How Important Is Each Dermoscopy Image?

Catarina Barata   Carlos Santiago
Institute for Systems and Robotics
Lisbon, Portugal
ana.c.fidalgo.barata@tecnico.ulisboa.pt

Abstract

Deep neural networks (DNNs) have revolutionized the field of dermoscopy image analysis. Systems based on DNNs are able to achieve impressive diagnostic performances, even outperforming experienced dermatologists. However, DNNs strongly rely on the quantity and quality of the training data. Real world data sets, including those related to dermoscopy, are often severely imbalanced and of reduced dimensions. Thus, models trained on these data sets typically become biased and fail to generalize well to new images. Sample weighting strategies have been proposed to overcome the previous limitations with promising results. Nonetheless, they have been poorly investigated in the context of dermoscopy image analysis. This paper addresses this issue through the extensive comparison of several sample weighting methods, namely class balance and curriculum learning. The results show that each sample weighting strategy influences the performance of the model in different ways, with most finding a compromise between correctly classifying the most common classes or biasing the model towards the less represented classes. Furthermore, the features learned by each model differ significantly, depending on the training strategy.

1. Introduction

In the last few years, the field of dermoscopy image analysis has seen an impressive growth. The yearly release of larger and more challenging data sets (*e.g.*, [6, 8]) has fostered new problems, such as the diagnosis of multiple types of skin lesions (as opposed to the most common two class problem - melanoma vs benign) or the fusion of multiple imaging modalities [5]. Additionally, the emergence of deep neural networks (DNNs) facilitated the development of better diagnostic systems that do no require expert knowledge to define the most suitable image features [19, 2].

However, as the complexity of the problems increases (*e.g.*, more classes, class-novelty, or segmentation of dermoscopic structures), it becomes apparent that the amount of available data is a limiting factor in the performance of DNNs [28]. As a matter of fact, these models require large and balanced data sets to learn effectively and have been shown to underperform in long-tailed scenarios, where the distribution across classes is not uniform (class imbalance) [30]. Additionally, data sets need contain sufficient variability to allow DNNs to generalize well. Unfortunately, acquiring sufficient data may not be possible, since collecting and annotating medical data requires a huge effort both from researchers and doctors. Thus, a relevant question that this work aims to address is: **How to make the most of the available data?**

To answer this question we will explore the hypothesis that the samples in a data set are not equally relevant. Recently, there has been an increased interest in developing more efficient learning strategies for DNNs that take into account the relevance of each sample during the training process. Some methods are based on the concept of importance sampling, which modifies the batch building step in order to select some samples more often [15, 14]. The sampling frequency depends on the gradient norm associated to each sample, and makes it possible for samples to be selected more than once during one training epoch. Unfortunately, computing gradient values is computationally expensive, forcing most importance sampling methods to rely on gradients from previous epochs. This prevents the adoption of online data augmentation strategies, which are commonly used in many works. Alternative approaches rely on sample weighting schemes that assign non-uniform weights to training sample losses, modifying their importance to update the DNN [32]. Weighting approaches are easier to implement and do not increase the computational burden, thus they will be the focus of our work.

Although weighting schemes have the potential to improve the performance of DNNs in the dermoscopy field, to the best of our knowledge they have been poorly addressed [10] and a proper evaluation of their impact is needed. Moreover, several weighting strategies have been proposed in the literature, but we do not know if any of them are suitable for dermoscopy data. In this work, we address the aforementioned issues, and conduct an extensive evaluation of several sample weighting schemes. To demonstrate that these approaches can be incorporated in the training of different DNN architectures, we perform our experiments in two different models: a VGG-16 for skin lesion classification and a multi-task DNN with a VGG-16 backbone that performs a hierarchical diagnosis of the skin lesions. Additionally, we compare two popular optimizers (SGD with Nesterov and ADAM) that are commonly used in parallel with the weighting schemes. Our experiments using ISIC 2018 data set [27, 8] achieved interesting results, and we expect that this work may point to new directions of research.

2. Related Work

Nowadays, various strategies are used to deal with small and imbalanced data sets. In the following paragraphs we will discuss the ones that have been adopted in the dermoscopy field.

One of the most popular methods is data augmentation. This approach has been used to tackle imbalanced data sets, increase their variability, and improve the robustness of DNNs to acquisition conditions. Different augmentation approaches have been adopted, ranging from simple geometric and color transformations [22], to more complex formulations based on Generative Adversarial Networks [23, 4] or on the deformation of the lesion's shape [29]. Geometric and color transformations introduce small variability on the properties of the skin lesions, but increase the robustness of DNNs to acquisition changes, such as the position of the camera or lighting conditions. On the other hand, generative and deformation strategies are used to create new and unseen samples, thus increasing the representativeness of the data set. However, due to the intra-class variability of some types of skin lesions, it is not possible to guarantee that the created examples will belong to the desired class.

On par with data augmentation, another common strategy is to import a state-of-the-art architecture that was pretrained on a larger data set (*e.g.*, VGG for classification or FCN for segmentation), and fine-tune it to the dermoscopy-specific task. Nonetheless, this still requires the training from scratch of at least the last layer of the DNN, and the final performance will always depend on the quality of the data. A few works have tried to leverage their data set through the manipulation of DNN architectures. In particular, two lines of research have emerged: multi-task DNNs and network ensembles. The main goal of multi-task networks is to improve the overall performance of a DNN by enforcing it to address more than one problem, such as combining in the same model the segmentation and classification of skin lesions [34, 33]. Other examples of multi-task networks include simultaneously trying to classify the lesions and detect various dermoscopic structures [16], as well as networks that have been designed to perform a hierarchical classification of skin lesions, from the higher level grouping of melanocytic/non-melanocytic to the differential diagnosis [3]. Combining more than one task has been shown to improve the performance in the joint tasks. However, this requires the acquisition of additional information for each dermoscopy images, which may be difficult to obtain. On the other hand, ensemble networks are usually designed to perform a single task. These approaches fine-tune several state-of-the-art architectures to perform the same task and then combine their outputs [7, 21]. Nevertheless, DNN ensembles may require too much computational effort to be applied in a real-world situation, such as a hospital, where resources are limited.

A less explored direction in the dermoscopy field is sample weighting, where non-uniform weights are assigned to training samples, in order to define their relative importance during training. An example is balanced cross-entropy (BCE) [10, 35], where the weight of each sample usually captures the distribution of its class on the training set. This strategy has been used in dermoscopy to deal with the imbalance problem. However, it may not only create a bias of the network towards the less represented classes, but also disregards the possibility that samples of the same class may not be equally relevant, due to intra-class variability.

To the best of our knowledge, BCE is the only weighting strategy that has been adopted in dermoscopy. However, several other strategies have been proposed in the literature. Thus, the goal of this work is to shed some light on alternative weighting strategies and assess their impact on skin lesion diagnosis. Moreover, we want to assess if weighting strategies will lead to different results, depending on the DNN model (single vs. multi-task) and optimizer (SGD with Nesterov vs ADAM).

3. Sample Weighting Strategies

DNNs are trained using a data set of N samples, where each sample is a tuple $\{(\mathbf{x}_i, y_i)\}_{i=1}^N$. Let us assume that $\mathbf{x} \in \mathbb{R}^D$ is a (dermoscopy) image and $y \in \mathcal{Y}$ is the corresponding class label. During the training of a DNN, the goal is to estimate the set of parameters θ such that, given a sample $\mathbf{x}_i$, the label predicted by the network $\tilde{y}_i = \psi(\mathbf{x}_i|\theta)$ matches the ground truth y_i. Thus, the network parameters θ are obtained by minimizing the empirical loss between the estimated $\tilde{y}_i$ and true labels y_i

$$\theta^\star = \arg\min_\theta \frac{1}{N} \sum_{i=1}^{N} \ell\left(\psi(\mathbf{x}_i|\theta), y_i\right), \tag{1}$$

where ℓ represents the loss function. The minimization is usually performed using the stochastic gradient descent (SGD) algorithm or one of its variants (*e.g.*, ADAM [17]), with mini-batches of size $M \ll N$, leading to the update equation

$$\theta_{t+1} = \theta_t - \eta \frac{1}{M} \sum_{j=1}^{M} \nabla_{\theta_t} \ell \left(\psi(\mathbf{x}_j|\theta_t), y_j \right), \qquad (2)$$

where η represents the learning rate, t denotes the step number, and the samples $j = 1, \ldots, M$ in the mini-batch are chosen uniformly at random from the training set.

The previous update equation assumes that all samples are equally relevant. However, it is possible to modify (2) to enforce the network to pay more attention to some samples than others. In particular, weighting strategies attribute different weights $w_j^t \in \mathbb{R}^+$ to each sample in the batch, such that the update equation becomes

$$\theta_{t+1} = \theta_t - \eta \frac{1}{M} \sum_{j=1}^{M} w_j^t \nabla_{\theta_t} \ell \left(\psi(\mathbf{x}_j|\theta_t), y_j \right). \qquad (3)$$

Several approaches have been proposed to compute the weights w_j^t. One of the first methods was based on curriculum learning, which is inspired in the way humans learn: a student (the network) learns better when the curriculum (the sample weights) is chosen wisely by a teacher (the human expert). However, defining a good curriculum is a challenging task. Thus, instead of relying on human input, popular approaches define the curriculum directly from the data. To achieve this, they estimate the weights through the following joint optimization problem

$$\arg\min_{\mathbf{w},\theta} \quad \frac{1}{M} \sum_{j=1}^{M} w_j \ell \left(\psi(\mathbf{x}_j|\theta), y_j \right) + G(\mathbf{w}; \lambda) + R(\theta), \qquad (4)$$

where $R(\theta)$ encodes the regularization on the network parameters (such as weight decay), and $G(\mathbf{w}; \lambda)$ defines the curriculum (*i.e.*, the importance of each sample), based on a hyper-parameter λ.

The above equation must be solved using an alternating minimization strategy, where $\mathbf{w}$ and θ are alternatively minimized while the other is assumed to be fixed. When $\mathbf{w}$ is fixed, it is easy to see that the optimization will be the same as in (2). Different strategies have been proposed to define $G(\mathbf{w}; \lambda)$, ranging from methods that force the weights to be either 0 or 1, as in the case of self-paced learning [18], to more complex approaches where G is a trainable neural network [13].

Alternatively, the weights may be used to express the distribution of the different classes on the training set, which can be useful to deal with imbalanced data sets. In this case, the weight of a sample may be defined as the inverse class frequency [24] or a measurement of the effective number of samples [9].

From (2) it is possible to see that weighting strategies can be combined with any type of empirical loss function. In this work, we will focus on classification losses. The baseline is the **softmax cross-entropy loss (CEL)**

$$\ell \left(\psi(\mathbf{x}_j|\theta), y_j \right) = -\log p(y_j), \qquad (5)$$

where $p(y_j)$ is the probability that the DNN outputs for class y_j, assuming a softmax activation in the final classification layer.

The following subsections describe in detail the weighting strategies compared in this work.

3.1. Focal Loss

By itself, the CEL is unable to deal with imbalanced data sets and does not discriminate well between easy and hard examples. Therefore, Lin et al. [20] proposed the **focal loss (FL)**, which comprises a modulating factor to account for the different values of $p(y_j)$

$$\ell \left(\psi(\mathbf{x}_j|\theta), y_j \right) = -(1 - p(y_j))^\gamma \log p(y_j), \qquad (6)$$

where $\gamma \geq 0$ is a tunable hyperparameter. The idea of the modulating factor $(1 - p(y_j))^\gamma$ is that if $p(y_j)$ is small, then the sample is more challenging and the DNN should pay attention to it. On the other than, a $p(y_j)$ close to 1 means that the DNN chooses the true label with a high confidence and, thus, the loss for that sample will be down-weighted.

3.2. Class-balanced Losses

Both CEL and FL can be combined with different weighting strategies. The first one adopted in this work is the **class-balanced (CB)** weighting typically used in BCE [24], in which the sample weights translate the distribution of the classes in the training set, *i.e.*,

$$w_j = \frac{N}{N_{y_j}}, \qquad (7)$$

where N is the size of the training set, and N_{y_j} is the total number of samples that belong to class y_j.

Recent works have pointed out that the simple heuristic of CB may not be suitable for real world data sets. Therefore, we also compare an approach based on the **effective number of samples (ES)** per class [9]. This method assumes that the effective number of samples is an exponential function of N_{y_j} and set the weights to be to be inversely proportional to the effective number of samples for each class

$$w_j = \frac{1 - \beta}{1 - \beta^{N_{y_j}}}, \qquad (8)$$

where $\beta = (N - 1)/N$.

3.3. Curriculum Learning

The final weighting strategies evaluated are based on curriculum learning. In particular, we focus on the **self-paced learning (SPL)** [18] and **online hard example mining (OHEM)** [25] approaches.

Recalling (4), SPL sets $G(\mathbf{w}; \lambda) = -\lambda \|\mathbf{w}\|_1$, forcing the sample weights to be either zeros (for more challenging samples) or ones (for easier samples). This framework filters the hardest samples, which the network classifies incorrectly at the beginning of the training process. As the number of epochs increases and the DNN starts to learn the easiest samples, the hyper-parameter λ progressively changes, in order to allow the inclusion of a few challenging samples in the training. One of the problems of SPL is to find the best decay value for λ. In our experiments, we follow an approach similar to [11], where we start by assigning $\mathbf{w} = 1$ to the $m = M/2$ samples in the batch with lowest loss and gradually increase m as

$$m = M \frac{\exp(\alpha e)}{1 + \exp(\alpha e)}, \tag{9}$$

where e is the epoch (commonly referred to as the age of the model), with $e = 0$ being the first epoch, and α is defined so that all sample weights are set to 1 for the second half of training.

For the OHEM approach, we use the same framework described above, but focusing first on the most challenging samples. Specifically, instead of assigning $\mathbf{w} = 1$ to the m samples with lowest loss, we choose the samples with highest loss and gradually allow easier samples to be considered.

4. Experimental Setup

In this section we describe the experimental evaluation of the sample weighting strategies presented in Section 3.

4.1. DNN Architectures

We propose to evaluate the sample weighting strategies on two DNN architectures: a traditional CNN for skin lesion diagnosis, similar to the ones adopted in [12], and a hierarchical network, based on [3].

The first model uses the convolutional structure of VGG16 [26] as the backbone for feature extraction. These layers are followed by a CBAM attention module [31], which sequentially performs channel and spatial attention on the output of the last max-pooling layer. Although attention modules have been shown to improve the performance of DNNs [31], our main goal is to infer if the sample weighting strategies have some impact on the features learned by the convolutional layers. The final layer of the network is a fully connected, with the same number of neurons as the number of classes in the data set. For simplicity, we will herein refer to this architecture as *flat model*.

Table 1. Class distribution ISIC 2018 training set.

ISIC 2018 Training Set

Classes	**# Training Samples**
Nevus	6741
Melanoma	1119
Dermatofibroma	116
Actinic keratosis	331
Beningn keratosis	1101
Basal cell carcinoma	517
Vascular	143

The hierarchical model also uses a VGG16 for feature extraction and adopts a long short-term memory (LSTM) network to sequentially diagnose the dermoscopy images from the level of melanocytic/non-melanocytic until the differential diagnosis. Similarly to the flat model, this architecture comprises a CBAM attention module between the last convolutional layer and the LSTM. However, in this case the channel and spatial attention are dynamic, changing according to the hierarchic decision level. This type of attention module differs from the one used in [3], which was based only on spatial attention.

Although other CNN architectures could have been used for feature extraction, we selected VGG16 because it offers a good trade-off between representation power and computational burden.

4.2. Data Set

All of the experiments were conducted on the ISIC 2018 dermoscopy data set [8, 27]. This is a large and complex data set that contains 11,527 examples of melanocytic (melanoma and nevi) and non-melanocytic (basal cell carcinoma, actinic keratosis, benign keratosis, dermatofibroma, and vascular lesions) lesions. The non-melanocytic lesions can additionally be grouped into benign (keratosis, dermatofibroma, and vascular) and malignant (basal cell carcinoma and actinic keratosis) lesions, adding one more level to the hierarchial model on this branch. The data set is divided into training (10,015) and test sets (1,512). The training set is severely imbalanced, as shown in Table 1.

We do not augment the training set with external data and, in order to deal with image color variability induced by different acquisition setups, all of the images were normalized using the color normalization scheme proposed in [1]. We also resize all images to 299×299.

4.3. Model Training

We trained the aforementioned models end-to-end, using the sample weighting strategies described in Section 3. Additionally, we train baseline models with the CEL using (5).

Table 2. Average results for the validation splits using the **flat classification model**. All of these scores were obtained using the ADAM optimizer.

Loss		Acc	BAcc	mPR	mF1
CEL	**-**	87.5	72.5	80.0	75.5
	CB	82.0	78.8	73.0	75.0
	ES	83.5	78.0	77.5	77.0
	SPL	87.0	74.5	80.0	77.0
	OHEM	84.5	74.0	77.0	74.5
FL	**-**	87.0	73.5	79.5	75.5
	CB	81.0	78.8	73.0	75.0
	ES	82.0	78.7	74.0	75.5
	SPL	86.0	71.9	80.0	75.0
	OHEM	87.0	75.6	80.5	77.5

Table 3. Average results for the validation splits using the **hierarchical classification model**. All of these scores were obtained using the ADAM optimizer.

Loss		Acc	BAcc	mPR	mF1
CEL	**-**	87.5	75.5	80.0	77.0
	CB	84.0	78.4	76.0	76.5
	ES	84.5	76.7	77.0	76.0
	SPL	85.5	68.8	76.5	72.0
	OHEM	87.0	76.4	79.0	76.5
FL	**-**	87.0	74.5	79.0	76.0
	CB	83.0	76.9	73.0	75.0
	ES	83.5	78.0	74.0	75.5
	SPL	84.5	65.3	71.5	67.5
	OHEM	88.0	75.7	80.5	77.5

All of the experiments were performed using an NVIDIA Titan Xp.

In the case of the hierarchical model, the loss function must account for the sequential decisions. Thus, the generic form for the loss associated with sample j is the following

$$\ell\left(\psi(\mathbf{x}_j|\theta), y_j\right) = \sum_{l=1}^{L} \ell^l\left(\psi(\mathbf{x}_j|\theta), y_j^l\right), \qquad (10)$$

where $l \in \{1, \dots, L\}$ is the hierarchical level ($L = 2$ for melanocytic lesions and $L = 3$ for non-melanocytic) and y_j^l is the ground truth hierarchical class.

We compare two variants of the gradient descent method: SGD with Nesterov and ADAM, using mini-batches of size $M = 20$. The initial learning rates are 10^{-4} for SGD and 10^{-5} for ADAM, both scheduled to decay by 1/10 at 50% and 75% of the total number of epochs. This value is set to 150 for the flat model and 250 for the hierarchical.

In order to improve the generalization of the model, we have adopted the following strategies: i) online data augmentation (random crops, flips, and color transformations); ii) incorporation of dropout in several layers with a probability of 50%; and iii) initialization of the VGG16 with the weights from the model pre-trained on ImageNet.

4.4. Performance Evaluation

All of the trained models were evaluated using the following class-specific metrics: recall/sensitivity (RE), precision (PR), and F1-score (F1). Additionally, we also report the overall accuracy (Acc), as well as the overall balanced accuracy (BAcc), which corresponds to the average RE.

The performance on the real test set of ISIC 2018 can only be assessed through an online platform. Thus, we do not have access to the ground truth labels of the images on the test set. In order to have a better understanding of the impact of each sample weighting strategy, we performed two random splits of the training set into two sets of 80% for training and 20% for evaluation. Since some of the images in ISIC 2018 are repetitions of the same lesion from different viewpoints and/or time periods, we ensured that those cases appeared only in one of the sets. All of the results reported in this paper are the average of the two splits.

5. Experimental Results

5.1. Diagnostic Results

Tables 2 and 3 show the average performance for the validation splits using the ADAM optimizer, while Tables 4 and 5 show similar results, but using SGD. The following

Table 4. Average results for the validation splits using the **flat classification model**. All of these scores were obtained using the SGD optimizer.

Loss		Acc	BAcc	mPR	mF1
CEL	**-**	78.5	46.1	57.0	49.0
	CB	70.5	69.6	55.5	60.0
	ES	71.5	68.9	54.5	59.5
	SPL	78.0	45.4	55.0	48.0
	OHEM	78.0	44.0	61.0	47.0
FL	**-**	76.0	42.4	56.5	44.5
	CB	71.0	69.0	53.5	59.0
	ES	71.5	68.0	54.0	58.5
	SPL	77.0	39.0	60.0	44.0
	OHEM	78.0	46.1	59.5	50.0

Table 5. Average results for the validation splits using the **hierarchical classification model**. All of these scores were obtained using the SGD optimizer.

Loss		Acc	BAcc	mPR	mF1
CEL	**-**	80.5	51.8	56.0	52.5
	CB	75.0	74.4	60.5	65.5
	ES	77.0	73.8	62.5	66.5
	SPL	80.5	52.1	57.0	53.0
	OHEM	82.5	55.3	75.0	52.5
FL	**-**	78.0	49.0	55.0	51.0
	CB	70.5	69.8	55.0	60.0
	ES	72.5	70.2	55.5	60.5
	SPL	77.0	41.5	52.0	43.0
	OHEM	79.5	49.2	57.5	52.0

observations can be made from these tables. First, ADAM leads to significantly better scores than SGD for all models. This is in line with the results obtained in other works (*e.g.*, [21]). Another observation is that the impact of sample weighting is more noticeable with SGD, although it also influences the performance of the models trained with ADAM. Finally, the hierarchical model seems to achieve better overall performances than the flat one.

Comparing the models trained with CEL and FL, it is possible to see that the latter leads to overall worse performances. This is due to the modulating term, which may be penalizing the easy samples too much. In almost all of the experiments, the best γ for the modulating term in (6) was equal to 1, meaning that it is preferable to use the CEL.

The class balancing strategies, CB and ES, lead to an improvement in the BAcc, while reducing the Acc. This was expected, since both strategies give more importance to less represented classes, thus improving their RE at the cost of reducing the RE of the most common classes. This, in turn, will result in a lower Acc, that accounts for the proportion of samples that was correctly classified. CB and ES achieve similar performances, but the latter seems to lead to a smaller decrease on the Acc, without penalizing too much the BAcc. These results show that ES may be more suitable to deal with imbalanced dermoscopy sets. Thus, we will use this approach in the remainder of the paper.

Regarding the curriculum learning strategies, it is clear that OHEM outperforms SPL. Both of these approaches make the network focus only on a subset of samples during the first epochs of the training. OHEM provides the network with the challenging samples, while SPL presents the easiest ones. According to our results, it seems that presenting the hardest samples is a better strategy, since it improves both the Acc and BAcc of the model, while SPL lowers these values w.r.t the baseline. A possible explanation is that OHEM is making the network focus more on the less represented classes, which are harder to learn due to lack of examples.

In order to perform a detailed analysis of the impact of some weighting strategies on each class, we compare their average RE scores in Figures 1 and 2. We report the RE scores for the ADAM optimizer, but SGD shows a similar trend.

The first interesting result is that the baseline models trained with CEL and the ones trained with FL already achieve good performances on some of the less frequent classes, namely vascular and basal cell carcinoma (BCC). This is particularly visible on the hierarchical model, suggesting that imposing sequential decisions allows the model to capture better representations. More-

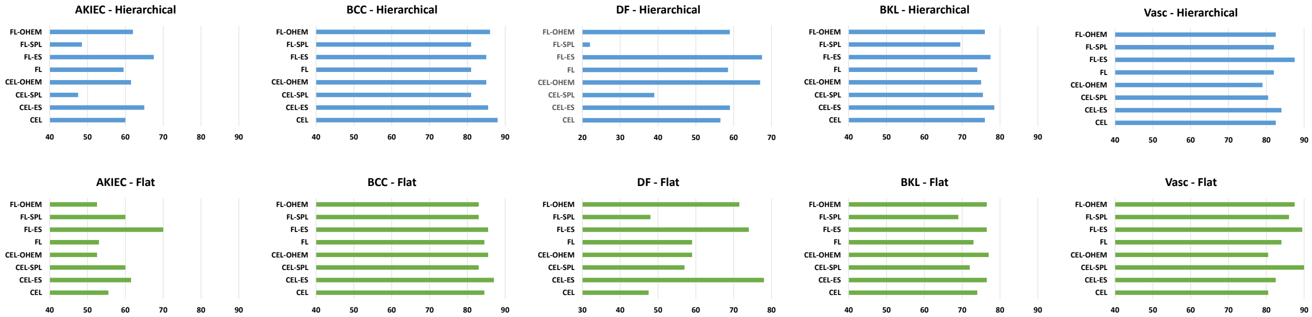

Figure 1. Average RE scores for non-melanocytic lesions, using various sample weighting strategies and the ADAM optimizer.

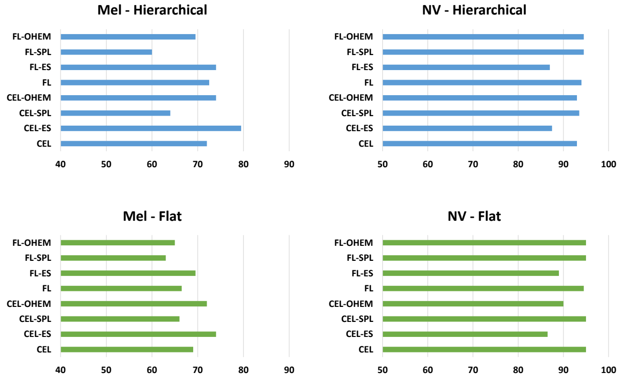

Figure 2. Average RE scores for melanocytic lesions, using various sample weighting strategies and the ADAM optimizer.

over, this approach also achieves consistently better performances for melanoma class, which is the most dangerous type of skin lesion.

When analyzing the RE values for the vascular and BCC classes, these are even better than for melanoma and keratosis, which have at least twice the number of examples, suggesting that the latter are more challenging for the models to learn. This hypothesis is validated by the results obtained with the models trained using SPL, which makes the network focus on the easiest samples. In this case, the RE of BCC is marginally degraded, while for vascular it increases (flat model) or remains the same (hierarchical). On the other hand, the RE of melanoma is severely affected. This is in line with the characteristics of melanoma, which often mimic other types of skin lesions. Actinic and dermatofibroma also seem to be very challenging.

The results for the vascular lesions (the second less represented class) suggest that class distribution does not necessarily lead to biased models. However, using the class balancing scheme, ES improves the performance of the models for the less represented classes, particularly in the case of the flat model. The drawback is that the nevus class is severely penalized, with performance drops of almost 10%. This is undesirable, since the model is becoming biased towards the less represented classes.

A good trade-off seems to be achieved when OHEM is used to determine the sample weights. It maintains the performance of easier classes (nevus, BCC, and vascular), while improving the scores of the other classes. Although these curriculum learning strategies do not explicitly deal with class imbalance, OHEM seems to be able to tackle this issue. We hypothesize that by adding some information about class distribution, such as guaranteeing that OHEM selects the hardest samples of each class, instead of selecting the most challenging samples across all classes, the results for this strategy could improve.

5.2. Analysis of CBAM Maps

The inclusion of the CBAM attention module allows us to identify the most relevant features and image regions for the models output. Moreover, the visual inspection of this information makes it possible to understand how each weighting strategy influences the learning of the convolutional features.

In Figures 3 and 4, we analyze the output of the CBAM from the hierarchical model, trained using some of the weighting strategies. In particular, we inspect the first selected channel and the spatial attention map for the last decision.

Both images clearly show that all of the sample weighting strategies significantly influence the features learned by the model. An inspection of the spatial attention highlights the variability across models, especially for the classes of melanoma, nevus, and actinic. However, the biggest differences are associated with the channels selected by CBAM. Depending on the weighting strategy, the best channel corresponds to the presence of a feature/structure, while in others it is the opposite. This is particularly visible in the case of the melanoma.

All the hard examples (Figure 3) are incorrectly classified by the model in at least one of the weighting strategies, while only BCC is incorrectly classified once in the easiest examples (for the model trained with CEL-SPL).

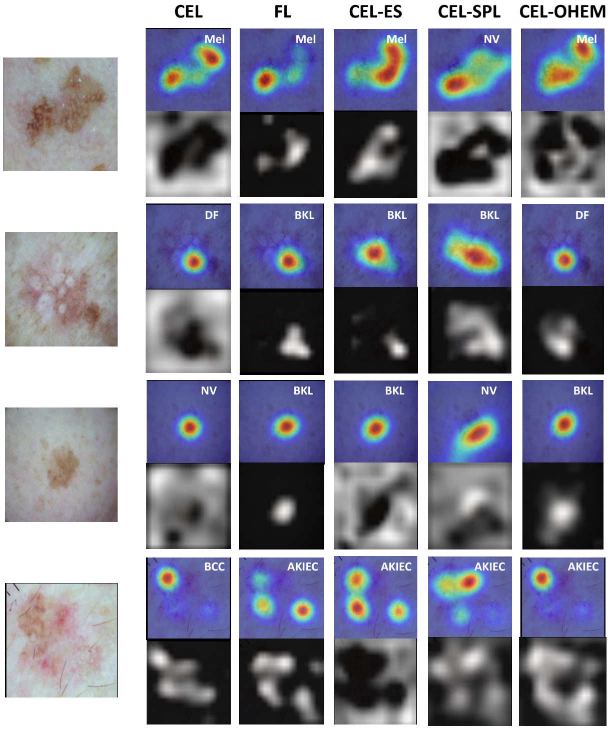

Figure 3. Output of the CBAM (spatial attention and best channel) for the hierarchical model, trained with different sample weighting schemes. These examples correspond to the most challenging classes: melanoma (top), dermatofibroma (mid-top), keratosis (mid-bottom), and actinic (bottom).

6. Conclusions and Future Work

This paper performed a comparison of several sample weighting strategies on the task of skin lesion diagnosis. In particular, we evaluated the impact of these methods on the performance of two diagnostic approaches: flat and hierarchical. Our experimental results showed that weighting methods significantly affect the performance of a classification model and may even induce bias. We also observed that the features learned by the models were highly variable, possibly caused by the sampling weighting approaches. Finally, the results suggest that OHEM achieves the best trade-off in terms of various performance metrics.

To the best of our knowledge, this is the first work that uses OHEM to train DNNs in dermoscopy data, but we believe that this approach has further potential. Thus, we plan to address the hypothesis of combining OHEM with class balancing in future work.

Acknowledgments

This work was supported by the FCT project and multi-year funding: [CEECIND/ 00326/2017] and LARSyS - FCT Plurianual funding 2020-2023.

The Titan Xp used in this project was donated by the NVIDIA Corporation.

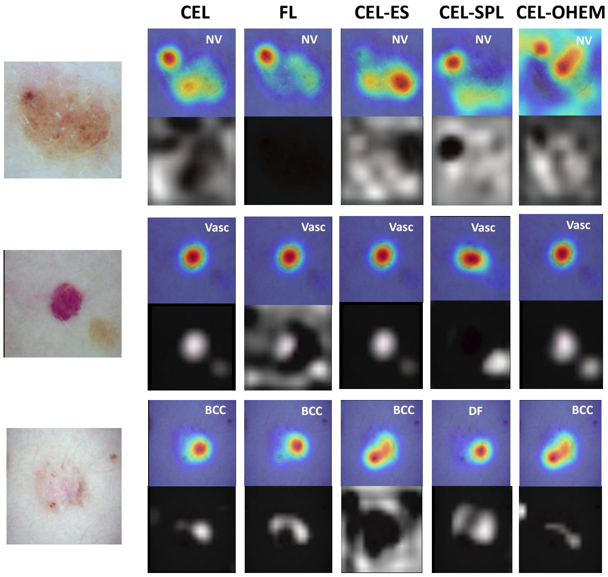

Figure 4. Output of the CBAM (spatial attention and best channel) for the hierarchical model, trained with various sample weighting schemes. These examples correspond to the easiest classes: nevus (top), vascular (mid), and actinic (bottom).

References

[1] Catarina Barata, M Emre Celebi, and Jorge S Marques. Improving dermoscopy image classification using color constancy. *IEEE journal of biomedical and health informatics*, 19(3):1146–1152, 2014.

[2] C. Barata, M. E. Celebi, and J. S. Marques. A survey of feature extraction in dermoscopy image analysis of skin cancer. *IEEE journal of biomedical and health informatics*, 23(3):1096–1109, 2018.

[3] C. Barata, J. S. Marques, and M. E. Celebi. Deep attention model for the hierarchical diagnosis of skin lesions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019.

[4] D. Bisla, A. Choromanska, R. S. Berman, J. A Stein, and D. Polsky. Towards automated melanoma detection with deep learning: Data purification and augmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019.

[5] M Emre Celebi, Noel Codella, and Allan Halpern. Dermoscopy image analysis: overview and future directions. *IEEE journal of biomedical and health informatics*, 23(2):474–478, 2019.

[6] N. Codella, D. Gutman, M. E. Celebi, B. Helba, M. A. Marchetti, S. W. Dusza, A. Kalloo, K. Liopyris, N. Mishra, H. 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 *2018 IEEE 15th Interna-*

tional Symposium on Biomedical Imaging (ISBI 2018), pages 168–172. IEEE, 2018.

[7] N. Codella, Q. Nguyen, S. Pankanti, D. A. Gutman, B. Helba, A. C. Halpern, and J. R. Smith. Deep learning ensembles for melanoma recognition in dermoscopy images. *IBM Journal of Research and Development*, 61(4/5):5–1, 2017.

[8] N. Codella, V. Rotemberg, P. Tschandl, M. E. Celebi, S. Dusza, D. Gutman, B. Helba, A. Kalloo, K. Liopyris, M. 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.

[9] Yin Cui, Menglin Jia, Tsung-Yi Lin, Yang Song, and Serge Belongie. Class-balanced loss based on effective number of samples. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9268–9277, 2019.

[10] N. Gessert, T. Sentker, F. Madesta, R. Schmitz, H. Kniep, I. Baltruschat, R. Werner, and A. Schlaefer. Skin lesion classification using cnns with patch-based attention and diagnosis-guided loss weighting. *IEEE Transactions on Biomedical Engineering*, 2019.

[11] Kamran Ghasedi, Xiaoqian Wang, Cheng Deng, and Heng Huang. Balanced self-paced learning for generative adversarial clustering network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4391–4400, 2019.

[12] B. Harangi. Skin lesion classification with ensembles of deep convolutional neural networks. *Journal of biomedical informatics*, 86:25–32, 2018.

[13] Lu Jiang, Zhengyuan Zhou, Thomas Leung, Li-Jia Li, and Li Fei-Fei. Mentornet: Learning data-driven curriculum for very deep neural networks on corrupted labels. *arXiv preprint arXiv:1712.05055*, 2017.

[14] Tyler B Johnson and Carlos Guestrin. Training deep models faster with robust, approximate importance sampling. In *Advances in Neural Information Processing Systems*, pages 7276–7286, 2018.

[15] Angelos Katharopoulos and Francois Fleuret. Not all samples are created equal: Deep learning with importance sampling. In *International Conference on Machine Learning*, pages 2530–2539, 2018.

[16] J. Kawahara, S. Daneshvar, G. Argenziano, and G. 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.

[17] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[18] M Pawan Kumar, Benjamin Packer, and Daphne Koller. Self-paced learning for latent variable models. In *Advances in Neural Information Processing Systems*, pages 1189–1197, 2010.

[19] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553):436, 2015.

[20] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017.

[21] A. Mahbod, G. Schaefer, I. Ellinger, R. Ecker, A. Pitiot, and C. Wang. Fusing fine-tuned deep features for skin lesion classification. *Computerized Medical Imaging and Graphics*, 71:19–29, 2019.

[22] F. Perez, C. Vasconcelos, S. Avila, and E. Valle. Data augmentation for skin lesion analysis. In *OR 2.0 Context-Aware Operating Theaters, Computer Assisted Robotic Endoscopy, Clinical Image-Based Procedures, and Skin Image Analysis*, pages 303–311. Springer, 2018.

[23] F. Pollastri, F. Bolelli, R. Paredes, and C. Grana. Augmenting data with gans to segment melanoma skin lesions. *Multimedia Tools and Applications*, pages 1–18, 2019.

[24] F. Provost. Machine learning from imbalanced data sets 101. In *Proceedings of the AAAI'2000 workshop on imbalanced data sets*, pages 1–3, 2000.

[25] Abhinav Shrivastava, Abhinav Gupta, and Ross Girshick. Training region-based object detectors with online hard example mining. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 761–769, 2016.

[26] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[27] P. Tschandl and *et al.* The ham10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Scientific data*, 5:180161, 2018.

[28] G. Van Horn and P. Perona. The devil is in the tails: Fine-grained classification in the wild. *arXiv preprint arXiv:1709.01450*, 2017.

[29] C. N. Vasconcelos and B. N. Vasconcelos. Experiments using deep learning for dermoscopy image analysis. *Pattern Recognition Letters*, 2017.

[30] Y. Wang, D. Ramanan, and M. Hebert. Learning to model the tail. In *Advances in Neural Information Processing Systems*, pages 7029–7039, 2017.

[31] 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.

[32] Saining Xie and Zhuowen Tu. Holistically-nested edge detection. In *Proceedings of the IEEE international conference on computer vision*, pages 1395–1403, 2015.

[33] Y. Xie, J. Zhang, Y. Xia, and C. Shen. A mutual bootstrapping model for automated skin lesion segmentation and classification. *IEEE Transactions on Medical Imaging*, 2020.

[34] J. Yang, F. Xie, H. Fan, Z. Jiang, and J. Liu. Classification for dermoscopy images using convolutional neural networks based on region average pooling. *IEEE Access*, 6:65130–65138, 2018.

[35] C. Yoon, G. Hamarneh, and R. Garbi. Generalizable feature learning in the presence of data bias and domain class imbalance with application to skin lesion classification. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 365–373. Springer, 2019.