EfficientARL: improving skin cancer diagnoses by combining lightweight attention on EfficientNet

Miguel Nehmad Alche[1], Daniel Acevedo[1,2], Marta Mejail[1,2]
[1]Universidad de Buenos Aires, Facultad de Ciencias Exactas y Naturales,
Departamento de Computación. Buenos Aires, Argentina.
[2]CONICET-UBA. Instituto de Investigación en Cs. de la Computación (ICC). Buenos Aires, Argentina.
mikealche@gmail.com, {daniel, marta}@dc.uba.ar

Abstract

Melanoma is a very dangerous form of skin cancer. Early diagnosis is crucial to increase the chances of its cure. Based on this, computer vision algorithms can be used to analyze dermoscopic images of skin lesions and decide if these correspond to benign or malignant tumors. In this work we propose the adaptation of the attention residual learning designed for ResNets to the EfficientNet networks, and compare this mechanism with attention mechanisms that these networks already have. We maintain the efficiency of these networks since only one extra parameter per stage needs to be trained. We also test several preprocessing methods on the dataset improving the final performance.

1. Introduction

Melanoma is one of the forms of skin cancer with the highest mortality rates [6]. That is why it is vitally important that proper treatment is carried out as soon as possible. Various computer vision techniques have emerged in order to detect this type of skin lesions in an early stage. The purpose of these algorithms is to be able to bring as many people as possible a reliable way to carry out periodic checks: either by distributing them in mobile applications for end users, as well as the creation of specific tools for health professionals specialized in skin treatment.

The research on classification of skin lesions has grown greatly in recent years due to the high rates achieved with deep learning techniques. Also, it has benefited from international competitions [2] which make dermoscopic image datasets with associated classification labels available to researchers.

Regarding image classification algorithms, after the success of ResNet networks [4], several improvements and variations have been tested. Among them, the EfficientNet [11] points to a good trade-off between precision and computational cost.

The attention mechanism has been applied so as to strengthen the discriminative ability of a convolutional network. Zhang *et al*. [13] proposed the addition of a small number of parameters to the ResNet that allows a simple but powerful attention mechanism with low computational cost. Another form of attention in the channel dimension has also been introduced by Hu *et al*. [5] to improve performance of convolutional networks.

Techniques that combine several models (ensemble models) were studied by Xie *et al*. [12], however this type of models usually require high computational costs that are to be avoided if a lightweight implementation is desired such as mobile apps.

In this paper we improve state of the art results on melanoma image classification. Inspired by the work of Zhang *et al*. [13], where an attention residual learning mechanism is added to the Resnet, we incorporate a similar mechanism into the EfficientNet. In line with the economy of resources posed by the EfficientNet, this mechanism uses very few parameters.

Also, it has been shown that skin lesion images benefit from color preprocessing [1]. For that, we apply preprocessing algorithms that normalize the colors of the images by applying color constancy algorithms, as well as removing variations in hue from images by applying Ben Graham preprocessing [3].

Our results show that significant improvements are achieved when both the Ben Graham preprocessing method as well as the addition of the attention residual learning mechanism to the EfficientNet networks are used. The paper is organized as follows. In Section 2 the proposed methodology is presented. Next on Section 3 we present and analyze the results that verify our hypothesis. Concluding remarks are presented in Section 4.

2. Methodology

The improvements we obtain on skin lesion classification are mainly achieved by the introduction of the attention residual mechanism on the EfficientNet along with image preprocessing. For that, in this section we describe the base methods of our proposal. First, preprocessing techniques are introduced. We then describe the attention residual mechanism on ResNets followed by the EfficientNets and how we insert this attention mechanism on them.

2.1. Preprocessing

2.1.1 Color correction and Color Constancy.

Color preprocessing methods aim to achieve greater uniformity in the images without discarding valuable and particular information about each one which facilitates the task of classification for neural networks. Color constancy methods [1] transform the colors of an image that have been captured under an unknown light source, so that the image appears to have been obtained under a canonical light source. The implementation of this transformation consists of two steps.

First, it is necessary to estimate the light source under which the image was taken, called estimated illuminant and represented by a vector $\mathbf{e} = [e_R\ e_G\ e_B]^T$. Two algorithms are implemented:

Max-RGB: This algorithm forms the estimated illuminant by selecting for each channel of an image the maximum value that appears in it by the equation $\max_{\mathbf{x}} I_c(\mathbf{x}) = ke_c$.

Shades of Gray: This method computes the estimated illuminant from the equation $\left(\frac{\int (I_c(\mathbf{x}))^p d\mathbf{x}}{\int d\mathbf{x}}\right)^{1/p} = ke_c$, where I_c represents the channel c of the image; $\mathbf{x} = (x, y)$ is the spatial position of the pixel and k is a normalization constant.

As a second step, the image colors have to be recalibrated by means of the equation $I_c^t = I_c * 1/(\sqrt{3}e_c)$ for each channel c, once the estimated illuminant vector $\mathbf{e} = [e_R\ e_G\ e_B]^t$ is computed.

In this work, the color constancy transformation is performed prior to any other data augmentation transformation.

2.1.2 Ben Graham Preprocessing.

This preprocessing comes from the winner of the diabetic retinopathy competition on the Kaggle platform [3]. It resembles the unsharp masking method for image sharpening, since it is based on obtaining the unsharp mask from an image. It can be summarized on Eq. (1) where the input image I_{in} is subtracted from its convolved version with the Gaussian kernel G (whose variance is determined automatically from the image size); then it is scaled and shifted.

$$I_{out} = 4(I_{in} - G * I_{in}) + 128 \tag{1}$$

2.2. Attention Residual Learning

In the paper by Zhang *et al*. [13] authors are able to simulate the effect of an attention layer without the extra computational cost that comes comes from the addition of significant number of new parameters. This technique is based on adding a second skip connection to the ResNet blocks, where the original input is multiplied pointwise by a Softmaxed version of the block's output.

The final output of the block is then formed by the typical output of a ResNet block (with its regular skip-connection) added to this new ARL aggregate which is controlled by a scalar α that regulates its effect. This can be seen formally in Eq. (2) where x is the input, F is the convolutional block, α is the scalar that regulates the intensity of the effect, '$\cdot$' is the point wise multiplication and finally $\mathfrak{N}$ is the softmax function applied spatially.

$$y = x + F(x) + \alpha \cdot \mathfrak{N}[F(x)] \cdot x \tag{2}$$

The softmax function $\mathfrak{N}$ is defined in Eq. (3) where O is the input, $m_{i,j}^c$ refers to the value at position (i, j) from channel c of output m.

$$\mathfrak{N}^S(O) = \left\{ m \mid m_{i,j}^c = \frac{e^{o_{i,j}^c}}{\sum_{i',j'} e^{o_{i',j'}^c}} \right\} \tag{3}$$

The classical residual block and the ARL is shown in Fig. 1. It is worth mentioning that only one parameter α is added to the training process for each ResNet block.

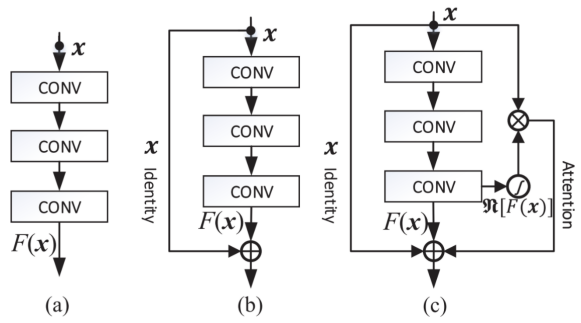

Figure 1. Comparison among different blocks used in deep neural networks. (a) classical convolutional block, (b) a block with the skip-connection as used by ResNets and (c) the block with the skip-connection and the added attention residual learning [13].

2.3. EfficientNets

EfficientNet architectures give a defined way of how to scale the architecture models when more computing power

is available. These networks are comparable to ResNets and outperform them in several tasks [11].

The decision of either choosing to add channels to the layers (width scaling), or to add layers to the model (depth scaling), or choosing to add resolution to the layers (resolution scaling) is specified by a constraint that involves taking full advantage of the computational power available. This restriction is defined by the following inequalities: (depth) $d = \alpha^\phi$, (width) $w = \beta^\phi$, (resolution) $r = \gamma^\phi$, subject to $\alpha \cdot \beta^2 \cdot \gamma^2 \approx 2$ and $\alpha \geq 1, \beta \geq 1, \gamma \geq 1$, where ϕ is a user-adjustable parameter that regulates available resources and α, β, γ are parameters determined by a small grid search. According to the convention used by the authors, a value of $d = 1$ implies 18 convolutional layers and a value of $r = 1$ implies a size of 224×224 for the images. This restriction leads to a family of networks called EfficientNet-b1 to EfficientNet-b7, depending on their capacity and complexity.

In this work we use the EfficientNet-b0 model. The choice of scaling that defines this model is done using Neural Architecture Search similar to [10] optimizing an objective function that is defined by $ACC(m) \cdot (FLOPS(m)/T)^w$ where $ACC(m)$ and $FLOPS(m)$ refer to the accuracy and the number of FLOPS of the m model respectively, T is the target of FLOPS to achieve and w is a parameter that controls the importance of the FLOPS in the optimization.

The base element of the EfficientNet-b0 is the MBConv block [7]. The first appearance of these blocks in neural networks occurs with the MobileNet architecture [7]. By means of Depthwise Separable Convolutions they reduce the amount of FLOPS required, without significantly impairing model accuracy. It is based on splitting the standard convolution layer in two: the first layer, called a *depthwise convolution*, performs a lightweight filtering by applying a single convolutional filter per input channel; the second layer is a 1×1 convolution, called a *pointwise convolution*, which is responsible for building new features through computing linear combinations of the input channels. Another significant change that MobileNet brings is the use of inverted residuals. In the bottleneck blocks of the ResNet, a channel reduction is made prior to applying the convolution operation. However, in MobileNet this dynamic is reversed: the channels are expanded prior to performing the convolution operation and then reduced again, instead of reducing the channels and then expanding them. See [7] for details.

The EfficientNet-b0 baseline network is mainly built from inverted bottlenecks MBConv. As in ResNets, this blocks have skip connections. In line with the computational savings that these networks put forward, we propose the addition of a light attention mechanism as in attention residual learning, i.e., we add the last term of Eq. (2) to each MBConv. It should be noticed that, as with ResNets, each MBconv block now only has a single extra parameter α to be trained.

The ARL attention mechanism implemented on EfficientNet-b0 can be introduced where skip-connections exist. The skip-connection only appears in the MBConv blocks which make up 16 of the 18 layers. However, the skip-connection cannot be applied in the 16 existing layers since 7 of them double the number of channels of the output with respect to the input (making it impossible to match input and output dimensions). That is why we will apply the ARL mechanism in only $16 - 7 = 9$ layers and we will then have 9 new α parameters to train. This can be clearly seen in Table 1 where the ARL column marks with the symbol '$\star$' the specific layers that have the ARL mechanism incorporated.

The EfficientNet-b0 baseline network also has an attention mechanism added to it called Squeeze & Excitation [5], which can be combined with our proposed mechanism. In the experiments we test the inclusion and removal of both mechanisms.

i	Operator $\hat{\mathcal{F}}_i$	Resol. $\hat{H}_i \times \hat{W}_i$	#Chan $\hat{C}_i$	ARL
1	Conv3x3	224×224	32	
2	MBConv1, k3x3	112×112	16	
3	MBConv6, k3x3	112×112	24	
	MBConv6, k3x3	112×112	24	$\star$
4	MBConv6, k5x5	56×56	40	
	MBConv6, k5x5	56×56	40	$\star$
5	MBConv6, k3x3	28×28	80	
	MBConv6, k3x3	28×28	80	$\star$
	MBConv6, k3x3	28×28	80	$\star$
6	MBConv6, k5x5	14×14	112	
	MBConv6, k5x5	14×14	112	$\star$
	MBConv6, k5x5	14×14	112	$\star$
7	MBConv6, k5x5	14×14	192	
	MBConv6, k5x5	14×14	192	$\star$
	MBConv6, k5x5	14×14	192	$\star$
	MBConv6, k5x5	14×14	192	$\star$
8	MBConv6, k3x3	7×7	320	
9	Conv1x1 & Pooling & FC	7×7	1280	

Table 1. EfficientNet-B0 Arquitecture. Each row describes a layer of type $\hat{\mathcal{F}}_i$, with $\langle \hat{H}_i, \hat{W}_i \rangle$ input resolution, and $\hat{C}_i$ output channels. The last column indicates with symbol '$\star$' the layers where the insertion of the ARL mechanism is possible.

3. Experiments and results

We work with ISIC's 2017 dataset [2] which consists of 2000 dermoscopic images. The images have been resized to a size of 224×224 pixels. The following data augmentation

Dataset	mean	std	25%	50%	75%	max
Baseline	0.874000	**0.006630**	0.868333	0.873333	0.878333	0.886667
Max RGB	0.876667	0.009558	0.868333	0.880000	0.885000	0.886667
Shades of Gray	0.858667	0.011244	0.853333	0.860000	0.865000	0.880000
Ben Graham	**0.887333**	0.012746	**0.881667**	**0.886667**	**0.893333**	**0.906667**

Table 2. Mean, standard deviation and quartiles accuracies for 10 runs on the dataset without any color correction, and with the 3 preprocessing described algorithms.

Dataset	mean	std	25%	50%	75%	max
Baseline	0.885389	0.014831	0.880555	0.889861	0.894861	0.905000
Max RGB	0.891306	0.011410	0.882153	0.891528	0.900695	0.908333
Shades of Gray	0.861222	0.018135	0.849931	0.859028	0.866111	0.903056
Ben Graham	**0.898667**	0.016924	**0.883681**	**0.900833**	**0.907153**	**0.928333**

Table 3. Mean, standard deviation and quartiles AUROC values for 10 runs on the dataset without any color correction, and with the 3 preprocessing algorithms.

techniques are applied as a basis: rotation of 180 degrees for both sides, zoom of up to 30% in different regions of the image and alterations in the luminosity. A batch size of 16 is used and Oversampling is used as a mechanism to solve the class imbalance.

The trainings are carried out in 2 stages. In the first stage, only the head of the model (the fully connected layers) are trained. The head has an output size of length 2 corresponding to the 2 classes that we are trying to predict: melanoma or others. This training is carried out for 4 epochs, applying a learning rate with a One Cycle [9] policy of using a maximum value of $3 \cdot 10^{-3}$. In the second stage, the entire model is trained for 20 epochs, applying the One Cycle policy with a maximum value of $3 \cdot 10^{-4}$. This process is repeated 10 times, each time with a different seed, in order to have robust data.

3.1. Impact of the preprocessing methods in the classification

In this experiment we hypothesize that the differences in luminosity on which the pictures were taken introduces noise into the classification process which hardens the task for the neural network. Therefore, we try to verify if the application of color constancy algorithms —Max RGB and Shades of Gray— in the process of homogenizing the dataset images, also facilitates the task of classification.

In turn, in addition to the color constancy algorithms, a third method is tried: Ben Graham preprocessing. This preprocessing is rather something close to a high pass filter. However, in the process of suppressing the low frequencies, it is believed that the differences introduced by the different illuminations when the images have been taken will also be attenuated.

We train 4 ResNet-50 networks which were pretrained in ImageNet. The first will serve as a baseline for comparison, while the second one will be trained on the dataset corrected with Max-RGB, the third one on the dataset corrected with Shades of Gray and the fourth on the dataset preprocessed with the Ben Graham method.

3.1.1 Results and interpretation.

Both tables 2 and 3 are formed by taking the maximum accuracy and AUROC respectively reached by each run and then calculating the mean, variance and quartiles of them. As can be seen, the training using Ben Graham's method surpasses all the other methods, by giving a maximum average accuracy between all runs of 0.887, while the training without any processing obtains 0.874.

Likewise, the Max RGB method also seems to have a positive impact on the results with average maximum accuracy between runs of 0.877, albeit to a lesser extent than the Ben Graham method. However, the Shades of Gray method seems in this case to harm performance.

It is understood that this may be due to the way each preprocessing method corrects images. In Fig. 2 it can be seen how the images processed by Shades of Gray are inclined —in some cases— towards a blue tint, while those processed by Max-RGB maintain the reddish tone that characterizes them at the same time as they amalgamate. Regarding the reason of why the Ben Graham method gives such good results, it can be hypothesized that it is because its action of filtering the low frequencies of the image allows the network to focus on the finer details of the lesions.

As a second reason, it can be seen from the comparison images that the result of applying the Ben Graham method also achieves a certain color correction effect: all images (not just some as in the Shades of Gray method) are now found leaning towards blue and brown tints. In other words, the global information on the tone of the image is homogenized together with the reduction of low frequencies.

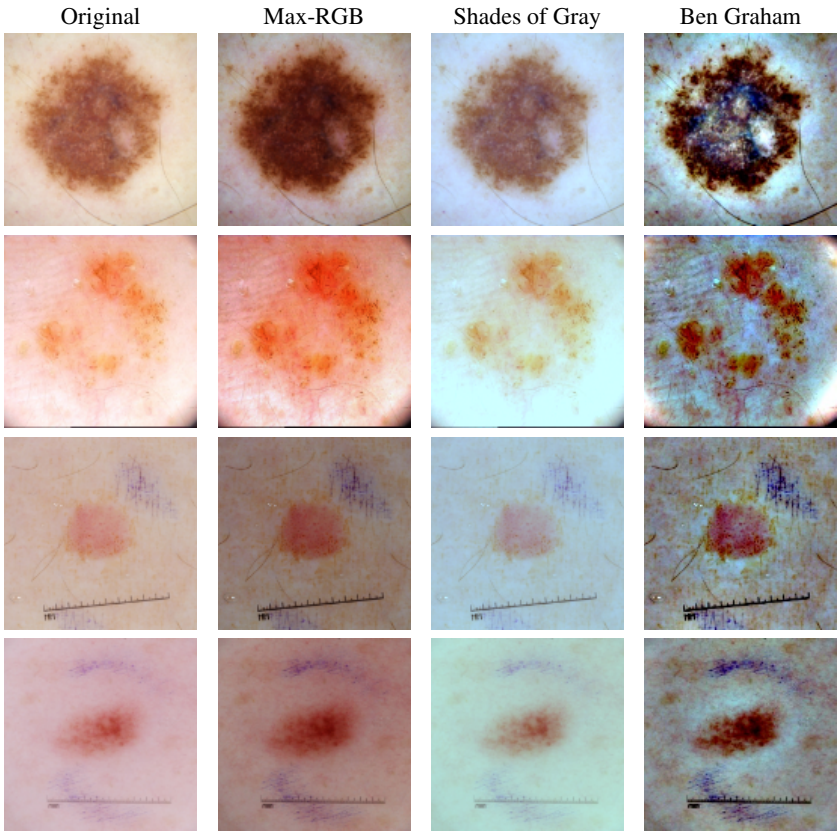

Figure 2. Comparison between different image preprocessing methods.

Dataset	mean	std	25%	50%	75%	max
Baseline(SE)	0.857333	0.007166	0.853333	0.860000	0.860000	0.866667
No attention	0.834667	0.007569	0.833333	0.833333	0.840000	0.846667
ARL	0.829333	0.012649	0.821667	0.833333	0.838333	0.846667
SE and ARL	**0.862000**	0.007730	**0.860000**	**0.863333**	**0.866667**	**0.873333**

Table 4. Accuracy results on the baseline EfficientNet-b0 and their combinations of attention mechanisms.

Dataset	mean	std	25%	50%	75%	max
Baseline(SE)	0.852667	0.009516	0.848472	0.850972	0.861042	0.865278
No attention	0.788500	0.016842	0.780347	0.794306	0.797639	0.813056
ARL	0.786250	0.012818	0.769722	0.785555	0.796875	0.804445
SE and ARL	**0.853611**	0.009903	**0.848958**	**0.855694**	**0.861250**	**0.866667**

Table 5. AUC-ROC results on the baseline and their combinations of att. mechanisms.

3.2. Attention Residual Learning on EfficientNet

In this experiment we try to study the effect of adding the Attention Residual Learning (ARL) mechanism on the Efficient-Net models. Specifically, it is not only interesting the addition of the mechanism to the base model, but we also study the way it relates to the attention mechanism already present: Squeeze & Excitation (SE). For this, we study how 4 EfficientNet variants behave (corresponding to the possible combinations of having these two attention mechanisms activated or not).

We employ the EfficientNet-b0 variant. Each of the four networks is pretrained on ImageNet. The first one will serve as a baseline for comparison, on the second one we will suppress the SE mechanism, on the third one not only we will suppress the SE but will also add the ARL mechanism and finally the fourth and last network will have both attention mechanisms activated. EfficientNet-b0 has only 9 blocks on which a skip connections is used, therefore when adding ARL we will have only 9 new α parameters to train.

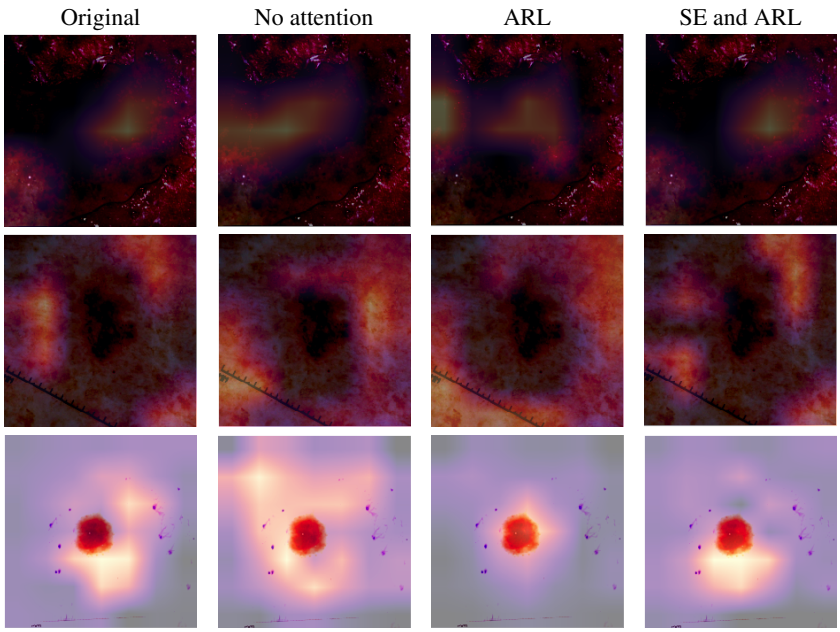

Figure 3. GradCAM heatmap comparison for models with or without ARL. In the first and second row examples, only the model with SE and ARL correctly predicted the presence of melanoma. In the example in the third row, all models correctly predicted the absence of melanoma.

3.2.1 Results and interpretation

As can be seen on Tables 4 and 5, after 10 runs with different initial seeds, the maximum average accuracy reached by the model using ARL and SE outperforms the base model. It can be seen a considerable difference between models that have SE enabled from the models that don't. This gives us the insight that a big part of EfficientNet's great performance comes from the attention mechanism rather than from the new architecture.

3.2.2 Qualitative analysis with GradCAM.

In addition to the metrics that allow a quantitative analysis, it is possible to perform a qualitative analysis of the models by observing the result of GradCAM [8]. The GradCAM technique allows viewing heat maps on images to understand which sections of the images have the greatest influence on the classification. Fig. 3 shows a comparison between the heat maps generated by the original model versus those generated by the model with ARL.

Looking at Fig. 3 the third row is the one that seems to expose the differences most clearly. The introduction of ARL appears to significantly reduce the area of the image that the network considers relevant for classification. That is, one can visually see that the attention mechanism is working properly. This is especially noticeable in the model that has ARL as the only attention method (EfficientNet model from which the SE mechanism is removed), although it is also observed in a more subtle way in the model that it has both ARL and SE mechanisms.

A second striking aspect to notice appears in the example in the second row where one can see how all the models seem to focus on the area surrounding the injury rather than the injury itself. This would be an indication that the tissue surrounding the lesion also provides valuable information. Particularly in the original image corresponding to the example of the second row, the surrounding tissue is covered with red marks, which does not appear to be information that can be ruled out.

4. Conclusions

In the present work, various mechanisms have been studied to improve the performance of the classification of skin lesions using convolutional neural networks.

Regarding the experiments on dataset preprocessing in section 3.1 it has been found that Ben Graham's preprocessing notably improves the accuracy of the classification. This is not the case with the Max RGB and Shades of Gray color correction algorithms, which give no considerable increases or even inferior results respectively to the original control dataset.

Regarding the experiment in section 3.2 which studies the impact of introducing attention mechanisms in networks —particularly the ARL mechanism— it has been found that the introduction of ARL improves the accuracy in the models of the EfficientNet family.

References

[1] C. Barata, J. S. Marques, and M. E. Celebi. Improving dermoscopy image analysis using color constancy. In *2014 IEEE Int. Conf. on Image Processing (ICIP)*, pages 3527–3531, 2014.

[2] Noel C. F. Codella, M. Emre Celebi, Kristin Dana, David Gutman, Brian Helba, Harald Kittler, Philipp Tschandl, Allan Halpern, Veronica Rotemberg, Josep Malvehy, and Marc Combalia. Int. Skin Imaging Collaboration (ISIC) Challenge: using dermoscopic image context to diagnose melanoma, 2020.

[3] Benjamin Graham. Kaggle diabetic retinopathy detection competition report., 2015.

[4] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *2016 IEEE Conf. on Comp. Vision and Pat. Recog. (CVPR)*, pages 770–778, 2016.

[5] J. Hu, L. Shen, and G. Sun. Squeeze-and-excitation networks. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7132–7141, 2018.

[6] Dora Loria, Abel González, and Clara Latorre. Epidemiología del melanoma cutáneo en argentina: análisis del registro argentino de melanoma cutáneo. 16, 03 2010.

[7] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L. Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *2018 IEEE/CVF Conf. on Comp. Vision and Pat. Recog.*, pages 4510–4520, 2018.

[8] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *2017 IEEE Int. Conf. on Comp. Vision (ICCV)*, pages 618–626, 2017.

[9] Leslie N. Smith. A disciplined approach to neural network hyper-parameters: Part 1 – learning rate, batch size, momentum, and weight decay, 2018.

[10] M. Tan, B. Chen, R. Pang, V. Vasudevan, M. Sandler, A. Howard, and Q. V. Le. Mnasnet: Platform-aware neural architecture search for mobile. In *2019 IEEE/CVF Conf. on Comp. Vision and Pat. Recog. (CVPR)*, pages 2815–2823, 2019.

[11] Mingxing Tan and Quoc Le. EfficientNet: Rethinking model scaling for convolutional neural networks. In *Proc. Int. Conf. on Machine Learning*, volume 97, pages 6105–6114, 2019.

[12] F. Xie, H. Fan, Y. Li, Z. Jiang, R. Meng, and A. Bovik. Melanoma classification on dermoscopy images using a neural network ensemble model. *IEEE Trans. on Medical Imaging*, 36(3):849–858, 2017.

[13] J. Zhang, Y. Xie, Y. Xia, and C. Shen. Attention residual learning for skin lesion classification. *IEEE Trans. on Medical Imaging*, 38(9):2092–2103, 2019.