Machine Learning for Automated Diagnosis of Skin Lesions

Fábio Santos Department of Electronics
Telecomunications and Informatics
University of Aveiro
Aveiro, 3810-193
Email: fmts@ua.pt

Abstract—Machine learning, specifically, deep learning is a fast-growing field in medicine which is being used for multiple medical imaging related problems, such as early detection of skin cancer. Even though, such systems don't lay out a 100% accurate diagnostic they aim to provide support for both dermatologists in the decision making process and for patients that don't have access to skin professionals. In this paper we will focus on the current state of skin cancer recognition using CNN's (Convolutional Neural Networks) and on skin cancer recognition applications for dermatology decision support. We aim to integrate these concepts into real world use by demonstrating how these systems have potential to change the landscape of medical imaging.

1 Introduction

1.1 Background

Skin cancer is the most common cancer in the United States and worldwide, particularly, in America 1 in 5 persons will develop skin cancer by the age of 70 [1]. However, skin cancer represents a international problem for the health community. For instance, in Europe, over 100,000 people are diagnosed with melanoma and 22,000 deaths are caused by this form of skin cancer annually [2]. But the most interesting fact about skin cancer is that when detected early, the 5-year survival rate for melanoma is 99 percent, as opposed to 23 predicted percent rate from late stages [1].

Initially, automated diagnosis of skin lesions was made based on predefined techniques well known by dermatology professionals, but often failed to either generalize to new cases or lacked the accuracy of a human. However in more recent years, a lot of work has put into medical applications of machine learning. The reason behind this shift are technical, namely:

- Huge amounts of data collected over the years, specifically, labelled skin cancer images
- Exponential computing power growth over the years
- Deep learning algorithm research

Contrary to other traditional machine learning algorithms, deep learning removes the need for feature engineering, which is a quite time consuming process which is both difficult to do and can introduce human error. In addition, it is relatively easy to adapt or modify existing deep learning architectures on new applications.

1.2 Objectives and Motivation

The main objective behind this work is to improve the current work on deep learning techniques through CNN's and to develop a production ready application which enables easy communication between patients and dermatologists. The first version of this application should be able to allow image sharing between patients and dermatologists as well as provide a deep learning based decision support tool for dermatologists. Because this tool has the intent of being production ready, it is highly important to improve the deep learning algorithm as much as possible to the point that it becomes comparable to a dermatologist in terms of accuracy.

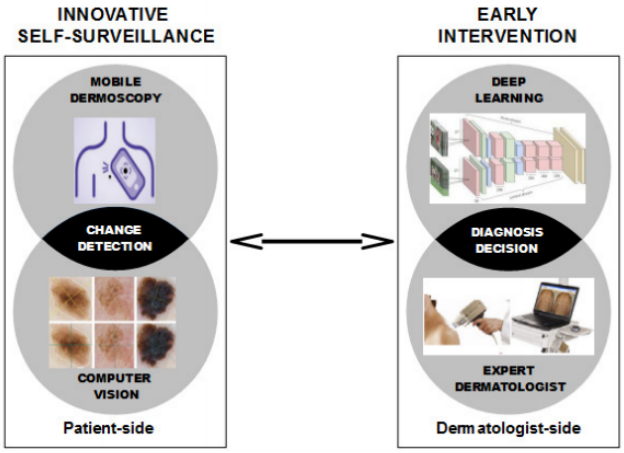

Fig. 1. Dissertation Work Plan

An important consideration for this work is to design an architecture which allows for future expansion of the system, for example for providing new features such as a self surveillance change detection which would allow users to monitor their own lesions.

2 LITERATURE REVIEW

2.1 eHealth and mHealth applications

The emergence of eHealth/mHealth applications presents exciting opportunities to enhance clinical care, health promotion, and disease prevention. However, it can be quite a challenge for professionals to adapt to these systems and integrate them into the clinical workflow, particularly, to seek guidance for decision making.

2.2 Deep neural networks

Deep learning algorithms are providing exciting solutions for medical image analysis problems and they are seen as a key method for future applications. The initial impact of deep learning for medical imaging was revealed through a special issue published in the IEEE Transactions on Medical Imaging (Greenspan, Ginneken and Summers, 2016). The surveys by Hu et al., (2018) and Litjens et al. (2017) contributes also to a clear understanding of the principles and methods of neural network and deep learning concepts, showing how the algorithms based on deep models are being applied to medical image in a wide variety of application areas. The successes of deep learning architectures such as deep neural networks (DNNs), deep belief networks (DBNs) and recurrent neural networks (RNNs) are now well reported in the areas of computer vision, speech recognition, natural language processing and gaming. A comprehensive and up to date approach to deep learning can be found elsewhere (Goodfellow, Bengio and Courville, 2016; LeCun, Bengio and Hinton, 2015).

2.3 Skin lesion classification using deep learning

The first Deep Convulutional Neural Network (a type of deep network) for skin cancer classification was first introduced in 2017 by Esteva et al. [3], which classified keratinocyte cancer and melanoma. The authors follow a transfer learning approach by leveraging the weights of the InceptionV3 network trained on ImageNet, on top of which they build their own classifier. Finally, they measured the network's performance by pitting it against 21 dermatologists that resulted in comparable accuracy to that of those board-certified dermatologists. This network used a very large set of labelled images in order to achieve high accuracy, namely, 129 450 clinical images (including 3374 dermoscopic images) [3]. This data was a combination of biopsy proven datasets from Edinburgh Dermofit Library, Stanford Hospital and from a initiative called ISIC.

2.3.1 International Skin Imaging Collaboration (ISIC)

One of the most important factors which determines a network's performance is the dataset used to train it on. Several public datasets such as ImageNet are quite useful to create generic models, however, for skin lesion classification datasets such as the BCN_20000 [4] are used.

Unfortunately, it is difficult and in many times impossible to compare the performance of published classification results since many authors use nonpublic datasets for training and/or testing [5]. Despite this, the HAM10000 dataset is the closest we have to a benchmark dataset for testing deep networks currently available and therefore we should use it to test ours [6].

International Skin Imaging Collaboration (ISIC) arose from the need for an open source public access archive of skin images and are trying. This archive serves as a public resource of images for teaching and for the development and testing of automated diagnostic systems and every year places a challenge around their datasets [7].

In part 3 of the ISIC 2018 challenge participants were asked to develop a classifier to distinguish between 7 different types of skin cancer. The provided dataset for the challenge was the HAM10000 but participants could complement that dataset by gathering their own. Finally, participant's were ranked based on their normalized multiclass accuracy (accuracy of the classifier on each of the classes averaged together) [8]. The top 3 submissions had balanced accuracies of about 0.885, 0.882, 0.871 respectively and were all submited by a company called Metaoptima [9], which as we will see later has products related to skin lesion classification on production environments. To train those networks they used the competition's HAM10000 dataset along with the ISIC Archive and other proprietary data. Additionally, they augmented the training data by performing random horizontal flips, random rotations, changes in brightness, saturation, and contrast. They used a method called transfer learning (which we will describe later in more detail) from several models trained on ImageNet (such as InceptionV3 or ResNet-50), and then choosed the best-performing and ensembled them together [9].

The 2019's version of this challenge asked participants to classify dermoscopic images among nine different diagnostic categories such as ”melanoma” or ”dermatofibroma”, however this time around one of the classes was unkown (none of the others). Similarly to the 2018's version participants could use their own data to improve the network's performance and were ranked based on a balanced multiclass accuracy [7]. The results turned out to be quite promising, with the best submission posted by Geesert et al. [10] scoring 92.6% accuracy. They trained their networks on the HAM10000, [6], BCN_20000 [4] and MSK [11] datasets. The following points represent their training process:

- Preprocessing methods such as cropping are applied based on the mean intensity differential between the mole area and the non mole area, they binarize the images, apply the shades of gray color constancy method (just like the 2018's top 3) and finally resize the larger images in the datasets.
- Data augmentation is performed by randomly changing brightness, contrast, rotation, scale, shear and flip.
- Two different input strategies are used. While the first takes a random crop from the preprocessed image, the second randomly resizes and scales the image when taking a crop from the preprocessed one.
- They use a transfer learning based approach relying on EfficientNets that were trained on the ImageNet dataset. For each model predictions are made based which input strategies was used.
- Finally, they find the optimal subset of models and ensemble them together

2.4 Teldermathology and deep learning

Because of the visual nature of a skin examination, teledermatology has the potential to become a powerful tool in the diagnosis and management of skin diseases, especially in rural areas where specialty services may not be available. Currently, several production ready skin lesion classification systems are currently available for both skin professionals and patients wishing to self monitor skin moles. However, almost none of them has shown to be sufficiently accurate and/or reliable enough for a clinical environment.

One of the most popular applications for this purpose is Metaoptima's Dermengine web application. Their Visual Search tool compares a user-submitted image with similar images in a database of thousands of pathology-labelled images gathered from expert dermatologists around the world. Deep learning techniques are used to search for related images based on visual features such as colour, shape, and patterns [12].

3 METHODS

3.1 Artifitial neural networks (ANN's)

Artificial Neural Networks (also known as ANNs) compose a category of machine learning algorithms which are inspired by biological neural networks from either humans or animals alike.

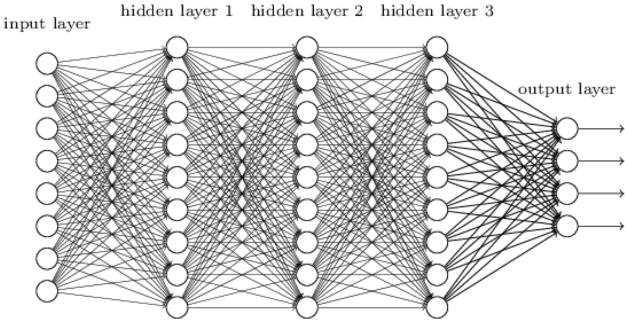

Fig. 2. Deep Neural Network

3.2 Convulutional neural networks (CNN's)

Deep convolutional neural networks or some close variant are used in most neural networks for image recognition problems [?]. They still retain the core concepts of ANN's, such as the way neurons operate and the layered architecture which flows data through the network in order to output a result. However, there are 3 different concepts which distinguish this variant from normal artificial neural networks:

- Local receptive fields: Each neuron in the first hidden layer will be connected to a small region of the input neurons (called a local receptive field).
- Shared weights: Weights and biases are shared across the hidden neurons so that convolutional networks become well adapted to translation variances in images. The shared weights and bias are often said to define a kernel or filter. To the map of shared weights from the input layer to the hidden layer we call feature map. A feature detected by a hidden neuron is some kind of input pattern that will cause the neuron to activate. To do image recognition we need more than one feature map in order to recognize multiple features.
- Pooling layers: Usually are used immediately after convolutional layers. What the pooling layers do is simplify the information in the output from the convolutional layer. A common pool layer is max pooling which provides a way to know if a given feature is found anywhere in a region of the image. [13]

3.3 Transfer learning vs learning from scratch

Supervised learning using deep neural networks requires large amounts of data and computational power in order to train models and determine the network parameters such as weights. However, both of which can either be impossible to have or quite difficult to acquire. Even when one has lots of data and computational power often times the fitting process takes a long time, especially while debugging the network to determine a good model fit.

A common way to solve this problem is the use of a technique called transfer learning. Transfer learning is a method of reusing a model or knowledge for another related task [14], in deep learning specifically, it means to carry weights and biases from a generic model trained from a generic dataset like ImageNet which contains 20000 categories like ”strawberry” or ”balloon” and using those for another model with a different purpose.

In CNN's, as inputs are passed along the network hidden layers closer to the input layer output generic features like shapes and curves, while hidden layers closer to the output layers represent more specific categories such as ”strawberry” or ”balloon” (like the models trained on the ImageNet dataset). Transfer learning aims to obtain parameters from layers that output generic features and build layers on top of those which output specific problem related classes, such as ”melanoma” or ”nevi”.

Often times, transfer learning techniques are used to solve image classification problems [15], such as the one that we are trying to solve. However, learning from scratch is also an option which can significantly decrease the number of parameters which define the network. To understand why that is, one can think that a lot of parameters in Transfer Learned models are being used to describe features which are not needed for specific problem domains. For instance, if the source model was trained to different foods, it might have learning to indentify its textures. for the and. Having less parameters means that the model is much smaller compared to transfer learning models (which often reach 100Mb) and can be used on devices such as smartphones which have less computational power.

3.4 Network Architectures

Over the years several convolutional neural networks architectures have been developed and tested against state of the art benchmark challenges such as the Imagenet Large Scale Visual Recognition Challenge (ILSVRC) [16]. We will describe a few widely used on image classification problems.

VGGNet is a network architecture that became quite popular by achieving excellent performance on the ImageNet dataset. There are some public variations of this network, one of which having 16 weight layers (VGG16) that generalizes well onto different datasets, making it a good candidate for our future network. The layers of this architecture are displayed in figure 2.

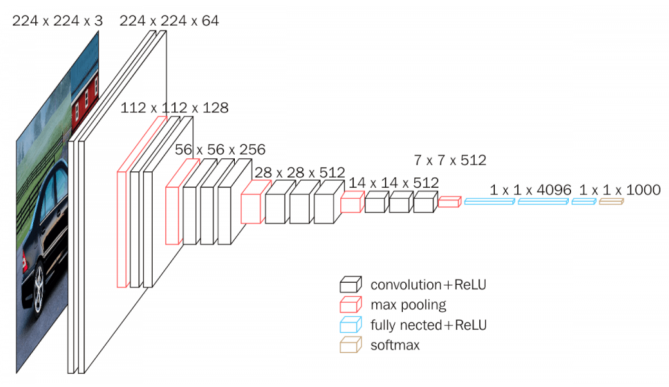

Fig. 3. VGG16 Architecture

3.5 Training neural networks

3.5.1 *Overfitting and underfitting*

The bias and variance trade off is a well known problem in deep learning. The bias of a model is the error caused by the assumptions made to approximate the model to the true predictions. In turn, the variance of a model is an error from sensitivity to small fluctuations in the training set. One must fine tune the model to both accurately make predictions from the training data while generalizing to new data, meaning, we must find a good trade off between bias and variance which minimizes the total error from the model. If the model underfits then it does not perform well even on the training data, and therefore has high bias and low variance. Simply put, the model doesn't predict well on either training data or new data. However, a common problem while training deep networks is to produce a model that performs well on the training data but that generalizes poorly to any new data [17], such that it starts to either memorize inputs or learn noise. In this case, we say that the model overfits and therefore has low bias but very high variance.

In order to evaluate whether a model is underfitting or overfitting one should use state of the art metrics which help describe what is happening while training. Multiple solutions to the overfitting problem have been proposed and tested over the years. Methods such as dropout or regularization techniques will certainly help us train our model by reducing complex co-adaptations between neurons, in order to deploy the proposed model into real world use.

3.5.2 *Expanding the training data*

No matter how hard we try to optimize the networks parameters, a model is highly dependent on its dataset. A good dataset is has to:

- Represent the real world
- Not be biased, meaning that it favors some classes
- Be diverse, even if it means that some entries contain noise
- Contain lots of examples

As we've seen there are several datasets available which are labelled for skin lesion diagnosis, however some of them are quite biased towards some specific class or lack large amounts of examples for a specific class. A bad dataset can easily cause the network to overfit because it does not provide enough proper real world examples for the network to produce a good bias variance trade off. As such, when some real world variation is introduced either by noise or some other factor the network fails to predict the class.

One way we intent to improve our dataset is through a concept called data augmentation. The main idea behind this concept is to expand the training data by applying operations that reflect real-world variation [13], which it turn introduces diversification and size to the dataset.

There is two ways of expanding the training data:

- General transformations
- Generative models

We will try to apply both in order to improve our classification accuracy.

4 Conclusion

The widespread of the e-Health and m-Health applications along the rise of mobile computing open new opportunities to develop and deploy applications which provide support for patients and health professionals. At the same time, deep learning shows significant improvements in skin lesion diagnosis compared with many other machine learning methods. Even though, there are still challenges to overcome, such as the requirement for large datasets in order to achieve near dermatologist it is expected that these models will continue to improve over the years as more and more data becomes publicly available.

References

[1] "Skin Cancer Facts & Statistics - The Skin Cancer Foundation," 2019. [Online]. Available: https://skincancer.org/skin-cancer-information/skin-cancer-facts/

[2] F. Bray, J. Ferlay, I. Soerjomataram, R. L. Siegel, L. A. Torre, and A. Jemal, "Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries," *CA: A Cancer Journal for Clinicians*, vol. 68, no. 6, pp. 394–424, nov 2018.

[3] A. Esteva, B. Kuprel, R. A. Novoa, J. Ko, S. M. Swetter, H. M. Blau, and S. Thrun, "Dermatologist-level classification of skin cancer with deep neural networks," *Nature*, vol. 542, no. 7639, pp. 115–118, feb 2017. [Online]. Available: http://www.nature.com/articles/nature21056

[4] M. Combalia, N. C. F. Codella, V. Rotemberg, B. Helba, V. Vilaplana, O. Reiter, C. Carrera, A. Barreiro, A. C. Halpern, S. Puig, and J. Malvehy, "BCN20000: Dermoscopic Lesions in the Wild," aug 2019. [Online]. Available: http://arxiv.org/abs/1908.02288

[5] T. J. Brinker, A. Hekler, J. S. Utikal, N. Grabe, D. Schadendorf, J. Klode, C. Berking, T. Steeb, A. H. Enk, and C. von Kalle, "Skin Cancer Classification Using Convolutional Neural Networks: Systematic Review." *Journal of medical Internet research*, vol. 20, no. 10, p. e11936, oct 2018. [Online]. Available: http://www.jmir.org/2018/10/e11936/ http://www.ncbi.nlm.nih.gov/pubmed/30333097 http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC6231861

[6] P. Tschandl, C. Rosendahl, and H. Kittler, “Data descriptor: The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions,” aug 2018.

[7] “ISIC 2019.” [Online]. Available: https://challenge2019.isic-archive.com/

[8] “Task 3: Lesion Diagnosis — ISIC 2018.” [Online]. Available: https://challenge2018.isic-archive.com/task3/

[9] A. Nozdryn-Plotnicki, J. Yap, and W. Yolland, “Ensembling Convolutional Neural Networks for Skin Cancer Classification.”

[10] N. Gessert, M. Nielsen, M. Shaikh, R. Werner, and A. Schlaefer, “Skin Lesion Classification Using Loss Balancing and Ensembles of Multi-Resolution EfficientNets.”

[11] N. C. F. Codella, D. Gutman, M. E. Celebi, B. Helba, M. A. Marchetti, S. W. Dusza, A. Kalloo, K. Liopyris, N. Mishra, H. Kittler, and A. Halpern, “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),” oct 2017. [Online]. Available: http://arxiv.org/abs/1710.05006

[12] “DermEngine — Visual Search.” [Online]. Available: https://www.dermengine.com/en-ca/visual-search

[13] M. Nielsen, *Neural Networks and Deep Learning*, 2018. [Online]. Available: http://neuralnetworksanddeeplearning.com/

[14] T. G. Dipanjan Sarkar, Raghav Bali, *Hands-On Transfer Learning with Python*, first, kin ed. Packt Publishing, 2018. [Online]. Available: https://books.google.pt/books?id=aPFsDwAAQBAJ&pg=PA155&lpg=PA155&dq=Transfer+learning+is+the+idea+of+overcoming+the+isolated+learning

[15] P. Ly, D. Bein, and A. Verma, “New Compact Deep Learning Model for Skin Cancer Recognition,” aug 2019.

[16] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, “ImageNet Large Scale Visual Recognition Challenge,” pp. 211–252, dec 2015.

[17] J. Grus, *Data Science From Scratch*. O’Reilly Media, 2015.