Deep Learning and Convolutional Neural Networks in the Aid of the Classification of Melanoma

Felipe Moure Cícero, Ary Henrique M Oliveira, Glenda Michele Botelho
Curso de Ciência da Computação
Universidade Federal do Tocantins (UFT)
E-mails: felipecicero@outlook.com aryhenrique@uft.edu.br glendabotelho@uft.edu.br

***Abstract*—Pattern recognition in digital images is a major limitation in machine learning area. But, in recent years, deep learning has rapidly been diffused, providing large advancements in visual computing by solving the main problems that machine learning imposes. Based on these advances, this study aims to improve results of a problem well-known by visual computing, the classification of melanoma, this one is classified as a malignant tumor, highly invasive and easily confused with other skin diseases. To achieve this, we use some techniques of deep learning to try to get better results in the task of classifying whether a melanotic lesion is the malignant type (melanoma) or not (nevus). In this work we present a training approach using a custom dataset of skin diseases, transfer learning, convolutional neural networks and data augmentation of the deep network *ResNet* (Deep Residual Network).**

***Keywords*-deep learning; convolutional neural networks; melanoma classification;**

I. Introduction

The performance upgrade in deep learning provided by the parallelization in CUDA GPUs [1] [2] allowed the use of a larger number of examples for the training, which consequently provided a large gain in the generalization error of the models [3]. Today, such models have been applied in various practical areas, such as robotics, autonomous automobiles and in many areas of medicine [4] [5].

Various deep learning models were created to perform different types of tasks. In the case of visual computing, the algorithms were created based on neuroscience with the intention of approximating the function and behavior of the human visual cortex. In these models, many layers of neurons fulfill different levels of abstraction of minimal image details, even the most complete and generic ones, combining characteristics of previous levels with the new ones. These are called Convolutional Neural Networks (CNNs) [6]. Recent research applying these algorithms of deep learning in medical areas has produced positive feedback [4] [5].

Based on these results, this work will address this new generation of artificial neural networks and techniques that compose the scope, such as convolution networks applied to the dermatology medical area, transfer learning and data augmentation, by proposing a model of group classification of 24 dermatological diseases with the intention of determining whether a skin wound is a Melanoma (malignant melanotic wound) or is a Nevus (non-malignant melanotic wound).

II. Technical background

The key concept about Convolutional Neural Networks is that is not necessary anymore to select feature vectors to be analysed by the network, a CNN automatically selects the more important characteristics of the object. This technique represents a great step and maybe a first glimpse of a generic type of machine intelligence. The Transfer Learning technique is evidence of this generalist characteristic of CNNs [7]. Beeing the CNNs a key concept about this work and all the remnant of the literature, state of the art and knowledge required for a full understanding of the concepts about Deep Learning were taken from Goodfellow et al [6].

Assuming that CNNs currently are the best tool to recognize patterns in digital images, we can also assume that the CNNs will perform well in Melanoma classification [6]. According the theory about melanoma, this is a malignant tumor and highly invasive, which can easily reach the blood stream and cause metastasis in the patient. For Azulay [8], be able to distinguish non-malignant melanotic lesions (Nevus) of malignant melanotic lesions (Melanoma) is one of the biggest challenges of Dermatology. Although both are diseases that share morphological characteristics and belongs to the same class of diseases, Melanoma can also take on features that are not common to their morphology and can be confused with completely different diseases of the melanoma itself.

Besides the CNNs, the four main works that motivated the development of this work are: The first one is Botelho's work [9], work on which was used two neurons of the perceptron type to classify digital images as lesions malignant melanotic or not and using the gray histogram of the images as input, it was also proposed as future work an analysis of other characteristic vector types as well as probabilistic estimate of melanoma classification.; The second is Alex Krisvesk's work [1] about the use of GPUs to optimize the training of a deep convolutional neural network, CUDA's support in the optimization of matrix operations provided a large increase in performance, which allowed a greater number of images could be used in training their model; The third is the Esteva's work [4], where a deep network with Transfer Learning used a dataset customized by the authors, which succeeded in obtaining a good result in the classification of 23 classes of skin diseases with 90% of accuracy on a Melanoma classification compared to Benign Tumors; The fourth work

is the proposed deep convolutional neural network, called Deep Residual Networks (ResNet) by Microsoft's South Korean team [10]. The ResNet, to date of this work, has the best performance currently in the classification of objects in ImageNet dataset.

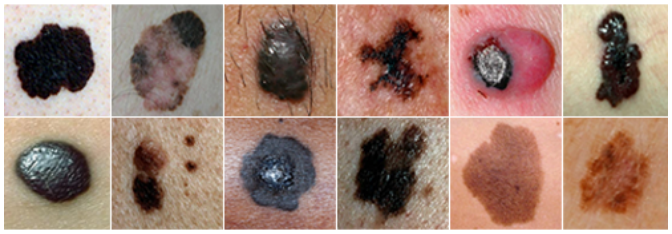


Fig. 1. Examples randomly chosen from the dataset created, the first line has some Melanoma examples and the second line has some Nevus examples.

III. Methodology

In this section we present the techniques used and adapted for the classification of non malignant melanotic lesions called Nevus, from the prejudicial ones called Melanomas. For this we create a dataset of images using digital atlas available on the Internet, we classify this set of images according to the 24 classes proposals. Then we transform the image dataset with rotations to make deep network invariant to rotations, thereby improving the generalization error. To evaluate the model we use the deep learning framework Caffe [11] and a deep network (ResNet) as features extractor, that is, the deep network already trained and converged to ImageNet dataset, and so we changed the last three network layers to train them in the classification of our dataset.

A. Organization and Distribution of the Dataset

It is the most important part and one of the largest deep learning problems. There are no open source datasets available for most areas of research that can help improve medical diagnosis. Due to the the non existence of a specific dataset for skin deseases as we have for example in the ImageNet [7], it was necessary to build an image dataset, which followed the classification proposed by *DermNet* [12] in which the dermatological pathologies are classified in 23 disease classes. For this study, we used 24 classes, as to separate the malignant melanotic wounds (Melanomas) from the non malignant melanotic (Nevus), because in the morphological classification both of these are part of the same class. Fig. 1 presents some examples of these images. The binary classification was discarded because this classification with 24 classes get full advantage of the ability of CNNs, which require a very large number of training examples and can classify up to 1000 distinct classes. In addition to providing the ability to evaluate and classify any dermatological disease in future work.

All the 27963 images acquired for the creation of the dataset used in this work was download from various dermatological digital atlas available on *DermWeb* [13], which consists in a list of sites facing dermatology professional education, including a relative number of good examples. We only downloaded images have a good quality, colored, representing directly a skin wound. It is emphasized that the cleaning and organization work of the dataset is hard and there are still some examples considered bad, or that do not show the actual wound, do not show the details of the wound, or show parts of the body that might confuse the neural network, that is, the network becomes detect the body part instead of the actual wound.

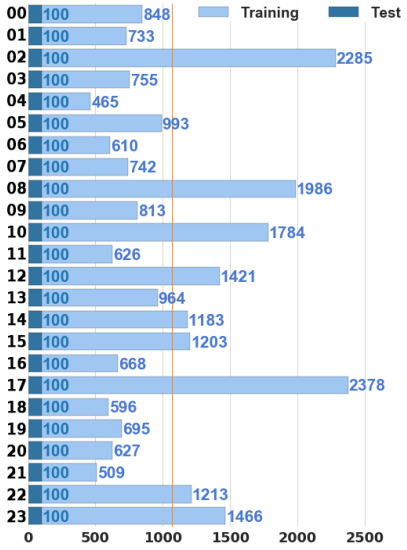


Fig. 2. Organization and Distribution of the dataset created for this work.

Fig. 2 shows the distribution of these classes of diseases ordered according to the labels vector. According to the number of available examples for training, the orange line represents the arithmetic mean of the distribution (1065). There are 25563 total images for training and 100 for tests for each class. Labels of the graphic are represented by: 00 - Acnes, 01 - Bacterial Infections, 02 - Benign Tumors, 03 - Bug Bites, 04 - Bullous Diseases, 05 - Connective Tissue Diseases, 06 - Contact Dermatitis, 07 - Dermatitis, 08 - Eczemas, 09 - Exanthems and Eruptions, 10 - Fungal Infections, 11 - Hair Diseases, 12 - Keratosis and Carcinomas, 13 - Melanomas, 14 - Nail Diseases, 15 - Nevus, 16 - Pigmentation Disorders, 17 - Psoriasis, 18 - STDs, 19 - Systemic Diseases, 20 - Urticaria, 21 - Vascular Tumors, 22 - Vasculitis and 23 - Warts.

B. Data Augmentations

A parameter adjusted differently from the one proposed by ResNet authors are the options of Data Augmentation [14]. The Caffe framework implemented natively the Mirror, the 10-crop, Scale and the subtraction of mean of all images. The rest of the parameters of the ResNet are the same as those proposed by the ResNet authors [15]. For the creation of a dataset, a pattern supported by a Caffe was used, along with a LMBD database [16]. By creating the dataset, the images are resized to 256 square pixels, meaning, the rectangular images end up distorted. This distortion promotes an interesting variable

factor. With this in mind, all the images of the dataset were rotated in 90 and 180 degrees, as the rectangular images are rotated and resized after that, they end up adding a factor of extra variance, which is very important, because it generates different distortion types to the same image. It is possible that other transformations have more interesting results, but we need to create is an optimized pipeline to generate these rotations without the runtime is compromised due to the high computational cost of rotating images.

C. Transfer Learning

To perform the training, we use the ResNet model already trained and provided by own model creators and available in Caffe's Model Zoo [15]. We exchange the last three layers of the model, being these layers Pooling, InnerJoin and Softmax. There is no technical change in these three layers. Only exchanged layers already trained and converged by three identical layers, but without any training or adjustment. The rest of the layers that have not been changed, their weights were fixed and the training was performed only in the three layers were exchanged, in other words, ResNet was used as a features extractor [7] and the last three layers are responsible for the classification of test and validation examples.

IV. Training

For training was chosen the ResNet 50 layer version, the training was performed by script generating using Google Protobuf in the Caffe Framework, and was used all the parameters suggested by ResNet authors [15]. This parameters are: *learning rate* set to 0.0001, *learning rate policy* set to "step", *weight decay* set to 0.0005, *step size* set to 160000 and *gamma* set to 0.95. Only the *Batch Size* was adjusted to suit the available training hardware, in case the VGA is a GTX TITAN X with 12GB of memory, which permits us to place a batch size of 30 images for training and four for tests. The model was trained at approximately 250 thousand iterations.

After training, all test dataset samples were tested through the deep network using Python/Caffe scripts and all results of softmax layer were stored in CSVs files. Then we used data analysis techniques to generate all statistics and graphs presented in the section VI.

V. Experiments

Only one test was created to analyze results for 100 examples of Nevus and Melanoma available in the test dataset. The test analyses all the outputs of the Sotmax layer for this 200 test cases. This test produced five possible results: the first is Melanoma classification hits, False Negative (the error of classification of a Melanoma as a Nevus), the Nevus classification hits, False Positive (the error of classification of a Nevus as a Melanoma) and the last possible result is when the error points out another diseases besides the two analyzed in this test.

VI. Results and Discussion

This section brings a series of statistical plots with the finality of presenting the results obtained in the test proposed on the section V. In Fig. 3 we have the results of the tests executed by the training module of Caffe Framework. On the X axis we have the iterations of model and on the Y axis we have the percentage accuracy on the model, on green is the Test Accuracy and on blue we have the Training Accuracy. The model converges well and the test result stabilized on accuracy of 60% analysing all the test examples.

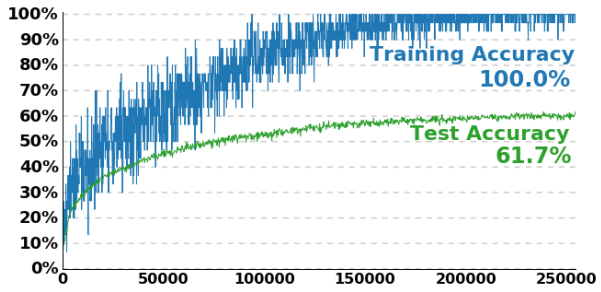


Fig. 3. Accuracy of the model. On the X axis we have the iterations of the model and on the Y axis we have the percentage accuracy of the model.

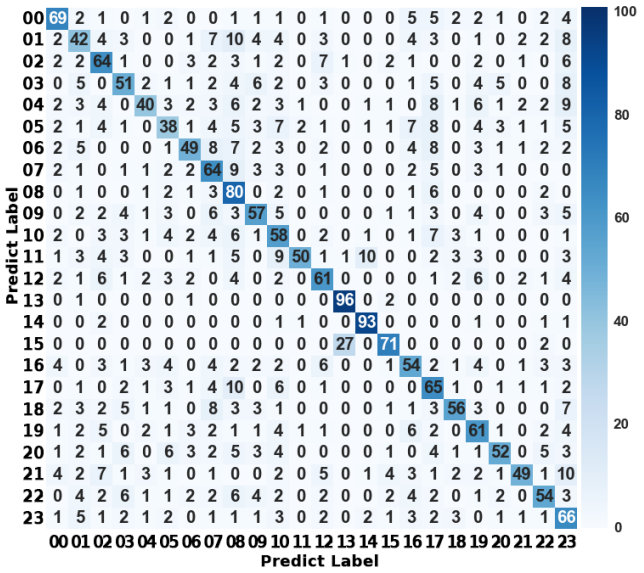
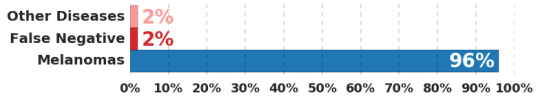


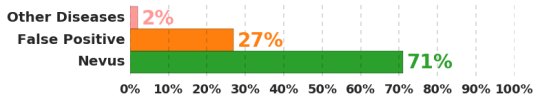
Fig. 4. Confusion Matrix of the all test examples of the dataset. Every number represents one label of the dataset listed on the section III-A.

In Fig. 4 we have the confusion matrix for the 100 examples of each of the 24 diseases classes. Our model has an accuracy of 60% in the 2400 test examples. We can clearly see that the Melanomas (class 13) has the highest result among all classes, although it is not the class with the largest number of samples available. Then, we discarded the trend due the distribution of examples in class, as the most numerous class of the dataset (Psoriasis) has a accuracy slightly above the

arithmetic average of the model. We can also see in Fig. 4 that the False Positive is the most common model error. The lower values are represented in white and light blue colors and the highest values are represented in dark blue color.



(a) Results of the 100 tests on the Melanoma examples.



(b) Results of the 100 tests on the Nevus examples.

Fig. 5. Quantitative results of the test proposed on the section V.

In Figure 5 we have the quantitative results of the test proposed on the section V. We see clearly that the model has a great convergence to classify Melanoma, because there are only 4 cases classified wrong, two of them are considered False Negatives. The high incidence of False Positives make explicit the difficult of this particulary task, it comes down to differentiate Melanoma from a Nevus.

Lastly, in Fig. 6 we have the probabilistic results of the test proposed on section V. The arithmetic mean values show a great convergence of the model, this results are present on the Falses Negative and Positive cases. Melanoma has the best arithmetic mean value while the best single result is a False Positive. The Other Diseases cases has no good results such as the others, this show the examples that are very difficult to classify correctly.

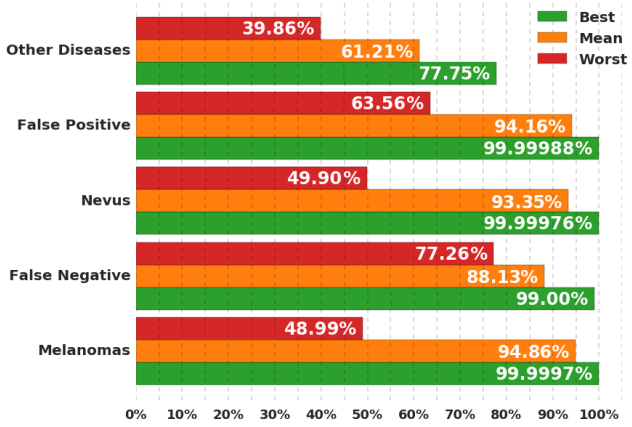


Fig. 6. Probabilistic results of the 5 possible results of the test proposed on section V.

VII. Conclusion

In the specific case of Melanoma, differentiating these from the Nevus is very difficult, according to a high level of false positives in contrast to a low level of false negatives produced by the model. Due to the quantity of available examples of Nevus being 239 more than the Melanoma class, a tendency to Melanoma was discarded. In other words, a tendency for the Nevus instead of the Melanomas was expected, even with the limitations already stated. The result is very exciting and part of this result is due to the use of a model that has better convergence than other models. Another factor that contributed to the good results was the implementation of image rotation as a Data Augmentation. In addition, an important characteristic noted during the executed tests is that the model is totally invariable to watermarks in the photos, because around 80% of the images possess watermarks. These images were mixed between the training and tests sets, and in none of the examples, the watermarks interfered in the results.

Thus, it is concluded that the deep learning can in fact come to solve very important problems regarding visual data processing and also possesses an enormous potential to be used in medicine as a tool to aid in diagnostics. For that, well patterned image datasets must be constructed with a large quantity of examples for each of the millions of diseases that exist. Such a measure may open up the possibility of executing a complete training of such images instead of using a Transfer Learning, as was used during this study, guarantying a better generalization of the model.

References

[1] A. Krizhevsky, I. Sutskever, and G. Hinton, "Imagenet classification with deep convolutional neural networks," *NIPS 2012*, pp. 1–9, 2012.

[2] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large scale image recognition," *ICLR 2015*, 2015.

[3] D. Strigl, K. Kofler, and S. Podlipnig, "Performance and scalability of gpu-based convolutional neural networks," *18th Euromicro Conference on Parallel, Distributed and Network-based Processing*, p. 317, 2010.

[4] A. Esteva, B. Kuprel, and S. Thrun, "Deep networks for early stage skin disease and skin cancer classification," *Stanford*, 2015.

[5] D. Kumar, M. J. Shafiee, A. Chung, F. Khalvati, M. Haider, and A. Wong, "Discovery radiomics for computed tomography cancer detection," *Cornell University Library*, 2015.

[6] I. Goodfellow, Y. Bengio, and A. Courville, "Deep learning," 2016, book in preparation for MIT Press. [Online]. Available: http://www.deeplearningbook.org

[7] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson, "How transferable are features in deep neural networks?" *Advances in Neural Information Processing Systems 27 (NIPS 14)*, 2014.

[8] R. D. Azulay, *Dermatologia*, 6th ed. Guanabara Koogan, 2006.

[9] G. Botelho, M. Batista, and M. Aurélio, "Detecção de câncer de pele do tipo melanoma utilizando redes neurais artificiais," *20th Brazilian Simposium on Computer Graphics and Image Processing*, 2007.

[10] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *NIPS2015*, 2015.

[11] Caffe, "Caffe - a deep learning framework," http://caffe.berkeleyvision.org/, 2015, last access 05/30/2016.

[12] S. P. Domain, "Dermnet nz: the dermatology resource," http://www.dermnetnz.org/, Jan. 2016, last access 1/1/2016.

[13] DermWeb, "Dermweb dermatology links and resources," http://www.dermweb.com/photo_atlas/default.htm, 2015, last access 05/30/2016.

[14] D. V. Dyk and X.-L. Meng, "The art of data augmentation," *Journal of Computational and Graphical Statistics*, pp. 1,50, 2001.

[15] K. He, X. Zhang, and J. S. Shaoqing Ren, "Deep residual networks," https://github.com/KaimingHe/deep-residual-networks, 2015, last access 05/30/2016.

[16] Symas, "Lightning memory-mapped database - lmdb," https://symas.com/products/lightning-memory-mapped-database/support-and-documentation/, 2015, last access 05/30/2016.