maps generated by the model proposed in Woo et al. [45]

The concept of attention was firstly introduced in recurrent neural networks (RNN) for machine translation [2]. The idea was rapidly extended to convolutional neural networks

(CNNs) and applied to various computer vision tasks such as image classification [44, 45], image caption [51, 48, 27], visual question answering [47, 1], etc. Whatever the model architecture and application, the key idea of visual attention is consistent: generating an *attention map* which assigns weight to each pixel. The higher the weight, the more contribution the pixel makes to the final result (i.e., paying *more attention* to that region). An illustration of visual attention mechanism is shown in Fig. 1.4.

Generally speaking, there are two kinds of visual attention: post hoc attention and learnable attention. *Post hoc attention* interprets fully-trained deep networks by post-processing. It doesn't touch the trained parameters of the model. Typical works include class activation maps [57] network dissection [3], and gradient-based saliency visualization [39, 38]. As the goal of this thesis is not merely interpretability, *learnable attention*, especially for image classification, is what we focus on. Woo et al. [45] proposed an attention module called CBAM which can be inserted into classic deep networks to refine the features. This thesis is most inspired by Jetley et al. [19] that leveraged attention modules to enhance intermediate features. This thesis extends the linear attention module proposed by Jetley et al. to more complex non-linear computations in order to handle images with much higher resolution. Most importantly, our model allows plug-in attention regularize terms to take advantage of any available pixel-level prior information.

1.4 Contribution

This thesis leverages attention mechanisms for skin cancer recognition. Besides, 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. The main 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 flexible method to efficiently utilize pixel-level 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 further 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.

This work was published in IPMI 2019 [49], and this thesis adds more experimental results on some latest datasets.

> Yiqi Yan, Jeremy Kawahara, and Ghassan Hamarneh. Melanoma recognition via visual attention. In *International Conference on Information Processing in Medical Imaging*, Lecture Notes in Computer Science, vol 11492, pages 793–804, Springer, 2019. DOI https://doi.org/10.1007/978-3-030-20351-1_62

Chapter 2

Methodology

2.1 Motivation

Previous research has pointed out that the shallow layers of deep networks capture some common patterns such as edges and texture, while the deep layers respond to more complex, class-specific features [55]. Unlike natural images of complicated scenes and objects, skin lesions have relatively simple structure, so even the shallow features can contain necessary information for skin cancer classification (e.g., different classes of lesions may have different shape and texture features). CNNs typically use the deepest global feature (i.e., the output of the final convolutional or pooling layer) for classification. Instead, we chose to directly leverage some intermediate features and combine them with the global one.

That being said, the shallow features do have too much "plain" information, and some may be irrelevant to the classification. This is where "attention" comes in. The attention map filters out the minor regions and produces a refined feature that would be more discriminative than the original one. The attention module should be differentiable so that the overall network can be trained end-to-end.

The proposed network architecture is illustrated in Fig. 2.1, with the attention modules shown as gray blocks. The inner structure of the attention module is shown in Fig. 2.2. In the next few sections, we will describe the details of our model.

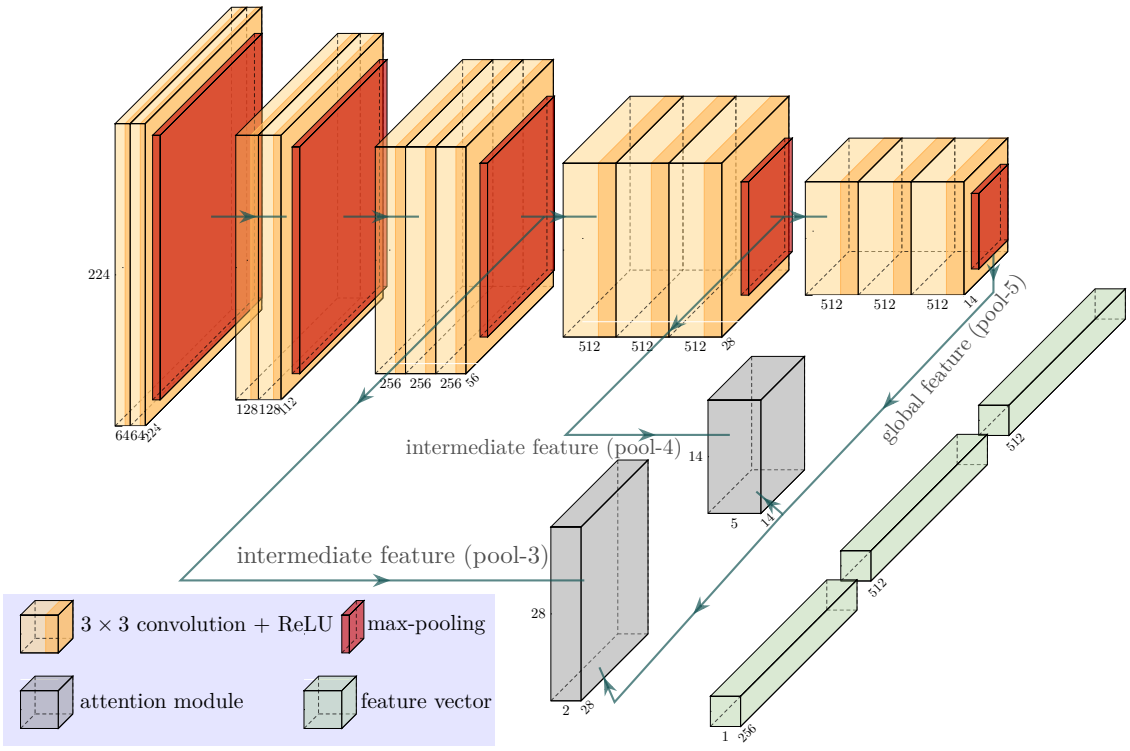


Figure 2.1: The overall network architecture. The backbone network is VGG-16 (the yellow and red blocks) without any dense layers. Two attention modules (described in Fig. 2.2) are applied (the gray blocks). The three feature vectors (green blocks) are computed via global average pooling and are concatenated together to form the final feature vector, which serves as the input to the classification layer. The classification layer is not shown here.

2.2 Network Architecture

We adopt VGG-16 [40], with all dense layers removed, as the backbone network of our model. We exploit the third and forth pooling features (pool-3 and pool-4) as the intermediate features. As the deepest feature (pool-5) contains the most compressed and abstracted information over the entire image, we use it as a global guidance (denoted as $\boldsymbol{\mathcal{G}}$) when computing the attention maps for pool-3 and pool-4. Let $\boldsymbol{\mathcal{F}} = (\boldsymbol{f}_1, \boldsymbol{f}_2, \ldots, \boldsymbol{f}_n)$ be any intermediate feature (either pool-3 or pool-4), 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. 2.2), yielding a one-channel response map $\boldsymbol{\mathcal{R}}$,

$$\boldsymbol{\mathcal{R}} = \boldsymbol{W} \circledast \mathrm{ReLU}(\boldsymbol{W}_f \circledast \boldsymbol{\mathcal{F}} + up(\boldsymbol{W}_g \circledast \boldsymbol{\mathcal{G}})), \tag{2.1}$$

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 $\mathcal{A}$ is then calculated as the per-pixel normalization of $\boldsymbol{\mathcal{R}}$,

$$\boldsymbol{\mathcal{A}} = Sigmoid(\boldsymbol{\mathcal{R}}). \tag{2.2}$$

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 refined feature map, denoted as $\hat{\boldsymbol{\mathcal{F}}}$, is then computed by "pixel-wise" multiplication. That is, each feature vector $\boldsymbol{f}_i$ is multiplied by the attention element,

$$\hat{\boldsymbol{f}}_i = a_i \cdot \boldsymbol{f}_i. \tag{2.3}$$

The above computation is applied separately and independently to pool-3 and pool-4, producing the refined features $\hat{\boldsymbol{\mathcal{F}}}^{(3)}$ and $\hat{\boldsymbol{\mathcal{F}}}^{(4)}$. 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. 2.1). A softmax classification layer is then formed based on this final feature.

2.3 Loss Function for Unbalanced Dataset

Typically skin cancer datasets are highly unbalanced, and the classic cross-entropy loss is prone to bias towards the more frequent classes. Focal loss [26] has proven to be effective in dealing with class imbalance.

The formula of the cross-entropy loss is

$$\mathcal{L}_{ce}(p_t) = -\log(p_t), \tag{2.4}$$

where p_t is the model's estimated probability for the ground-truth class. Focal loss adds a factor to the standard cross entropy

$$\mathcal{L}_{focal}(p_t) = -(1-p_t)^{\gamma}\log(p_t), \tag{2.5}$$

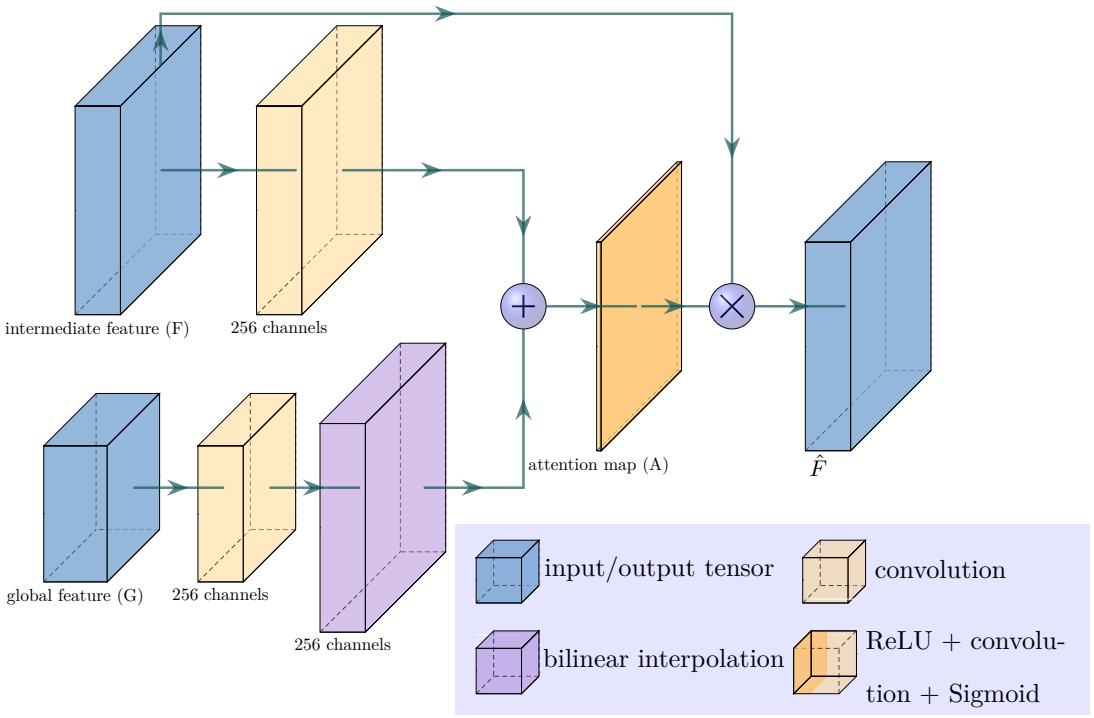


Figure 2.2: Inner architecture of the attention module (i.e., the gray blocks in Fig. 2.1). When the spatial size of global and intermediate features are different, feature upsampling is done via bilinear interpolation. The sum operation is element-wise, and the multiplication is “pixel-wise" following Eq. 2.3

where γ is a hyper-parameter.

When some sample is well classified ($p_t > 0.5$), the focal loss for it reduces so that more efforts will be put on other hard samples, thus training bias is avoided (Fig. 2.3).

2.4 Attention Regularization via Regions of Interest

The network is trained using focal loss when only image-level class labels are available. Sometimes we may have access to additional pixel-level annotations that specify regions of interest (ROIs), such as lesion segmentation and dermoscopic features (Fig. 2.4). Note that dermoscopic features can be pretty spares, and not all images have valid dermoscopic features. In fact it is a typical case for skin lesion datasets that a proportion of images have “empt” dermoscopic features (all-zero annotations).

Given binary or probability maps of some specific ROIs, we incorporate these maps as prior information to guide the attention maps. To this end, we introduce a regularization

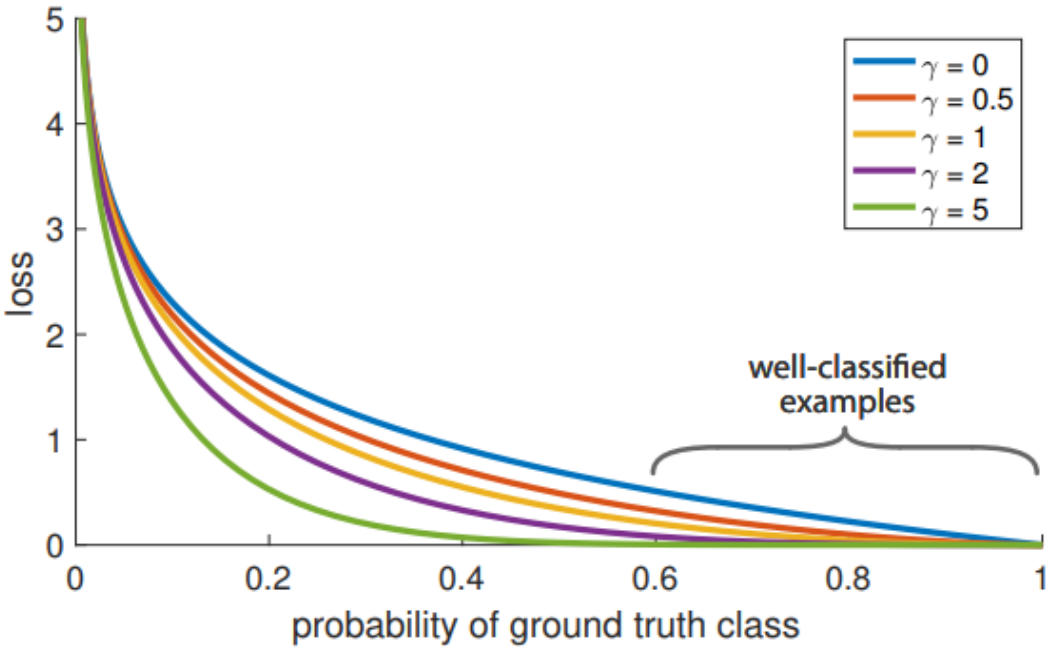


Figure 2.3: Focal loss reduces when the sample is well classified. The figure is borrowed from Lin et al. [26].

term where the ROIs serve as a reference. Inspired by Kawahara et al. [24], we minimize a negative Sørensen-Dice-F1 loss,

$$\mathcal{L}_D(\boldsymbol{\mathcal{A}}, \bar{\boldsymbol{\mathcal{A}}}) = 1 - D(\boldsymbol{\mathcal{A}}, \bar{\boldsymbol{\mathcal{A}}}) = 1 - \frac{2 \cdot \sum_{i=1}^{n} (a_i \cdot \bar{a}_i)}{\sum_{i=1}^{n} (a_i + \bar{a}_i)}, \tag{2.6}$$

where $\bar{\boldsymbol{\mathcal{A}}}$ is a reference map of ROIs and $\boldsymbol{\mathcal{A}}$ is the attention map produced in Eq. 2.2. We do not compute $\mathcal{L}_D$ per image to avoid division-by-zero when there exists $\bar{\boldsymbol{\mathcal{A}}}$ with all-zero values. Instead, we treat one batch of data as a high dimensional tensor and calculate $\mathcal{L}_D$ using tensors.

The proposed model generates attention maps in pool-3 and pool-4 layers ($\boldsymbol{\mathcal{A}}^{(3)}$, $\boldsymbol{\mathcal{A}}^{(4)}$), and we regularize both of them. Since the spatial dimension of attention maps are smaller than the original image, the given ROI maps are downsampled to the same size as $\boldsymbol{\mathcal{A}}^{(3)}$ and $\boldsymbol{\mathcal{A}}^{(4)}$ respectively. The downsampled maps are denoted as $\bar{\boldsymbol{\mathcal{A}}}^{(3)}$ and $\bar{\boldsymbol{\mathcal{A}}}^{(4)}$.

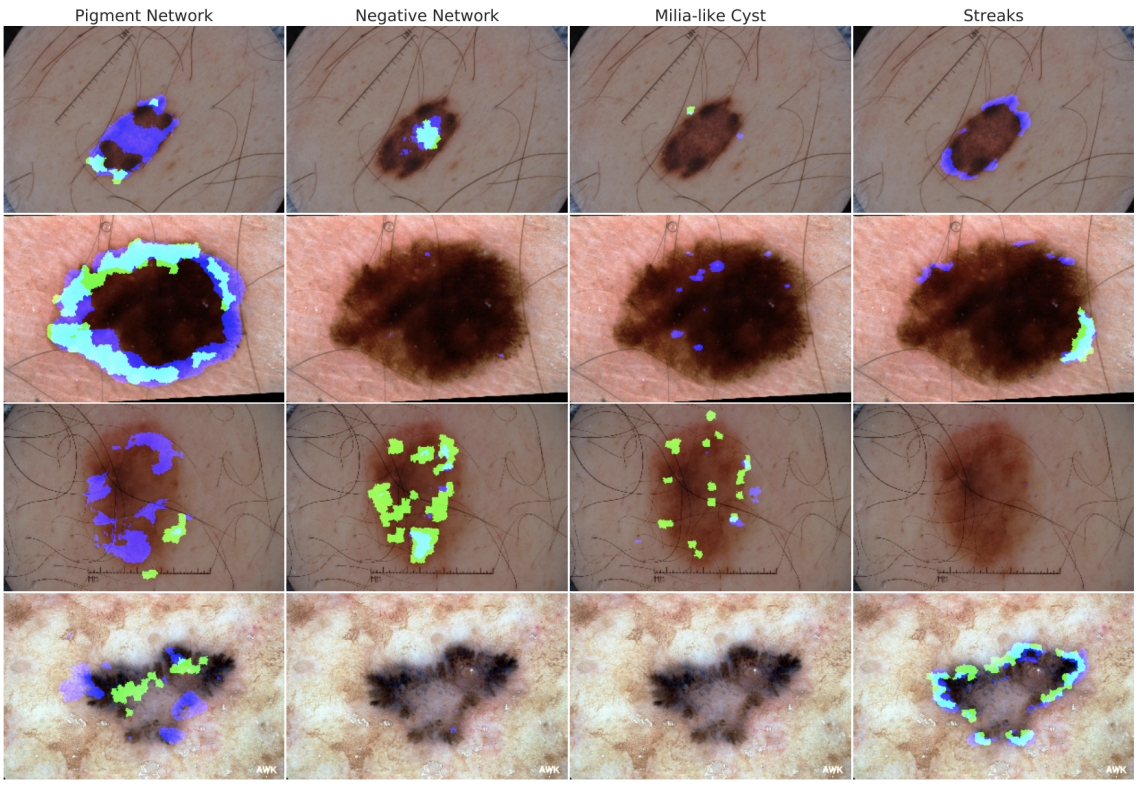


Figure 2.4: Examples of dermoscopic features. The figure is borrowed from Kawahara et al. [24]

The overall loss with regularization becomes

$$\mathcal{L} = \mathcal{L}_{focal} + \lambda_1 \mathcal{L}_D(\boldsymbol{\mathcal{A}}^{(3)}, \bar{\boldsymbol{\mathcal{A}}}^{(3)}) + \lambda_2 \mathcal{L}_D(\boldsymbol{\mathcal{A}}^{(4)}, \bar{\boldsymbol{\mathcal{A}}}^{(4)}), \tag{2.7}$$

where $\mathcal{L}_{focal}$ is the focal loss discussed in section 2.3. 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.

The loss function introduced in Ep. 2.7 is quite flexible in that the regularization terms can be plugged in whenever there are available ROI maps, and out when we only have class labels.

Chapter 3

Experiments and Results

3.1 Experimental Setup

3.1.1 Datasets and Evaluation Metrics

The international skin imaging collaboration (ISIC) has hosted several skin lesion analysis challenges [13, 7, 6]. We perform experiments on the datasets from three years' challenges.

- ISIC 2016 [13] contains two classes: benign and malignant (melanoma). There are 900 dermoscopic lesion images in JPEG format for training, and 379 images for testing. The participants were ranked by average precision score.[1]

- ISIC 2017 [7] has three classes: melanoma, nevus, and seborrheic keratosis. Participants were asked to perform two independent binary classification tasks: melanoma vs others, and seborrheic keratosis vs others. We do experiments on melanoma recognition, which is the harder task. The data is split into three parts: 2000 images for training, 150 from validating, and 600 for testing. The official metric is the area under the receiver operating characteristic curve (AUC).[2]

- ISIC 2018 [6] extends the dataset further to seven classes. There are 10015 images for training, 193 for validation, and 1512 for testing. The performance is evaluated by nor-

[1]The average precision can be computed using the function *average_precision_score* from scikit-learn toolbox (`https://scikit-learn.org`).

[2]The AUC score can be computed using the function *roc_auc_score* from scikit-learn toolbox (`https://scikit-learn.org`).

malized multi-class accuracy, which is the arithmetic mean of each class's classification accuracy.[3]

The statistic information of the training sets is shown in Fig. 3.1. Note that every dataset is highly unbalanced.

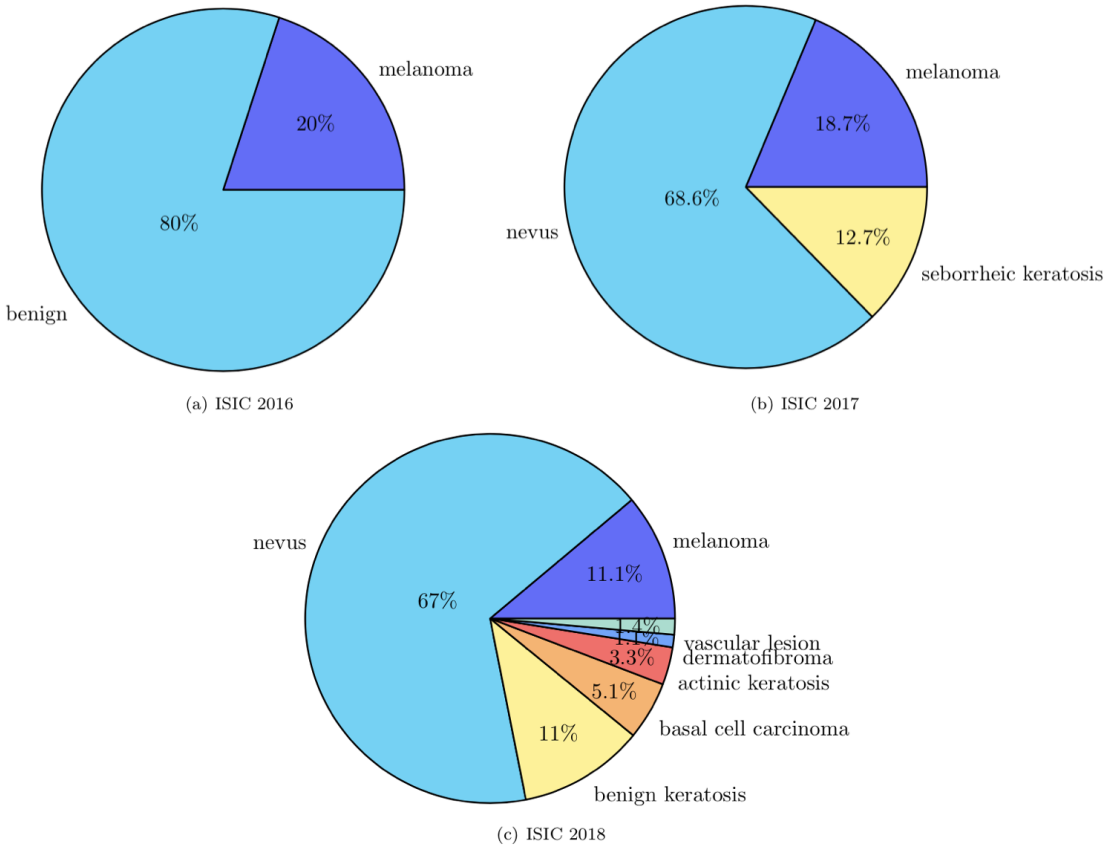


Figure 3.1: The statistic information of three training sets.

3.1.2 Implementation Details

Data Preprocessing 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.

[3]The multi-class accuracy can be computed using the function *recall_score* from scikit-learn toolbox (`https://scikit-learn.org`), with the *average* argument set to "macro".

Data Oversampling As was discussed in section 2.3, we use focal loss to tackle data imbalance. Additionally, we perform data oversampling by duplicating samples from the class with fewer instances.[4]

Network Training We implement our model using PyTorch [35]. The backbone network is initialized with VGG-16 pre-trained on ImageNet, and the attention modules are initialized using He's initialization [16]. 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's transform modules. For the datasets with validation set (ISIC 2017 and 2018), we do early-stopping and select the optimal model parameters depending on the performance on the validation set.

3.2 Binary Classification

In this section we'll discuss experimental results on ISIC 2016 and 2017. Both of them involve binary classification tasks. The results shown in this section are from our publication [49].

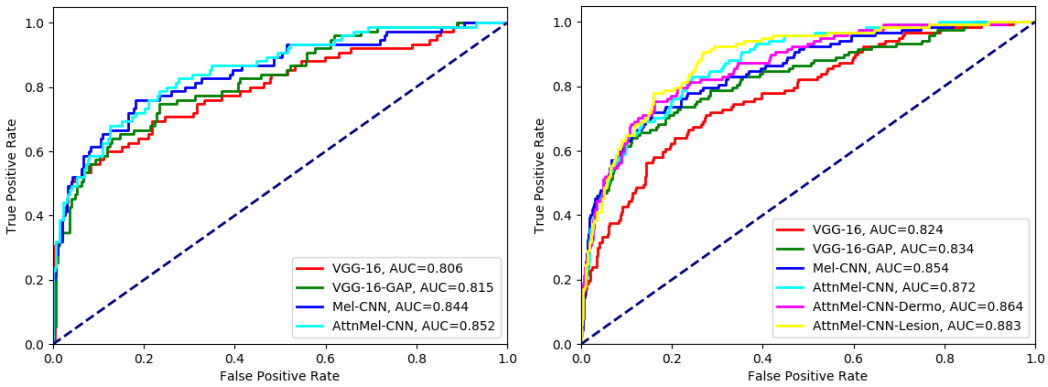


Figure 3.2: The receiver operating characteristic (ROC) curves of different models on dataset ISIC 2016 (left) and 2017 (right).

[4]For ISIC 2016 and 2017, we do file copy and save duplicated images on disk, preducing a "larger" dataset than the original one. As for ISIC 2018, file copy results in a 5 times larger dataset stored on disk (about 50k images), and this means around 1k iterations per epoch during training with a batch-size of 32. To avoid too much training burden, we do the trick of specifying a weighted sampler for the (un-duplicated) dataset. Suppose some class c takes a proportion of p in the dataset, the sampling weight for this class would be $1/p$.

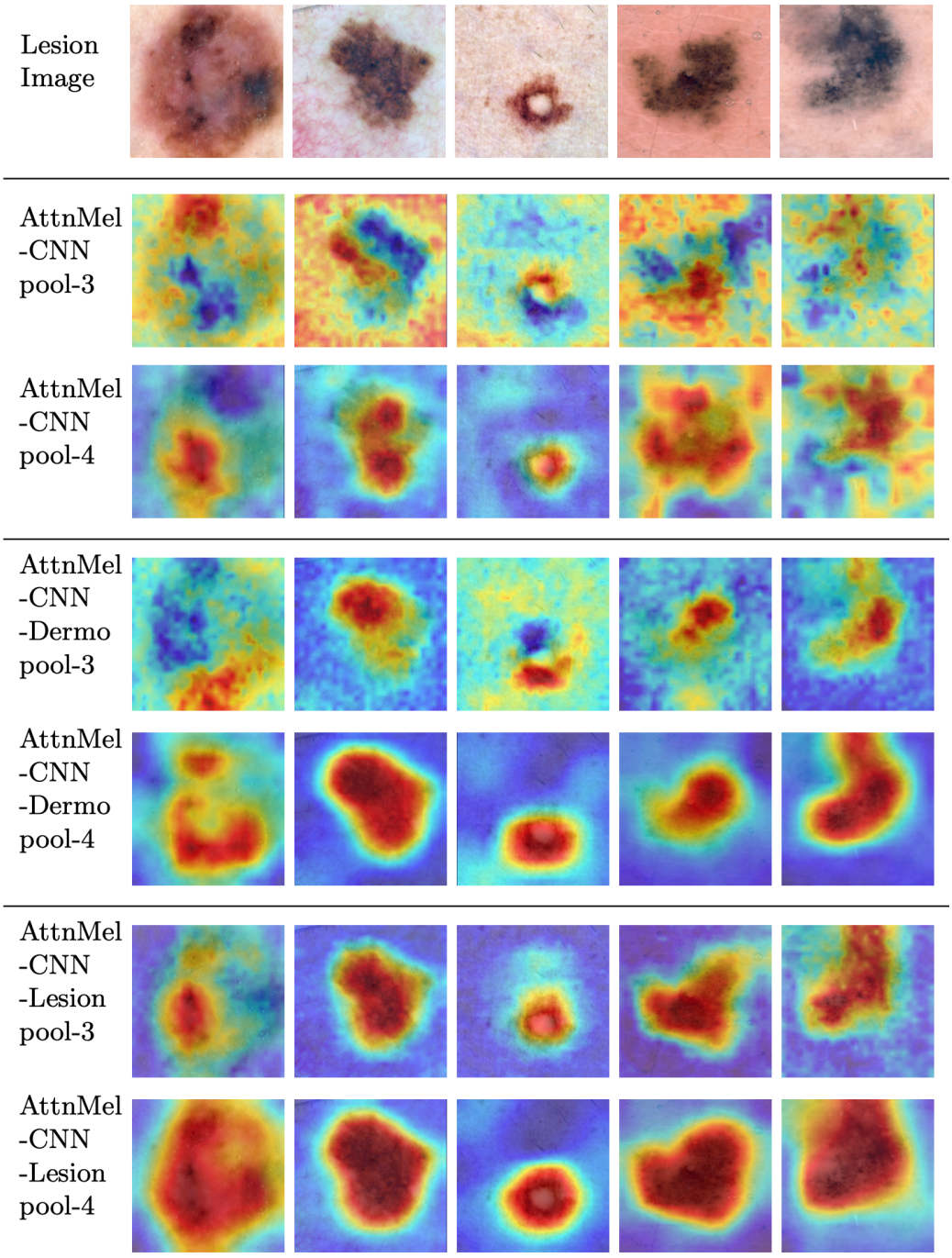


Figure 3.3: Visualization of attention maps for different models on ISIC 2017 test data. The deeper layer (pool-4) exhibits more concentrated attention to valid regions than the shallower layer (pool-3). The models with additional regularization (rows 4-7) produce more refined and semantically meaningful attention maps.

3.2.1 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* 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. The quantitative results are summarized in table 3.1 and 3.2. For a comprehensive view of the performance, we plot the ROC curves for different models in Fig. 3.2. We'll 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. 2.1), 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 3.1 (rows 4,7) and Table 3.2 (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 the two attention modules. Referring to Table 3.1 and 3.2, *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. The results also prove the hypothesis in section 2.1 that due to the simple structure of skin lesions, features from shallow layers can also contain import discriminative information.

		AP	AUC	Lesion	Interp	Ensemble
#1	Lequan et al. [52]	0.637	0.804	✓	✗	✗
#2	Codella et al. [8]	0.596	0.808	✗	✗	✓
#3	Yu et al. [53, 54]	*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**	✗	✓	✗

Table 3.1: 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.

Does attention help? After confirming the usefulness of exploiting 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 directly on pool-3 and pool-4 instead of their attention versions. We denote this baseline *Mel-CNN*. According to Table 3.1 and 3.2, 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.

How does the regularization influence the model? We re-train the network using the loss proposed in Eq. 2.7 with three different reference maps ($\bar{\boldsymbol{\mathcal{A}}}$): (1) *AttnMel-CNN-Dermo* uses the union of all dermoscopic features[5]; (2) *AttnMel-CNN-Lesion* uses the whole lesion segmentation map; and (3) *AttnMel-CNN-Bkg* uses image background (the inverse of the lesion segmentation map). Table 3.2 shows that encouraging attention to lesion or

[5]For ISIC 2017, we convert the superpixel labels to binary pixel labels in the same way as Kawahara et al. [24].

dermoscopic features yields better performance, while improper attention (*AttnMel-CNN-Bkg*) harms the performance.

		AP	AUC	Lesion	Dermo	Interp	Ensemble	External
#1	Winner 1 [29]	–	0.868	✗	✗	✗	✓	✓
#2	Winner 2 [9]	–	0.856	✓	✓	✗	✗	✓
#3	Winner 3 [31]	–	*0.874*	✗	✗	✗	✓	✓
#4	Harangi et al. [14]	–	0.836	✗	✗	✗	✓	✗
#5	Mahbod et al. [28]	–	*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	✓	✗	✓	✗	✗

Table 3.2: 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.

3.2.2 Benchmarking

We summarize previous work in Table 3.1 rows 1-3 and Table 3.2 rows 1-5. Comparison with Yang et al. [50] and Chen et al. [4] 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 attention map 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 [52, 50, 4, 9], our method is more robust and flexible in that: (1) The regularization terms can be plugged out when pixel-level annotations are not available, and our model can still work well. While these comparison methods must have segmentation masks for training. (2) The competing works can only utilize lesion segmentation, but our regularization method can efficiently use sparse maps such as dermoscopic features(*AttnMel-CNN-Dermo*).

3.2.3 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-Dermo* and *AttnMel-CNN-Lesion* on the ISIC 2017 test data by upsampling $\boldsymbol{\mathcal{A}}$ (Eq. 2.2) to align with the input image. The results are shown in Fig. 3.3. 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 3.2.

Quantitative Analysis We quantify the "quality" of the learned attention map by computing its overlap with the ground truth lesion segmentation. First, 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) [57] from *VGG-16-GAP* and follow the same procedure as above to compute the Jaccard index value. The results reported in Table 3.3 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.

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

Table 3.3: Jaccard index of (binarized) attention maps and class activation maps with respect to the ground truth lesion segmentations.

3.3 Multi-class Classification

This thesis extends our paper [49] with experiments on ISIC 2018. The test set for ISIC 2018 is kept private, and the online platform is restricted to a limited number of submissions per week. To perform enough experiments for ablation study, we do 5-fold cross-validation using the training set and report the average value of the five folds for each model. What's more, the data of lesion segmentation and dermoscopic feature detection tasks don't overlap with the classification task, which means attention map regularization (*AttnMel-CNN-Dermo* and *AttnMel-CNN-Lesion*) is not feasible. Despite this, we train a simple U-Net [36] with the data provided in the lesion segmentation task (only 2594 images) and use it to generate masks of the classification training set. Even though these masks are far from perfect, we show that our model can still take advantage of them.

3.3.1 Ablation Study

Similar to section 3.2, we train various models including *VGG-16*, *VGG-16-GAP*, *Mel-CNN* and *AttnMel-CNN*. The results are shown in table 3.4. Again, the model with attention yields the best performance. Other characteristics that are observed from the results include:

- The classification difficulty varies for different classes. There is around 40% gap in accuracy between the easiest class (VASC) and the most difficult one (DF).

- The least frequent class is not necessary the hardest one to recognize. For example, though vascular lesion (VASC) takes only 1.4% of the dataset (Fig. 3.1), it proves to be the easiest class, with almost every model achieving 100% accuracy.

		MEL	NV	BCC	AKIEC	BKL	DF	VASC	AVG
#1	VGG-16	**0.829**	0.848	0.902	0.750	0.782	0.545	**1.0**	0.808
#2	VGG-16-GAP	0.811	0.870	0.863	0.813	0.845	0.545	**1.0**	0.821
#3	Mel-CNN	0.811	0.861	0.902	**0.906**	0.827	0.545	0.929	0.826
#4	AttnMel-CNN	0.784	**0.896**	**0.941**	0.813	0.818	**0.636**	**1.0**	0.841
#5	AttnMel-CNN-Lesion*	0.801	**0.896**	0.922	0.750	**0.873**	**0.727**	**1.0**	**0.853**

Table 3.4: Quantitative results on the ISIC 2018 test set. Accuracy on each class (recall) and the average recall are recorded. The highest rankings are highlighted in **bold**. **Notation:** *MEL*: melanoma; *NV*: melanocytic nevus; *BCC*: basal cell carcinoma; *AKIEC*: actinic keratosis / Bowen's disease (intraepithelial carcinoma); *BKL*: benign keratosis (solar lentigo / seborrheic keratosis / lichen planus-like keratosis); *DF*: dermatofibroma; *VASC*: vascular lesion.

3.3.2 Attention Regularization with Imperfect Reference

Though we don't have pixel-level labels to perform the same experiments in section 3.2, we managed to "fake" lesion segmentation "groundtruth" by ourselves. We use the training data from ISIC 2018 task 1 (lesion boundary segmentation) to train a simple U-Net model. Then we use the U-Net to generate segmentation maps for the classification training set. These automatically generated prediction maps are treated as ROI reference maps ($\bar{\boldsymbol{\mathcal{A}}}$ in Eq. 2.6) to train the model *AttnMel-CNN-Lesion** (the star symbol indicates that "fake" lesion segmentation is used).

Since the size of the segmentation training set is much smaller than the classification dataset (2594 vs. 10015 images), the generated ROI maps are of low quality in terms of segmentation accuracy, though they roughly indicate the region of the lesion (Fig 3.4). According to the results shown in table 3.4, the proposed model takes advantage of the ROI maps despite that they are imperfect, leading to better overall classification performance. We visualize the attention maps in Fig. 3.5, from which we can tell that regularization yields refined attention.

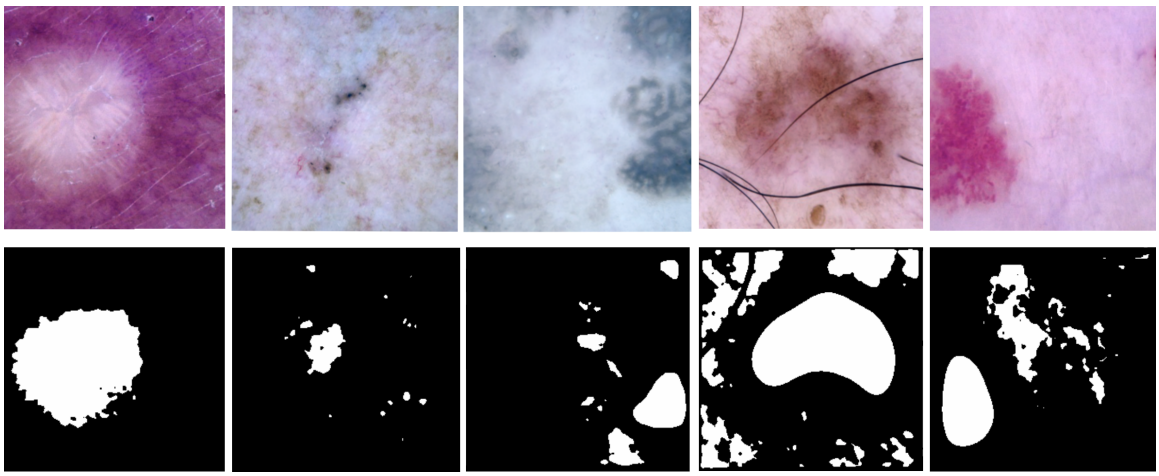


Figure 3.4: The generated ROI masks roughly highlight the lesion region but are of low quality.

3.3.3 Summary

In general, the experiments on multi-class dataset agree with the results in section 3.2.

- The plain *VGG-16* has the worst performance due to too many parameters and overfitting, and reducing parameters do help with the performance (*VGG-16-GAP*)
- Explicitly extracting some intermediate features is effective (*Mel-CNN*), but intermediate features may contain too much raw information, and attention helps via ruling out irrelevant regions (*AttnMel-CNN*).
- Attention regularization further boosts the performance, as well as refines the learned attention maps (*AttnMel-CNN-Lesion**).

What's more, we show that our model is well tolerant of imperfect ROI maps for attention regularization. As long as the maps roughly indicate the target region, the model can capture enough useful information to learn high-quality attention maps.

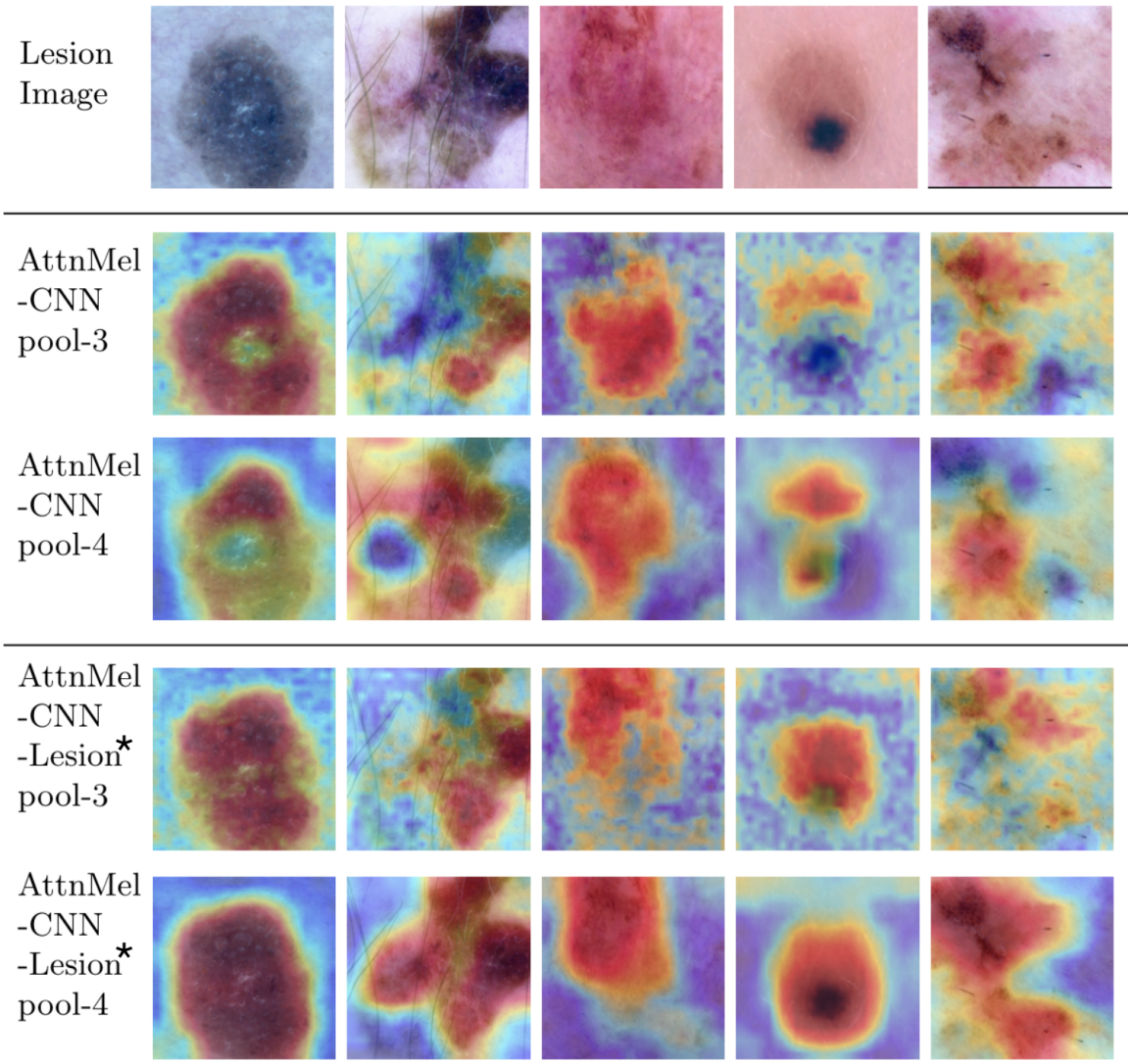


Figure 3.5: Visualization of attention maps for different models on ISIC 2018. Even though *AttnMel-CNN-Lesion** is trained with imperfect ROI maps, the attention maps are refined compared with *AttnMel-CNN*.

Chapter 4

Conclusion

This thesis proposed an attention-based deep network with attention regularization for skin cancer classification. Attention modules not only improve classification performance but also produce interpretable results. Furthermore, our novel attention regularization terms were shown effective, robust and flexible. They can be easily plugged into the loss function whenever ROI reference maps are available. The ROI maps do not necessarily need to indicate the whole lesion area, nor need they be perfect annotations. As long as ROIs give hints on the dedicated attention region, the attention regularization can take advantage of them.

Our work has witnessed extensions since published. For example, Zhao et al. [56] used estimated segmentation masks as attention priors for predicting retinopathy, wherein automatic vessel segmentation guided attention resulted in improved classification accuracy.

Chapter 5

Future Work

5.1 Further Exploration of Attention

In this thesis, ablation study was performed to show the performance improvement when the VGG-16 network gradually "evolves" into the proposed model. There is another series of ablation studies that can be done to explore the effects of attention, especially attention regularization. Specifically, we can set four "dimensions" for the "model space":

- Quantity: the proportion of the training set of images with attention prior, i.e., whether all or a subset of training images have pixel-level annotations as attention prior.
- Quality: the quality of the attention prior, i.e., whether it's annotated by an expert, a novice, or automatically generated, and the level of detail provided by the segmentation mask, e.g. precise delineation, approximating polygon, or a bounding box.
- Type: the type of the attention prior, i.e., is it lesion segmentation, dermoscopic features, or others.
- Architecture: the number of intermediate layers to apply attention and which specific layers to apply.

Some of the four "dimensions" are discrete. For example, the type of attention prior can only be one of several options, and there are finite options in terms of which layer to apply attention. On the other hand, quantity and quality are continuous dimensions. We can explore the influence when we gradually increase the fraction of images with attention

prior. It is also interesting to start from very rough attention prior (e.g., rectangles or circles around the lesions) and refine it step by step to most accurate human annotations.

In this thesis, we applied two types of priors to ISIC2017 (lesion segmentation and dermoscopic features), and used generated lesion segmentation masks as prior for ISIC2018. Nevertheless, a complete exploration over the above dimensions are to be done in the future.

5.2 Data bias

ISIC datasets contain some "divergent" samples where there are rulers or stickers in the images. There is an question to be answered: is the dataset biased w.r.t those visual outliers? For example, is it the case that all or most of the images with stickers belong to one class, so that the classifier actually learns the bias towards stickers?

Further, if the answer to the above question were "yes", another question would occur: does attention or attention prior reduce such bias? Manual examination of datasets is needed to answer these questions.

5.3 User Study

Since one of the most important products of attention is visual interpretability, one meaningful further work would be performing a user study to explore the relation between the attention maps and the regions of the image that human experts actually look at when making the diagnosis.

Bibliography

[1] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6077–6086, 2018.

[2] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.

[3] David Bau, Bolei Zhou, Aditya Khosla, Aude Oliva, and Antonio Torralba. Network dissection: Quantifying interpretability of deep visual representations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6541–6549, 2017.

[4] Sheng Chen, Zhe Wang, Jianping Shi, Bin Liu, and Nenghai Yu. A multi-task framework with feature passing module for skin lesion classification and segmentation. In *IEEE International Symposium on Biomedical Imaging*, pages 1126–1129, 2018.

[5] Noel Codella, Junjie Cai, Mani Abedini, Rahil Garnavi, Alan Halpern, and John R Smith. Deep learning, sparse coding, and svm for melanoma recognition in dermoscopy images. In *International Workshop on Machine Learning in Medical Imaging*, pages 118–126. Springer, 2015.

[6] Noel Codella, Veronica Rotemberg, Philipp Tschandl, M Emre Celebi, Stephen Dusza, David Gutman, Brian Helba, Aadi Kalloo, Konstantinos Liopyris, Michael Marchetti, et al. Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (ISIC). *arXiv preprint arXiv:1902.03368*, 2019.

[7] Noel CF Codella, David Gutman, M Emre Celebi, Brian Helba, Michael A Marchetti, Stephen W Dusza, Aadi Kalloo, Konstantinos Liopyris, Nabin Mishra, Harald Kittler, 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*, pages 168–172, 2018.

[8] Noel CF Codella, Q-B Nguyen, Sharath Pankanti, DA Gutman, Brian Helba, AC Halpern, and John R Smith. Deep learning ensembles for melanoma recognition in dermoscopy images. *IBM Journal of Research and Development*, 61(4):1–15, 2017.

[9] Iván González Díaz. Incorporating the knowledge of dermatologists to convolutional neural networks for the diagnosis of skin lesions. *arXiv preprint arXiv:1703.01976*, 2017.

[10] Andre Esteva, Brett Kuprel, Roberto A Novoa, Justin Ko, Susan M Swetter, Helen M Blau, and Sebastian Thrun. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, 542(7639):115, 2017.

[11] C Fink and HA Haenssle. Non-invasive tools for the diagnosis of cutaneous melanoma. *Skin Research and Technology*, 23(3):261–271, 2017.

[12] Zongyuan Ge, Sergey Demyanov, Rajib Chakravorty, Adrian Bowling, and Rahil Garnavi. Skin disease recognition using deep saliency features and multimodal learning of dermoscopy and clinical images. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 250–258. Springer, 2017.

[13] David Gutman, Noel CF Codella, Emre Celebi, Brian Helba, Michael Marchetti, Nabin Mishra, and Allan Halpern. 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.

[14] Balazs Harangi, Agnes Baran, and Andras Hajdu. 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*, pages 2575–2578. IEEE, 2018.

[15] Roderick J Hay, Nicole E Johns, Hywel C Williams, Ian W Bolliger, Robert P Dellavalle, David J Margolis, Robin Marks, Luigi Naldi, Martin A Weinstock, Sarah K Wulf, et al. The global burden of skin disease in 2010: an analysis of the prevalence and impact of skin conditions. *Journal of Investigative Dermatology*, 134(6):1527–1534, 2014.

[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision*, pages 1026–1034, 2015.

[17] MetaOptima Technology Inc. Molescope. `https://molescope.com`.

[18] Saeed Izadi, Zahra Mirikharaji, Jeremy Kawahara, and Ghassan Hamarneh. Generative adversarial networks to segment skin lesions. In *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*, pages 881–884. IEEE, 2018.

[19] Saumya Jetley, Nicholas A Lord, Namhoon Lee, and Philip HS Torr. Learn to pay attention. *arXiv preprint arXiv:1804.02391*, 2018.

[20] Agnieszka Kardynal and Malgorzata Olszewska. Modern non-invasive diagnostic techniques in the detection of early cutaneous melanoma. *Journal of dermatological case reports*, 8(1):1, 2014.

[21] Jeremy Kawahara, Sara Daneshvar, Giuseppe Argenziano, and Ghassan Hamarneh. Seven-point checklist and skin lesion classification using multitask multimodal neural nets. *IEEE journal of biomedical and health informatics*, 23(2):538–546, 2018.

[22] Jeremy Kawahara, Sara Daneshvar, Giuseppe Argenziano, and Ghassan Hamarneh. Seven-point checklist and skin lesion classification using multitask multimodal neural nets. *IEEE Journal of Biomedical and Health Informatics*, 23(2):538–546, 2019.

[23] Jeremy Kawahara and Ghassan Hamarneh. Multi-resolution-tract cnn with hybrid pretrained and skin-lesion trained layers. In *International workshop on machine learning in medical imaging*, pages 164–171. Springer, 2016.

[24] Jeremy Kawahara and Ghassan Hamarneh. Fully convolutional neural networks to detect clinical dermoscopic features. *IEEE journal of biomedical and health informatics*, 23(2):578–585, 2018.

[25] Harold Kittler, H Pehamberger, K Wolff, and MJTIO Binder. Diagnostic accuracy of dermoscopy. *The lancet oncology*, 3(3):159–165, 2002.

[26] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollar. Focal loss for dense object detection. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 2999–3007. IEEE, 2017.

[27] Jiasen Lu, Caiming Xiong, Devi Parikh, and Richard Socher. Knowing when to look: Adaptive attention via a visual sentinel for image captioning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 375–383, 2017.

[28] Amirreza Mahbod, Gerald Schaefer, Isabella Ellinger, Rupert Ecker, Alain Pitiot, and Chunliang Wang. Fusing fine-tuned deep features for skin lesion classification. *Computerized Medical Imaging and Graphics*, 71:19–29, 2018.

[29] Kazuhisa Matsunaga, Akira Hamada, Akane Minagawa, and Hiroshi Koga. Image classification of melanoma, nevus and seborrheic keratosis by deep neural network ensemble. *arXiv preprint arXiv:1703.03108*, 2017.

[30] Laura M McIntosh, Randy Summers, Michael Jackson, Henry H Mantsch, James R Mansfield, Marilyn Howlett, A Neil Crowson, and John WP Toole. Towards noninvasive screening of skin lesions by near-infrared spectroscopy. *Journal of Investigative Dermatology*, 116(1):175–181, 2001.

[31] Afonso Menegola, Julia Tavares, Michel Fornaciali, Lin Tzy Li, Sandra Avila, and Eduardo Valle. Recod titans at isic challenge 2017. *arXiv preprint arXiv:1703.04819*, 2017.

[32] Zahra Mirikharaji and Ghassan Hamarneh. Star shape prior in fully convolutional networks for skin lesion segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 737–745. Springer, 2018.

[33] Zahra Mirikharaji, Yiqi Yan, and Ghassan Hamarneh. Learning to segment skin lesions from noisy annotations. In *Domain Adaptation and Representation Transfer and Medical Image Learning with Less Labels and Imperfect Data*, pages 207–215. Springer, 2019.

[34] Aleksey Nozdryn-Plotnicki, Jordan Yap, and William Yolland. Ensembling convolutional neural networks for skin cancer classification. *International Skin Imaging Collaboration (ISIC) Challenge on Skin Image Analysis for Melanoma Detection. MICCAI*, 2018.

[35] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*, pages 8024–8035, 2019.

[36] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

[37] JK Schofield, D Fleming, D Grindlay, and H Williams. Skin conditions are the commonest new reason people present to general practitioners in england and wales. *British Journal of Dermatology*, 165(5):1044–1050, 2011.

[38] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 618–626, 2017.

[39] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*, 2013.

[40] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *International conference on learning representation*, 2015.

[41] Kien T Tran, Natalie A Wright, and Clay J Cockerell. Biopsy of the pigmented lesion—when and how. *Journal of the American Academy of Dermatology*, 59(5):852–871, 2008.

[42] Pieter Van Molle, Miguel De Strooper, Tim Verbelen, Bert Vankeirsbilck, Pieter Simoens, and Bart Dhoedt. Visualizing convolutional neural networks to improve decision support for skin lesion classification. In *MICCAI Workshop on Understanding and Interpreting Machine Learning in Medical Image Computing Applications*, pages 115–123. Springer, 2018.

[43] Shyamal Wahie and Clifford M Lawrence. Wound complications following diagnostic skin biopsies in dermatology inpatients. *Archives of dermatology*, 143(10):1267–1271, 2007.

[44] Fei Wang, Mengqing Jiang, Chen Qian, Shuo Yang, Cheng Li, Honggang Zhang, Xiaogang Wang, and Xiaoou Tang. Residual attention network for image classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3156–3164, 2017.

[45] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. CBAM: Convolutional block attention module. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 3–19, 2018.

[46] Junyan Wu, Xiaoxiao Li, Eric Z Chen, Hongda Jiang, Xu Dong, and Ruichen Rong. What evidence does deep learning model use to classify skin lesions? *arXiv preprint arXiv:1811.01051*, 2018.

[47] Huijuan Xu and Kate Saenko. Ask, attend and answer: Exploring question-guided spatial attention for visual question answering. In *European Conference on Computer Vision*, pages 451–466. Springer, 2016.

[48] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*, pages 2048–2057, 2015.

[49] Yiqi Yan, Jeremy Kawahara, and Ghassan Hamarneh. Melanoma recognition via visual attention. In *International Conference on Information Processing in Medical Imaging*, pages 793–804. Springer, 2019.

[50] Xulei Yang, Hangxing Li, Li Wang, Si Yong Yeo, Yi Su, and Zeng Zeng. 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)*, pages 1263–1266. IEEE, 2018.

[51] Quanzeng You, Hailin Jin, Zhaowen Wang, Chen Fang, and Jiebo Luo. Image captioning with semantic attention. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4651–4659, 2016.

[52] Lequan Yu, Hao Chen, Qi Dou, Jing Qin, and Pheng-Ann Heng. Automated melanoma recognition in dermoscopy images via very deep residual networks. *IEEE Transactions on Medical Imaging*, 36(4):994–1004, 2017.

[53] Zhen Yu, Xudong Jiang, Tianfu Wang, and Baiying Lei. Aggregating deep convolutional features for melanoma recognition in dermoscopy images. In *International Workshop on Machine Learning in Medical Imaging*, pages 238–246. Springer, 2017.

[54] Zhen Yu, Xudong Jiang, Feng Zhou, Jing Qin, Dong Ni, Siping Chen, Baiying Lei, and Tianfu Wang. Melanoma recognition in dermoscopy images via aggregated deep convolutional features. *IEEE Transactions on Biomedical Engineering*, 2018.

[55] Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *European conference on computer vision*, pages 818–833. Springer, 2014.

[56] Mengliu Zhao and Ghassan Hamarneh. Retinal image classification via vasculature-guided sequential attention. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 0–0, 2019.

[57] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2921–2929, 2016.

[58] Jiaxin Zhuang, Weipeng Li, Siyamalan Manivannan, Roy Wang, Jianguo Zhuang, Jihan Liu, Jiahui Pan, Gongfa Jiang, and Ziyu Yin. Skin lesion analysis towards melanoma detection using deep neural network ensemble. *International Skin Imaging Collaboration (ISIC) Challenge on Skin Image Analysis for Melanoma Detection. MICCAI*, 2018.