Skin Lesion Recognition Using Randomly Regularized Convolutional Networks

***Abstract*— Deep learning architectures have become the leading instrument for image-based recognition in multiple fields, including dermatology. However, while the deep networks are data hungry, in medicine-related subjects the amount of annotated data is limited and often comes in imbalanced datasets, which can lead to memorization, and subsequent poor performance on the testing part. This paper proposes the introduction of a simple technique for regularization during training, implemented by the injection of a randomized quantity of controlled magnitude in the gradient of the classification loss function. The method is evaluated on a challenging dermatology database and is shown to significantly improve the performance of baselines and compares favorably with established prior art.**

Keywords—skin lesion; random; perturbation; gradient.

I. Introduction

Following their dominance in the ILSVCR grand image recognition challenges, the convolutional neural networks (CNNs) have become the dominant tool for biomedical image classification and analysis. They have been proposed as basis for instrument for fully automatic and semi-automatic (aided) clinical diagnosis in disciplines such as radiology, histology, ophthalmology, and dermatology [1]. In this paper we address the classification and recognition of skin lesions in dermatology.

In terms of actual motivation, we shall remind that due to the fact that skin is directly exposed, its diseases are some of the most common human illnesses as they affect the health of 30% to 70% of individuals, with higher rate for subpopulations [2]. For dermatology diagnosis, the most common tool is the dermatoscope which permits the observation of the latent structures of skin lesions and the extraction of visual cues needed for establishing the diagnosis. Yet, the dermatoscope is hard to access in poor regions [3], and moreover, it is not really necessary for many common skin diseases where appropriate light and a lens are sufficient for observation. Thus, developing an effective skin disease diagnosis system based on easily accessed clinical images would be beneficial and could provide low-cost, universal access. In this paper we envisage a solution based on a smartphone, using its camera to acquire dermatoscopic-like images and its processing power to run an algorithm for recommendation of the diagnosis. The algorithm is based on deep convolutional neural networks.

Compared to the standard image classification problems, such as included in the ILSVR competition, where huge amount of data was available on internet and limitation come only from the resources available for annotation, in medical imagining in general and in dermatology in particular, the number of training samples for each skin disease depends heavily on the incidence of that disease [4]. While there exists more than 1000 different kinds of skin diseases, both common and uncommon, it is also difficult to collect or annotate a rich and balanced datasets.

To address the limitations of deep learning classification on such datasets we propose a simple technique that assumes the injection of random perturbation in the loss function. Such injection acts as a regularization factor and the experiments performed on the challenging SD-198 dataset shows that the proposed approach improves the baseline performance with almost 3%.

The remainder of the paper is organized as follows: in the next section we review relevant prior art, and then follow with the presentation of the proposed method. The paper continues in section IV with details about implementation and achieved results and ends with some conclusions

II. Prior art

A. Skin Lession recognition

Initial works focused on the usage of dermatoscopic images. These images, taken under uniform, bright illumination conditions, are sufficiently clear for recognition. In this direction, works before deep learning used classical image analysis approaches: image segmentation on a multiscale resolution followed by extraction of standard features (location of center of mass, number of pixels, mean value of pixels etc.) [5], border detection followed by description with gray level co-occurrence matrix followed by classification with SVM [6].

IEEE978-1-6654-4000-4/21/$31.00 ©2021IEEE

In the deep learning era, solutions have multiple facets. Advances were built upon publication of large collection such as the one used in this work, SD-198 [7]. The class imbalance was further addressed in self-paced mechanism [8]. Noting the limited amount of data, there were several attempts to use together the network, using its layers as features coupled to an SVM classifier. For instance, Kawahara [9] used directly the features from a network with an SVM. With a more elaborate framework, Yu et al [10] started with a network used as feature extractor and followed with a shallow network, fully connected for classification.

B. Random strategies for backward update in deep learning.

On the technical side, this paper contributes by the introduction of a regularization term based on the injection of randomness which alters the gradient (given its path decided by the SGD algorithm). This work approaches a direction that contains several other notable works which are relying on various strategies that involve the use of random perturbations injection as a mean to prevent memorization in deep network training.

Into this category, some outstanding approaches are *dropout* and *shake–shake*. Srivastava et al. [11] introduced *dropout* as random choice of parameters at each training iteration where a group of weights stands still. *Shake-shake* regularization was proposed by Gastaldi et al. [12] in networks with parallel prances where the standard summation of branches with a random weight. In the same direction lies the *cutout technique* [13], which refers to regularization by randomly masking out square regions of input images during training iterations. Our solution is simpler and can be stacked with these ones.

In more recent developments, Frankle and Carbin suggested by the so-called *lottery ticket hypothesis* [14], that "large networks that train successfully contain sub–networks that, when trained in isolation, do converge in a comparable number of iterations to comparable accuracy". While seeking sub-networks, randomly selected weights are masked and the resulting performance is evaluated to determine the increase of performance. In contrast, our proposal is to use a single point of injection that in the final loss function gradient, while seeking the same purpose.

III. Method

We will describe here a training algorithm that is applied to any standard deep convolutional network during training. While this contribution focuses on deep networks, it is not intrinsically connected to it and may be extended to any other learning model.

For an efficient problem formulation, the learner, the dataset and the method are denoted and developed as follows. Data is denoted by standard association: $X = \{\mathbf{x}_1, \dots, \mathbf{x}_N\}$ are data instances and $Y = \{y_1, \dots, y_N\}$ their respective annotations. Data is drawn from a probability density function $p(X, Y)$, with marginals over the input $p(X)$ and respectively over the output $p(Y)$. Part of the dataset is used for testing, with data $Xt = \{\mathbf{x}_{N+1}, \dots, \mathbf{x}_{N+T}\}$ and labels $Y = \{y_{N+1}, \dots, y_{N+T}\}$, the pair being used for establishing the training performance.

The deep convolutional learner takes an input example from X and produces a vector of class confidence scores. This is denoted by $H_\Theta : X \rightarrow R^C$, where Θ are the weights sought in the training process and C is the number of classes. For instance, in the case of SD-198 database, $C=198$. The network can be seen as composed by an input part and an output part. The output part, $h_f : \mathrm{R}^d \rightarrow \mathrm{R}^C$ takes the final descriptors and produces a confidence score. The result of the composition, the overall function h, maps the input space onto the output probabilities, which encode what is the chance, from the learner point of view, that an input belongs to a class. Thus, the network outputs, for a given instance $\boldsymbol{x}_j$ is $\mathrm{H}(\mathbf{x}_j)$ and the prediction is the argument of the maximum confidence score according to the soft max paradigm:

$$y_j = arg\max_i H(x_i) \tag{1}$$

In eq. (1) the subscript i iterates through the classes denoted by the i-th dimension of the vector. Furthermore, weights of the networks are determined by the following minimization problem:

$$H_\theta(x) = arg\min_\theta(J) = arg\min_\theta\big(L(x;y;\theta) + R(\theta)\big) \tag{2}$$

where $R(\theta)$ is a the standard $L2$ regularization term (seeks the lowest magnitude for the weights, $R(\theta) = \left\|\sum_k \theta_k\right\|_2$, $L(\cdot)$ is the cross entropy loss function and J is the cumulative loss as it gathers everything in a single term. The process aims to compute the best set of weights Θ as they lead to optimum performance. The weights are found by gradient descent in the backward step:

$$\theta_i^{k+1} = \theta_i^k + \varepsilon\frac{\partial J(\theta)}{\partial\theta_i^k} \tag{3}$$

where $\theta_i \in \Theta_{opt}$ is a network weight found to lead to optimal values and ε is the learning rate. As the loss function J is complex, by both a sum and the composite function that models the learner, the chain rule is used to compute the derivative of the loss function with respects to sought weight. The chain cascades derivative of the top scaled with the derivative of the top with respect the derivative of the weights from lower layers:

$$\theta_i^{k+1} = \theta_i^k + \varepsilon\frac{\partial J(\theta)}{\partial\theta_{top}}\frac{\partial\theta_{top}}{\partial\theta_i^k} \tag{4}$$

where θ_{top} is the top weight, placed in the final, softmax layer of the network.

One basic assumption of this approach is that the training set if largely enough such that, in the many iterations of the stochastic gradient descent, the correct direction for minimization will come out. The stochastic gradient descent acts, from a point of view in a cumulative manner, as many local directions are cascaded; from another point of view, it acts in a greedy manner as the optimization is local: it seeks the optimum direction from the current point (set of weights).

Yet, there are cases where data is insufficient and deep networks are capable of simply memorizing the training set. In such a situation a regularization is needed, that will prevent it from being greedy.

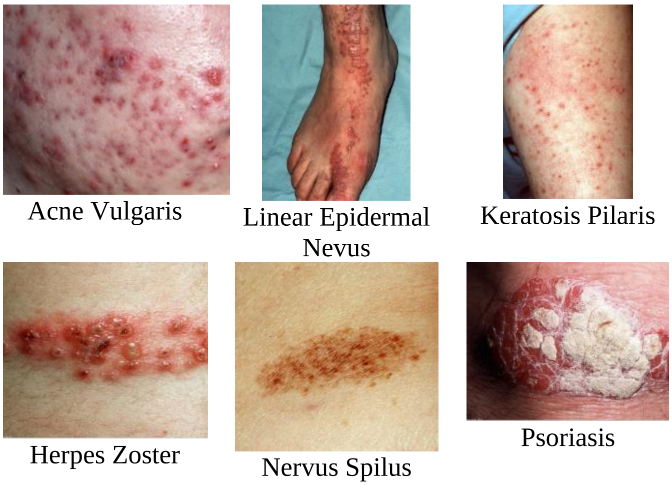

Fig 1. Example of various classes represented in the SD-198 database

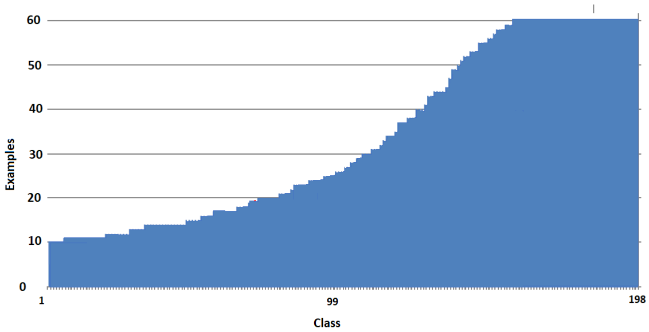

Fig. 2 Distribution of examples with respect to the class in the SD-198 database

The actual proposal is to add a random perturbation in the gradient and thus push towards generalization instead of memorization. In particular, we change the derivation of the backward pass as follows:

$$\theta_i^{k+1} = \theta_i^k + \varepsilon\left(\frac{\partial J(\theta)}{\partial \theta_{top}} + \xi\right)\frac{\partial \theta_{top}}{\partial \theta_i^k} \qquad (5)$$

In equation (5) above ξ is a random quantity drawn from a zero mean Gaussian distribution.

IV. RESULTS

A. Implementation

The proposed code is based on the PyTorch framework and is accelerated by Titan X GPU. The network architecture of choice is ResNet-18 without any regularization based on randomness. The resolution of choice for images is 224 x 224. The optimization has been carried using Stochastic Gradient Descent with batches of 64 images. Training took 100 epochs, divided over three periods upon which the learning rate was kept constant to 0.01, divided by 10 and respectively by 100. Standard weight decay was also included.

B. Dataset

The dataset of choice is (Skin disease) SD-198 [12]. It constructs the basics for automatic skin disease recognition and diagnosis problem. The data contains 6584 clinical images that are divided into 198 categories of skin diseases. The situation covered by this datasets are widely varied as it contains samples from various races (Caucasian, Asian, African) and thus color of the skin (white, black, brown, and yellow), both genders (male and female), various age (child, adult, and old), disease body location (head, nails, hand, and feet), and different periods of instatement of the lesions (early, middle, and late). The images contain variations in color, exposure, illumination, and scale as they have been acquired with digital still cameras and mobile phone in less restrictive conditions. The images were uploaded by patients to the dermatology Dermquest website and annotated by professional dermatologists. A selection of images may be followed in figure 1. The database is unevenly distributed as one can see in figure 2; this fact further increases the difficulty of correct recognition,

The partition between training and testing is the one from the introductory work [7]: 50% of the images went in training and an equal number (3,292) in testing. Other works used uneven division and reported higher performances, but we will focus on this simple scenario.

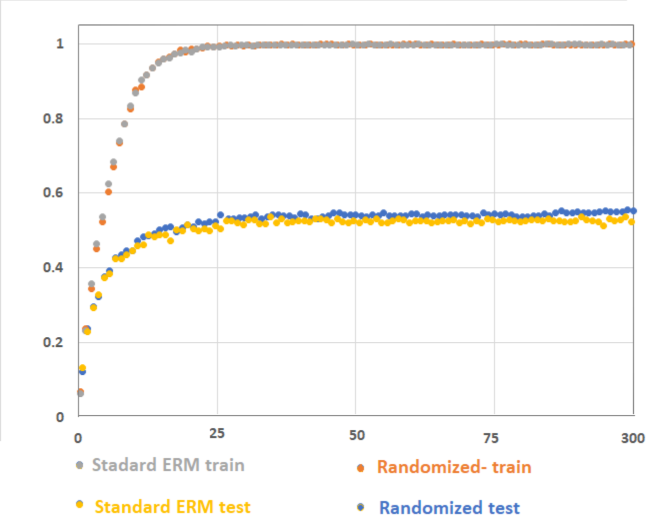

Fig. 3 The behavior in training and testing of the proposed technique compared to standard empirical risk minimization by SGD

C. Performance and comparison with state of the art

The task investigated was that of skin lesion classification based on deep learning solution. The performance is reported in table I. First we report the performance when standard classification is used (case in which empirical risk minimization is used), without any randomness injection; this is marked as baseline. Next, we report the performance of previous strong solutions. Furthermore, thanks to [8], which reported performance of doctors, we can compare the automatic solution with expert opinion. One may see that our approach leads to top performance, even compared to junior dermatologist doctor.

TABLE I. ACCURACY ON THE SD-198 DATABSE OF VARIOUS SOLUTION. "FT" STANDS FOR FINETUNING, WHEN THE NETWORK WAS PRETRAINED ON IMAGENET AND FINETUNED (LAST LAYER) ON THE SD-198 DATABASE

Method	Architecture	Performance
Proposed	**ResNet-18**	**55.26%**
Baseline	ResNet-18	52.38%
SIFT +SVM [7]	SVM	25.85%
COlorNames+SVM[7]	SVM	20.20%
CaffeNet [7]	AlexNet	42.31%
VGG+ft [7]	VGG-16	52.15%
GoogleNet[8]	GoogleNet	35.33%
GoogleNet+ft [8]	GoogleNet	46.48%
ResNet [8]	ResNet-34	48.78%
ResNet+ft [8]	ResNet-34	53.35%
Doctor – generalist	-	49.00%
Doctor – junior dermatologist	-	52.00%
Doctor – expert dermatologist	-	85.00%

The proposed solution notably improves the baseline performance by almost 3 percent, proving the method efficiency. It should be emphasized that the accuracy result reported must be correlated with the number of decision classes (198) and the (average) number of examples available per class, which is less than 33 in the SD-198 database. For comparison, we will recall that some related state of the art research dealing with the use of CNN classification of skin melanoma used on average from 369 images per class [15] up to 14380 images per class [16] for a 9-class decision.

V. CONCLUSIONS

In this paper we have proposed the usage of a random perturbation in the gradient of the loss function as a technique efficiently usable for the regularization of the training of a convolutional neural network when it is used on a small database (as compared to the number of classes required in the final decision). The proposed application is that of skin lesion detection within 198 disease classes, where the proposed regularization increases the baseline accuracy by almost 3%.

Acknowledgment

We gratefully acknowledge the support of NVIDIA Corporation with the donation of the Titan V GPU used for this research.

References

[1] F. Piccialli, V. Di Somma, F. Giampaolo, S. Cuomo, G. Fortino, A survey on deep learning in medicine: Why, how and when?, Information Fusion, Volume **66**, 2021, Pages 111-137.

[2] R. J. Hay, N. E. Johns, H. C.Williams, I.W. Bolliger, R. P. Dellavalle, D. J. Margolis, R. Marks, L. Naldi, M. A. Weinstock, S. K. Wulf, et al. The global burden of skin disease in 2010: an analysis of the prevalence and impact of skin conditions. *Journal of Investigative Dermatology*, 134(6), 2014.

[3] R. J. Hay and L. C. Fuller. The assessment of dermatological needs in resource-poor regions. *International journal of dermatology*, 50(5), 2011

[4] Tajbakhsh, N., Shin, J. Y., Gurudu, S. R., Hurst, R. T., Kendall, C. B., Gotway, M. B., & Liang, J., "Convolutional neural networks for medical imageanalysis: Full training or fine tuning?" IEEE Trans. Med. Imag., vol. 35,no. 5, pp. 1299–1312, May 2016

[5] Maglogiannis, I., Delibasis, K.K.: Enhancing classication accuracy utilizing globules and dots features in digital dermoscopy. Computer methods and programs inbiomedicine 118(2) (2015) 124-133

[6] Celebi, M.E., Kingravi, H.A., Uddin, B., Iyatomi, H., Aslandogan, Y.A., Stoecker,W.V., Moss, R.H.: A methodological approach to the classication of dermoscopy images. Computerized Medical Imaging and Graphics 31(6) (2007) 362-373

[7] X Sun, J. Yang, M Sun, K. Wang., "A benchmark for automatic visual classification of clinical skin disease images," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 206–222

[8] Yang, J., Sun, X., Liang, J. and Rosin, P.L., 2018. Clinical skin lesion diagnosis using representations inspired by dermatologist criteria. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 1258-1266).

[9] J. Kawahara, A. BenTaieb, and G. Hamarneh, "Deep features to classify skin lesions," in Proc. IEEE Int. Symp. Biomed. Imag., Apr. 2016,pp. 1397–1400.

[10] Yu, L., Chen, H., Dou, Q., Qin, J. and Heng, P.A,"Automated melanoma recognition in dermoscopy images via very deep residual networks," IEEE Trans. Med. Imag., vol. 36, no. 4,pp. 994–1004, Apr. 2017

[11] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever,and R. Salakhut-dinov, "Dropout: a simple way to prevent neural networks from over-fitting," *The journal of machine learning research*, vol.15, no.1, pp.1929–1958, 2014.

[12] X. Gastaldi,"Shake-shake regularization of 3-branch residual networks," in *Proc. of Int. Conf .on Learning Representations Workshops*, 2017.

[13] T. De Vries and G. W. Taylor, "Improved regularization of convolutional neural networks with cutout," *arXivpreprint arXiv:1708.04552*, 2017.

[14] C. Frankle and M. Carbin, "The lottery ticket hypothesis: Finding sparse, trainable neural networks," in *Proc. of Int. Conf .on Learning Representations*, 2019.

[15] Zhang, J., et al. Attention residual learning for skin lesion classification. IEEE Trans. on Medical Imaging, 2019, vol.38, no. 9, pp. 2092-2103.

[16] Esteva, A., et al. Dermatologist-level classification of skin cancer with deep neural networks. Nature, 2017, doi:10.1038/nature21056, vol. 542.7639: 115-118.