Melanoma Detection Using Deep Learning

Florent Favole[1(✉)], Maria Trocan[1], and Ercüment Yilmaz[2]

[1] Institut Supérieur d'Électronique de Paris, Paris, France
{florent.favole,maria.trocan}@isep.fr
[2] Karadeniz Technical University, Trabzon, Turkey
ercument@ktu.edu.tr

Abstract. In this paper, we describe a region of interest-based approach for the classification of dermoscopic images of skin lesions, which nowadays contributes to early identification of skin melanoma. Once the region of interest detected, it will be further processed in order to be used for training and hence classification using deep learning methods. The main goal is to compare three different convolutional neural networks (CNNs) models and determine the one which provides the best accuracy, knowing that only salient parts of the skin lesions images have been used for training.

Keywords: Image classification · Melanoma detection · Convolutional neural network

1 Introduction

Early detection of skin cancer is vital for patients. Differential diagnosis of skin lesions, especially malignant and benign melanoma is a challenging task even for specialist dermatologists. The diagnostic performance of melanoma has significantly improved with the use of images obtained via dermoscopy devices. With the recent advances in medical image processing field, it is possible to improve the dermatological diagnostic performance by using computer-assisted diagnostic systems. For this purpose, various machine learning algorithms are designed and tested to be used in the diagnosis of melanoma [7]. Deep learning models, which have gained popularity in recent years, have been effective in solving image recognition and classification problems. Concurrently with these developments, studies on the classification of dermoscopic images using CNN models are being performed.

In this study, the performance of AlexNet, Inception-V1 (a.k.a. GoogLeNet) and Resnet50 CNNs will be examined for the classification problem of skin lesions, especially benign, malignant and unknown melanoma cancers on dermoscopic images. Dermoscopic images of 23 906 lesions obtained from ISIC database

Dr. E. Yilmaz's contribution was supported by The Scientific and Technological Research Council of Turkey (TUBITAK) under Grant 1059B191802000.

N. T. Nguyen et al. (Eds.): ICCCI 2020, LNAI 12496, pp. 816–824, 2020.
https://doi.org/10.1007/978-3-030-63007-2_64

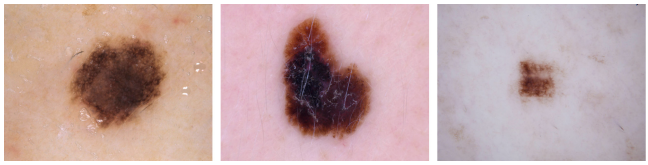

Fig. 1. Benign (left), malignant (center), unknown (right) melanoma

will be used in the experiments. After comparison of these 3 methods the one which will achieve the best accuracy will be determined. The main goal of this paper is to determine the most successful architecture having the best accuracy, to classify the images in three classes : benign, malignant and unknown.

2 Proposed Method

In the sequel, prior of introducing our images pre-processing and data augmentation methods, we describe the complexity of the dermoscopic database and introduce the deep-learning architectures used within this paper.

2.1 Datasets

Table 1. Datasets used in our approach

Dataset name	Number of images	Resolution of images
HAM10000	10015	600 × 450 pixels
SONIC	9251	3024 × 2016 pixels
MSK	3918	Varying
UDA	617	Varying
2018 JID Editorial	100	Varying

The main datasets used in our classification are provided by the ISIC archive database[1], composed by these datasets : HAM10000, SONIC, MSK, UDA and 2018 JID Editorial Images.

The HAM10000 dataset [5] is made up of 10 015 images, each one have a size of 600 × 450 pixels.

The SONIC dataset is made up of 9251 images, each one have a size of 3024 × 2016 pixels.

[1] ISIC Archive: ISIC Archive Database, https://www.isic-archive.com/#!/topWithHeader/onlyHeaderTop/gallery.

The MSK dataset [6] is composed of subdatasets with different image resolution size, in range 1024 × 720 pixels to 6600 × 4400 pixels. The dataset has approximately 3900 images.

Finally we have two others datasets (UDA and 2018 JID Editorial images) which have 700 images with several resolution image sizes.

We can already induce that the datasets which possess high resolution images will not be fed directly to the CNN models without apply a pre-processing on these images.

2.2 Considered Deep Learning Models

With 60 million parameters, AlexNet has 8 layers — 5 convolutional and 3 fully-connected. At the point of publication, the authors pointed out that their architecture was "one of the largest convolutional neural networks to date on the subsets of ImageNet". They were the first to implement Rectified Linear Units (ReLUs) as activation functions [2].

Inception-V1 (a.k.a. GoogLeNet) has 22 layers architecture with 5 million parameters. The Network In Network approach is heavily used, as mentioned in the paper [3]. This is done by means of 'Inception modules'. The design of an architecture with Inception modules is a product of research on approximating sparse structures. This design was novel by his building networks, using dense modules/blocks, instead of stacking convolutional layers, stacking modules or blocks which are convolutional layers.

ResNet50 with his 26 million parameters and 50 layers was consequently design for this problem, using skip connections (a.k.a. shortcut connections, residuals), while building deeper models. ResNet50 is one of the first adopters of batch normalisation [4].

Table 2. CNN models used in our approach

Model	Parameters	Layers	Year of publication
AlexNet	60 M	8	2012
Inception-V1	5 M	22	2014
RestNet50	26 M	50	2015

2.3 Dermoscopic Image Pre-processing

Each image undergoes the following pre-processing [1] before being fed into the classification models. Indeed the datasets MSK, SONIC and UDA, have images with several high resolutions (for example MSK-1 has 1000 images with 6400 × 4400 pixels resolution). Furthermore the models need images with the same resolutions. Therefore at the end of the pre-processing the images will be all to the resolution 256 × 256 pixels.

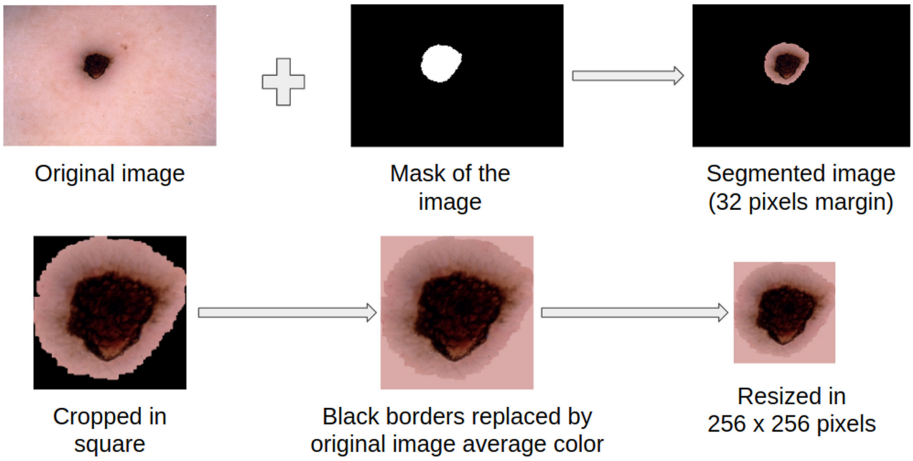

Fig. 2. The pre-processing procedure.

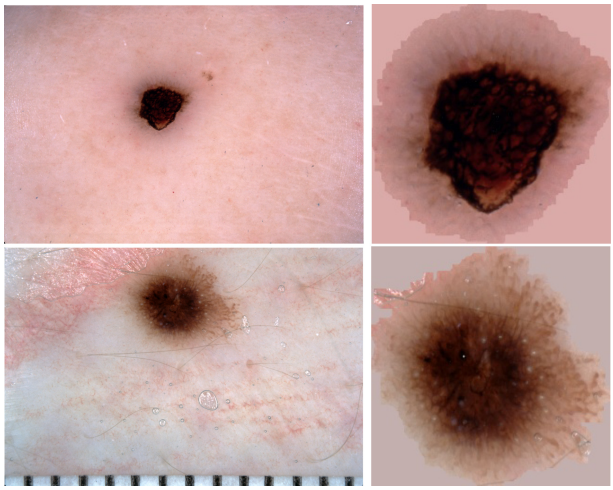

Fig. 3. Before (left) and after (right) pre-processing

Since the main ISIC archive database provides segmentation for these datasets, we started firstly by apply a skin lesion segmentation with a margin of 32 pixels.

The dataset HAM10000 will not have segmentation application, in place just a resizing to 256 × 256 pixels, because the images are already enough usable for classification.

Based on the segmentation image we remove, in rectangular format, the black borders until the remaining colored pixels. Then we keep the highest dimension to be able to crop in square format. Using this format allows to keep the original ratio of the image and consequently to loss the less information possible.

After the cropping we have an image which has still black pixels which originally has not, so to not misinform the models, we replace these black borders by the original image average color.

At the end, the image is resized to 256 × 256 pixels format. The results of this pre-processing can be see in the Fig. 4.

The images are then sorted following their classes (benign, malignant, unknown) and divided in training set 80% or validation set 20% randomly.

2.4 Data Augmentation

Since the data is extremely unbalanced, with 80% benign, 10% malignant and 10% unknown images, this can lead the models to predict always the same class result. To solve this problem we use the oversampling data balancing technique[2] on the training and validation sets by multiplying the data with data augmentation on the classes malignant and unknown.

The standard data augmentations used, as you can see in the Fig. 4, are: random zoom, brightness, vertical, horizontal, channel shifts, vertical and horizontal flips. At the end of the data augmentation we obtain approximately 33% division for each class.

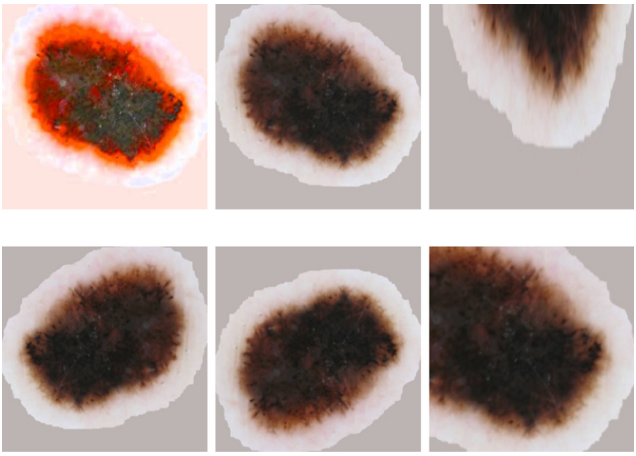

Fig. 4. Data augmentation examples.

[2] S. Chatterjee: Deep learning unbalanced training data? solve it like this. https://towardsdatascience.com/deep-learning-unbalanced-training-data-solve-it-like-this-6c528e9efea6.

3 Classification Performance Evaluation

Training a model simply means learning (determining) good values for all the weights and the bias from labeled examples. In supervised learning, a machine learning algorithm builds a model by examining many examples and attempting to find a model that minimizes loss; this process is called empirical risk minimization. Loss is the penalty for a bad prediction, namely loss is a number indicating how bad the model's prediction was on a single example. If the model's prediction is perfect, the loss is zero; otherwise, the loss is greater. The goal of training a model is to find a set of weights and biases that have low loss, on average, across all the training set[3].

In our case the models were train on 10 000 steps. Contrary to the approach in [7], and above mentioned, we consider only the saliency generated region of interest for training our data.

The value global step present on the x axis of the Fig. 5 and 6 is the actualization of the weights of the model after computed the batch size (of 32 images in our case). The loss graph of the three models is showed in the Fig. 5.

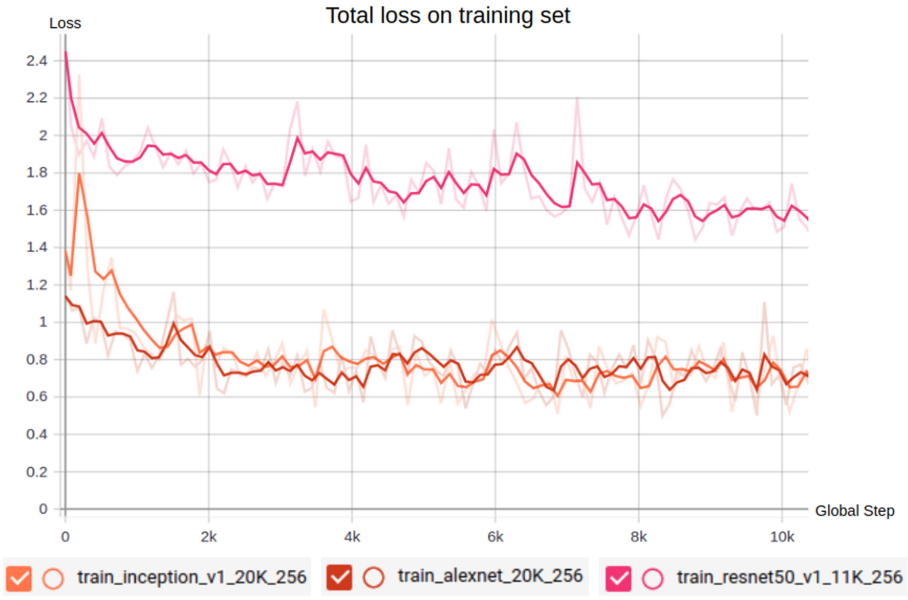

Fig. 5. Graph of the total loss for the 3 models.

On the Fig. 5, the total loss of ResNet50 model is superior by 0.9 points than Inception-V1 and AlexNet models. These differences are due to its architecture,

[3] Google: Descending into machine learning:Training and loss, https://developers.google.com/machine-learning/crash-course/descending-into-ml/training-and-loss.

containing two times more layers than the two others models. Moreover we can notice than the loss of the three models are still decreasing, consequently the models are still currently learning, and they are not overfitting, so we could have done the training on more steps. However for time issues the training was done only on 10 000 steps.

Accuracy is one metric for evaluating classification models. Informally, accuracy is the fraction of predictions our models got right. Formally for binary classification, accuracy has the following definition[4]:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

where: TP = True Positives, TN = True Negatives, FP = False Positives, and FN = False Negatives.

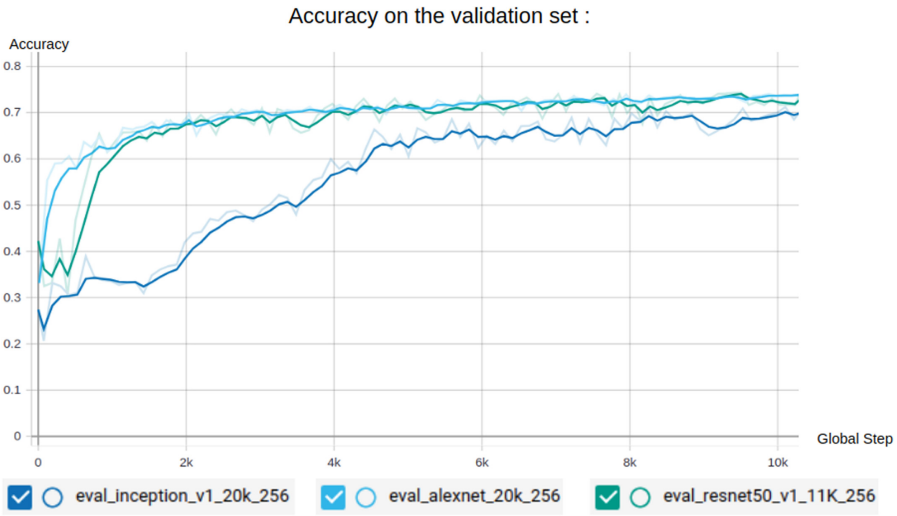

Fig. 6. Graph of the accuracy for the 3 models.

On the Fig. 6, the accuracy is in consonant (no presence of overfitting) with the total loss graph of the Fig. 5, namely the accuracy of the three models is not decreasing and it's still improving. AlexNet and ResNet50 are the models which gets the best final accuracy (0.74), Inception-V1 has the worst accuracy (0.70) but it's the fastest (in term of seconds) model to train and evaluate.

[4] Google: Classification: Accuracy, https://developers.google.com/machine-learning/crash-course/classification/accuracy.

Table 3. Final accuracy - no overfitting

Model	Accuracy	Training Time
AlexNet	0.74	55 min 36 s
Inception-V1	0.70	46 min 40 s
RestNet50	0.74	72 min 12 s

4 Conclusion

In this paper, an image region of interest based approach has been used on three deep learning architectures - both for training and validation- in order to evaluate their accuracy in melanoma detection. The networks ResNet50 and AlexNet have a similar and superior accuracy than Inception-V1. Nevertheless RestNet50 takes approximately double time of training than AlexNet. Therefore AlexNet seems to be the best CNN model for this skin lesion classification problem. However, some further improvements are possible, given the reported results. One can be the use of higher resolution for the images. We resized the images to (256, 256, 3). Using a higher resolution like (512, 512, 3) would improve the accuracy of the model.The different datasets have plenty of corrupted images therefore before to apply the pre-processing, cleaning the datasets by removing these might improve the training phase. Moreover, for the use architectures, the most sensitive layers are the last ones, so in freezing the first layers we can gain time in training. This will significantly decrease the computational load. For ISIC database, a model prediction for the metadata can be concatenated with the CNN models, resulting theoretically in higher accuracy.

References

1. Dat, T., et al.: Ensembled skin cancer classification (ISIC 2019 challenge submission)(2019)
2. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Advances in Neural Information Processing Systems, pp. 1097–1105 (2012)
3. Szegedy, C., et al.: Going deeper with convolutions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1–9 (2015)
4. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)
5. Tschandl, P., Rosendahl, C., Kittler, H.: The ham10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. Sci. Data **5**, 180161 (2018)

6. Codella, N.C., et al.: Skin lesion analysis toward melanoma detection: a challenge at the 2017 international symposium on biomedical imaging (ISBI), hosted by the international skin imaging collaboration (ISIC). In: 2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018), pp. 168–172. IEEE (2018)
7. Yilmaz, E., Trocan, M.: Benign and malignant skin lesion classification comparison for three deep-learning architectures. In: Nguyen, N.T., Jearanaitanakij, K., Selamat, A., Trawiński, B., Chittayasothorn, S. (eds.) ACIIDS 2020. LNCS (LNAI), vol. 12033, pp. 514–524. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-41964-6_44