Swiderski *et al.*
EURASIP Journal on Image and Video Processing (2022) 2022:3
https://doi.org/10.1186/s13640-022-00580-y

EURASIP Journal on Image and Video Processing

RESEARCH Open Access

Random CNN structure: tool to increase generalization ability in deep learning

Bartosz Swiderski[1], Stanislaw Osowski[2*], Grzegorz Gwardys[5], Jaroslaw Kurek[1], Monika Slowinska[3] and Iwona Lugowska[4]

*Correspondence: stanislaw.osowski@ee.pw.edu.pl
[2] Faculty of Electrical Engineering, Warsaw University of Technology and Faculty of Electronic Engineering, Military University of Technology, 75 Koszykowa Street, 00-662 Warsaw, Poland
Full list of author information is available at the end of the article

Abstract

The paper presents a novel approach for designing the CNN structure of improved generalization capability in the presence of a small population of learning data. Unlike the classical methods for building CNN, we propose to introduce some randomness in the choice of layers with a different type of nonlinear activation function. The image processing in these layers is performed using either the ReLU or the softplus function. This choice is random. The randomness introduced in the network structure can be interpreted as a special form of regularization. Experiments performed on the detection of images belonging to either melanoma or non-melanoma cases have shown a significant improvement in average quality measures such as accuracy, sensitivity, precision, and area under the ROC curve.

Keywords: Convolutional neural networks, Generalization, Melanoma recognition, Classification

1 Introduction

The generalization of artificial neural models refers to their ability to adapt to the new, previously unseen data that come from the same distribution as that used when the model was learned. It means to transfer the knowledge acquired in the learning process to the new situation by referring to previously unseen test data, thus combining the new experience with previous experiences that are similar in one or more ways.

Neural networks learn from examples of patterns that represent the training database. In the learning phase, a network adopts its structure and parameters to respond properly to the input signals. From the statistical point of view, it corresponds to understanding the mechanism, based on which the learning data have been created [1–3]. Such a mechanism can be significantly distorted by the noise that interferes with the data. Therefore, it is useful to reduce the noise level and recover the denoised image. Such an approach, using information from multiple views via neural networks for image retrieval, has been presented, for example, in the papers [4, 5].

However, the biggest problem is that the network may not be complex enough to properly learn the mechanism of data generation, or there may be a situation where the population of learning data is too scarce and does not represent the process being

modeled sufficiently well. The most important problem in obtaining good generalization properties of neural networks, especially deep structures, is the limitation of learning resources.

According to the theory of Vapnik and Chervonenkis [6], the population of learning samples should be sufficiently large in terms of the number of fitted parameters to produce a well-generalized neural model. In many cases, especially in deep learning, this condition is very difficult to satisfy [1]. Therefore, the test performance of a network may vary from one test set to another. To obtain the most objective measure of the generalization ability of the network, many repetitions of the learning and testing phases with different data are used, usually organized in K-fold cross-validation mode [7].

The generalization ability strongly depends on the relation between the size of learning data and the complexity of network architecture. The higher this ratio, the better probability of good performance of the network on the data not taking part in learning.

Many different techniques have been elaborated to improve the generalization ability of deep neural networks [8–10]. One of them is increasing the population of learning samples, based on the augmentation of data. Augmentation is a technique that is used to artificially expand the size of a training dataset by creating modified versions of data in the dataset. Different methods are proposed: flips, translations, rotations, scaling, cropping, adding the noise, non-negative matrix factorization, creating synthetic images using self-similarity, application of GAN technique or variational autoencoder, etc. [11–16]. However, in deep structures where the number of parameters is very high (millions of parameters), such a technique has limited efficiency.

A good way to increase the generalization is the regularization of the architecture. It is implemented by the modification of structure, as well as using different methods of learning. It was shown that the explicit forms of regularization, such as weight decay, dropout, and even data augmentation, do not adequately explain the generalization ability of deep networks [17, 18]. The empirical observations have shown that explicit regularization may improve the generalization performance of the network, but is neither necessary nor by itself sufficient for controlling the generalization error.

The important role fulfills the implicit regularization built into the learning algorithms. For example, stochastic gradient descent converges to a solution with a small norm, which might be interpreted as implicit regularization. A similar role performs early stopping and batch normalization in the learning procedure [19].

An important method for increasing generalization capability is the modification of network structures. It is especially popular when forming an ensemble of networks [15]. Different, independent team members, looking at the modeled process from a different point of view, form a so-called expert system, which makes it possible to generate a more objective decision.

Specific approaches have been proposed that allow increasing the independence of ensemble members. To such methods belong random choice of learning data used in training of particular units of an ensemble, application of mini-batches created randomly in the adaptation process of parameters, diversification of drop-out ratio of learning data, etc. Such techniques allow the creation of ensemble members that differ in operation in the hope of obtaining a more accurate classification of the test data that did not participate in the learning phase [9, 10]. All approaches: direct explicit regularization,

augmentation of data, and modification of network structures are usually combined to develop a better generalizing system.

In our work, we take a step further to implicit regularization of deep structure. It combines the ensemble approach and random integration of the results at each level of signal processing. Two parallel structures are created and learned simultaneously. Their integration is based on the introduction of randomness in the formation of the subsequent layers of the CNN network in both architectures. We show that such a method leads to the improvement of the generalization ability at the limited size of learning data.

In each stage of the final structure formation, we form two parallel layers that perform the same task. Both have a similar form (same number of filters, kernel size, and padding parameters), but differ in parameter values and type of nonlinear activation function (here ReLU and softplus). In the final structure formation of the network, only one of these two layers is chosen and this choice is completely random. This random selection occurs at each level of signal processing, up to the final classification level of softmax.

The idea of such an approach follows from the observation of gradient methods in optimization, applied to the problem with many local minima (typical case in deep learning). Fixed parameters of the structure tend to the closest local solution, which is not necessarily the best one. Introducing a random choice at the level of each layer allows us to explore a wider range of possible solutions and find a better result.

The numerical experiments performed on the medical data representing melanoma and non-melanoma cases have confirmed the superiority of such an approach over a standard one, relying on the same type of activation function in each step of signal processing.

2 Methods

The study aims to introduce a novel approach to building a convolutional neural network structure of the increased generalization ability. We propose the randomly constructed structure composed of two different independently working CNN networks. The proposed architecture of the network will be applied in the recognition of dermoscopic images representing cases of melanoma and non-melanoma lesions.

The idea of random structure is based on the assumption that in each level of signal processing the randomly chosen activation function will be used. In our solution, they are either rectified linear unit (ReLU) or softplus. It means also choosing the whole set of parameters associated with the chosen layer. This idea is illustrated for the two first local layers in Fig. 1.

The parallel sublayers apply the same number of filters, kernel size, and padding, but of different filter parameter values. They differ also by the activation function and this fact is the most differing both sublayers. The network consists of several layers. In each layer, only one sublayer is randomly selected with a specific activation function (ReLU or softplus) and the associated parameters. The procedure of drawing in a random process is performed using a uniform distribution in the range [0, 1], mean = 0.5, and standard deviation 1/12. The applied threshold of 0.5 is applied in the selection of the left or right channel in signal processing. Either the left or right layer, shown in Fig. 1, is chosen on each level in image processing. This is depicted by the dashed lines. Only the results of

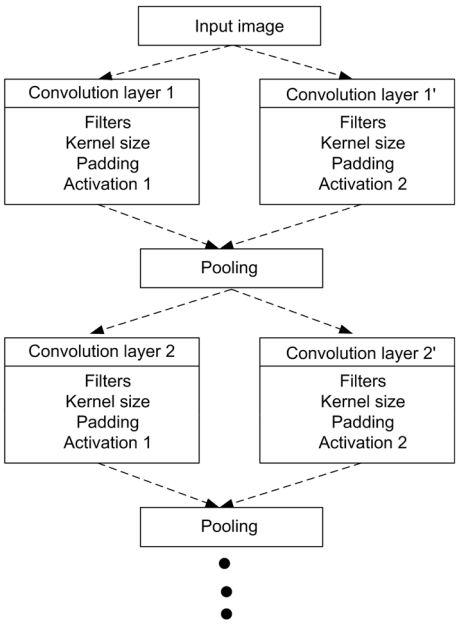

Fig. 1 The structure of the random selection of the layer parameters. The parallel layers on each level differ only by the activation function and parameter values of the filters. The input image is processed either by the left or the right side layers, and this selection is random (left or right side channel)

the chosen layer are subject to pooling operation. This process of a random choice of either left or right direction of signal flow is repeated in all layer levels.

The results of such multilayer local image processing are combined into the flat vector and delivered to the softmax classifier for developing the final decision in classification.

As a result, image processing is performed by the randomly chosen multilayer architecture. Each run of learning procedure applies a different network structure, although the hyperparameters of the layers (number of filters, kernel size, stride, and padding values) are fixed.

Two types of activation functions are considered: the rectified linear unit:

$$y(x) = \begin{cases} x \ dla \ x > 0 \\ 0 \ dla \ x \leq 0 \end{cases}, \tag{1}$$

and softplus

$$y(x) = \ln\left(1 + e^{x}\right). \tag{2}$$

The ReLU activation allows a network to obtain sparse representation. The observation of initialized weights using uniform initialization has shown that around 50% of units are close to zeros, increasing in this way the sparsity-inducing regularization. Moreover, because of linearity in the positive range of signal values, gradients flow well in active paths of neurons, mathematical investigations are easier and computations are also cheaper (no need for computing the exponential functions) [1, 2]. The softplus version of the rectifying function loses the exact sparsity. Although one might expect that softplus has the advantage over ReLU that it can be differentiated everywhere or is less

completely saturated, empirical studies have shown that this is not the case [2]. Our experiments, presented in this paper have confirmed this finding.

As a result of such a procedure of structure formation, the obtained network shows the randomness in its activity. Thus, it can be considered as a special form of regularization, which generally leads to an increase in generalization ability. However, the proposed approach can be treated as implicit network regularization—or perhaps a more precise limitation of degrees of freedom (because the weights in individual segments/network members have to be adjusted in such a way as to be compatible with each other. The performed experiments have shown that this network has a better capacity for generalization. Moreover, combining both activation functions in a random structure creates a unique value in the generalization property of the system.

3 Experimental data

Input data used in experiments were taken from Warsaw Memorial Cancer Center and Institute of Oncology, Department of Soft Tissue/Bone Sarcoma and Melanoma as dermoscopic images. The whole database of images was prepared by two medical experts (coauthors of the paper) with long-year experience in this field. The database was also medically assessed by other expert dermatologists by applying the ABCDE dermoscopic criteria and confirmed by exact pathomorphological inspection, including medical segmentation of the lesion and clinical and histological diagnosis. The database contained 112 RGB images of verruca seborrhoica representing non-melanoma and 101 images of basal cell carcinoma (melanoma). The total number of images was only 213. All images were acquired by dermatologists using a dermatoscopy of the magnification $20\times$. The registered images of the lesions stored in JPEG format were of different sizes extending from 767×576 to 4273×2848 pixels.

The examples of original sample images of both classes are shown in Fig. 2. The variety of shapes and colors of the lesions is visible. The colors are distributed in a different way within the images. The background in each image is different and the size, color, and

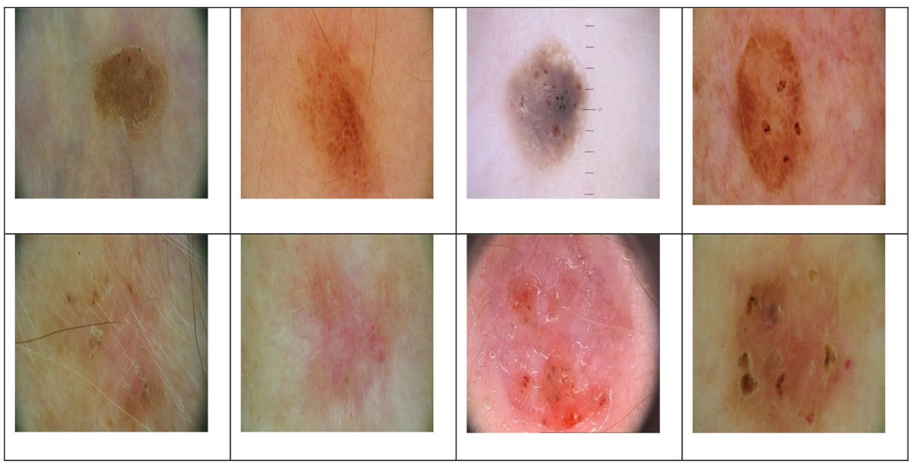

Fig. 2 The exemplary images represent both classes of images: the upper row—class of verruca seborrhoica, the lower row—basal cell carcinoma. We can also see the similarities of samples representing two opposite classes and at the same time large differences among samples belonging to the same class

shape of the region of interest (ROI) representing the skin lesions change from specimen to specimen. We can also see some similarities of samples representing two opposite classes.

The registered images differ significantly by size. Some of them were very large and some much smaller. All of them contain wide background areas of no interest in the recognition process. Therefore, in the first stage of processing, the region of interest (ROI) suggested by the medical experts was extracted from the images. In this way, the total size of all images was unified and reduced to only 32×32. In further analysis, the melanoma cases are referred to as class 1 and the non-melanoma cases as the second class. The recognition task is simplified into two classes.

3.1 Applied random CNN structure

The proposed CNN network construction method can be used to solve various classification problems. This work focuses on biomedical image processing, especially melanoma detection. The structure and hyperparameters of the network have been specifically adapted for this purpose. The final form of the solution is the result of many implementation experiments. In each design step, the numerical results expressed by the chosen quality measures (accuracy, sensitivity, precision, AUC) of the proposed random structure are compared with the corresponding classical approach for the CNN design to prove its advantages. The best structure with its associated hyperparameters is proposed as the best one.

As a result, the random CNN structure applied in the solution of this recognition problem was composed of 3 locally connected layers and two dense layers with a final softmax classifier. Two parallel sublayers on each architecture level differing by the filter parameters and the activation function are randomly chosen. These twin layers contain the same number of filters (however, of different parameter values), the same kernel size, and zero paddings. They differ by the nonlinear activation function: either ReLU (left branch) or softplus (right branch). On each level, either the left or the right path of the structure is randomly selected in image processing. It means that only one layer path (either left or right) on each level is included in the CNN structure. The general architecture of the network is depicted in Fig. 3. The dotted lines depict a random choice of the actual path in image processing.

The size of the input image is 32×32. The first two levels of locally connected layers apply 32 filters and in the third layer, the number of filters was increased to 64. The kernel size in all layers is 3×3 and has no zero padding. The max-pooling is applied to each level of local image processing. The pooling operation is performed on the linear convolution results generated either by the left or right path, chosen randomly. The structure and its hyperparameters have been obtained after a series of introductory experiments.

The results of the third locally connected layer (64 images of the size 2×2 after a pooling operation) are flattened into the 256-dimensional vector. The first fully connected layer reduces this size to 64 and the second layer to only two (the number of classes). The final classification decision is made by softmax. The softmax classifier is responsible for the recognition of melanoma and non-melanoma cases. The training of the CNN structure was performed using the ADAM algorithm with a mini-batch size equal to 10.

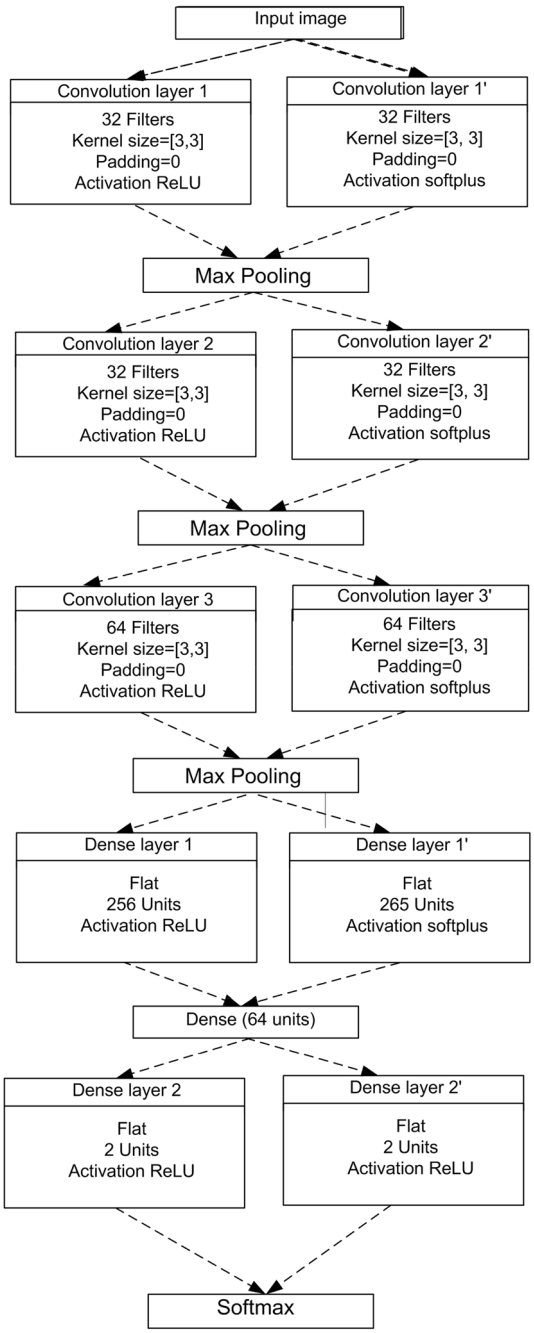

Fig. 3 The random structure of CNN applied in classification. On each level, there is a random choice of image processing: either left branch or right. The random choice is repeated for both locally and fully connected (dense) layers. Final recognition of classes is done by Softmax

The training procedures were performed by using Python 3.7 implementation of TensorFlow, with software executing on Windows. The process of single learning the whole system, performed on single PC (RAM 64 GB, processor: 12 × Intel(R) Xeon(R) CPU E5-2640 v3 @ 2.60 GHz, graphic card GeForce GTX 960-4 GB) took about 2 h for 100,000 epochs. This is a long time due to the very large number of learning epochs

applied in experiments. This excessive number of epochs was applied to show the effect of overlearning the CNN structure. In practice, we can use much fewer iterations and the learning time will be much shorter.

Moreover, the learning process could be significantly accelerated by applying cluster system architecture and parallel computing. After the training phase, the parameters of the network are fixed and the whole structure is ready for testing. The testing phase of a single image is very short and is measured in a fraction of a second.

4 Results and discussion

The results of the application of the proposed strategy are assessed based on statistics of particular runs of experiments. The most objective way of assessment is the application of the so-called tenfold cross-validation. The original data set is randomly partitioned into 10 equal size subsets, each containing approximately the same population of samples belonging to both classes. Of these 10 subsets, a single subset is retained as the testing data set and the remaining 9 are used in training the system. This process is then repeated 10 times (the folds), with each of 10 subsets used only once as the testing data. The results of testing are then averaged over all folds to produce a single final result of testing. The mean value plus standard deviation of results is presented. Thanks to such a strategy, all observations are used for both training and testing.

The training epochs were repeated 100,000 times and the averaged results of the testing were registered. Three different systems have been investigated. The first two were classical, with a single layer on each level of processing. One system applied ReLU activation and the second softplus. Their structures were stiff and represented either the left or right side of the system presented in Fig. 3. In the third set of experiments true random choice of structure, shown in Fig. 3, was investigated. It means that now some layers have applied ReLU and other softplus activation. In this way, each fold was performed using the structures that differ by layer parameter values and activations. The quality measures in the form of accuracy (ACC) and area under the ROC curve (AUC) were registered in all experiments.

Figure 4 shows the changes in the averaged test accuracy ACC (tenfold validation mode) in the following learning epochs (the results of a particular run of the

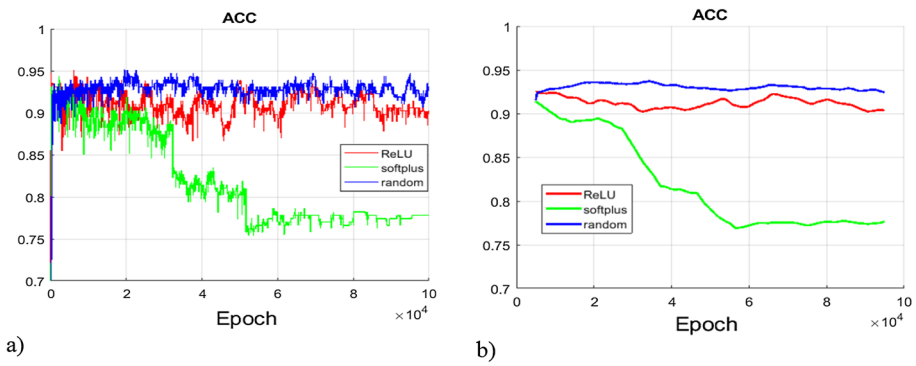

Fig. 4 The changes of testing accuracy in the following learning epochs: **a** temporary changes; **b** the smoothed moving average form. A very long learning procedure depicts a significant effect of overlearning in the case of softplus. ReLU and random structure do not experience such an overlearning effect

algorithm). A very high number of epochs was used to show the overlearning phenomena of the network.

The upper figure represents the temporary changes corresponding to each epoch and the lower one their smoothed moving average form. It is evident that the application of either softplus or ReLU, working independently, results in inferior ACC results.

Especially inefficient is the softplus activation. The inclusion of random choice between ReLU and softplus applied in the CNN leads to the improvement of accuracy and the stabilization of testing results at the prolonged learning time. Taking into account the smoothed curves at the end of learning epochs it is seen that the averaged ACC of random CNN was increased to the value of 92.3% from 90.5% in CNN with ReLU and from 77.5% with softplus activation. Especially high sensitivity to the length of learning has been observed in the case of softplus. The overlearning effect is seen very early and the stabilization is on a very low level of 77%.

The same tendency is observed in the case of the area under the ROC curve (AUC). The changes of this quality measure within the succeeding learning epochs are presented in Fig. 5 (the results of the particular run of the algorithm). The application of fixed softplus or ReLU (working independently) has shown much worse results compared to the random structure. Again, softplus activation is the least efficient, and the random choice between layers in the CNN structure significantly improves the results. The AUC corresponding to the random CNN reached the final smoothed value above 0.96, while the CNN based on ReLU was approximately 0.94. Once again the application of softplus has resulted in the worst AUC value of around 0.83. It should also be emphasized that the values of ACC and AUC are higher than the values shown in [16] for similar databases when using very sophisticated, however, classical data-processing systems.

Observe that the proposed solution applies randomness at each level of signal processing. Therefore, the important issue is the repeatability of the results. The next experiments were conducted to test how the quality of the solution changes with different repetitions of the experiments. Ten repetitions were performed (all arranged in the tenfold cross-validation mode). Their results in terms of average values and

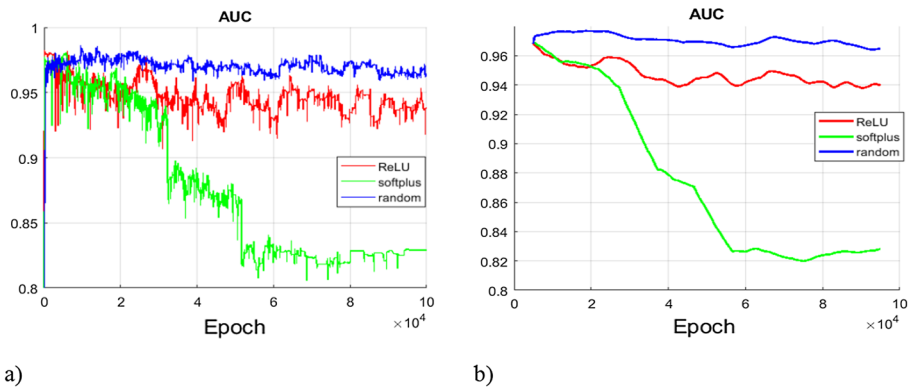

Fig. 5 The changes of AUC in the following learning epochs: **a** temporary changes; **b** the smoothed moving average form. Once again the advantage in generalization ability of the proposed random structure is visible

Table 1 The statistics (average ± STD) of the quality measures of the melanoma versus non-melanoma recognition obtained in 10 repetitions of the experiments

	ACC [%]	SENS[%]	SPEC[%]	PREC[%]	F1 [%]	AUC
ReLU	91.4 ± 5.92	91.2 ± 8.52	91.8 ± 8.88	92.6 ± 8.62	91.9 ± 5.84	0.955 ± 0.0414
Softplus	89.9 ± 5.79	89.4 ± 7.63	90.5 ± 6.21	91.4 ± 7.32	90.4 ± 5.75	0.947 ± 0.0487
Random	91.8 ± 5.40	92.4 ± 6.71	91.1 ± 8.55	92.3 ± 7.13	92.3 ± 5.02	0.971 ± 0.0312

Table 2 The comparison of values of score margin for three compared structures of CNN

	ReLU	Softplus	Random
γ	0.830 ± 0.124	0.800 ± 0.117	0.836 ± 0.109

standard deviation are presented in Table 1. They show accuracy ACC, sensitivity (SENS), specificity (SPEC), the precision of melanoma (PREC), F1 measure, and AUC, where these measures have been defined in a standard way, as given in [7].

The most important quality measures (accuracy, sensitivity, F1, and AUC) obtained in experiments show an advantage of random networks over classical solutions. This advantage is especially well seen concerning the network applying softplus activation function. The interesting phenomenon is also the smallest standard deviation value in all categories of quality measures. Concluding, all these results confirm the important role of randomness in increasing the generalization ability of CNN structure.

The presented approach to the designing problem of CNN represents a very specific form of implicit regularization. The results show that applying randomness in choosing the type of activation function in the layers helps in getting better generalization ability of the CNN network. This fact was confirmed also by comparing the margins of the score [19] represented by our random network and the classical CNN structures. The margin of score represents the difference between the score of the true label and the maximum score of other labels obtained for learning data. This margin is defined by [19]:

$$\gamma = p(\mathbf{x})_{[y=true]} - \max\left(p(\mathbf{x})_{[y\neq true]}\right), \tag{3}$$

where $p(\mathbf{x})$ represents the score of the network (probability of a class) at excitation of vector $\mathbf{x}$. It was proved in [19] that the statistical capacity of the network, defined in terms of several examples required to ensure generalization (when the test errors are close to the training errors) is inversely proportional to squared γ. The higher this difference, the wider the tolerance range, and good generalization can be achieved with a smaller population of learning data. Table 2 shows the average values ± standard deviation of the margin for our random network and the classical CNN structures using ReLU and softplus for the learning data.

The highest value corresponds to the random structure of the CNN. Note that the standard deviation in this solution has also reached the smallest value (the highest repeatability of the results). It confirms our supposition that randomness introduced into the network formation forms a special form of implicit regularization.

5 Conclusions

The paper has presented the new approach to designing the CNN structure of the improved generalization ability at a very small population of learning samples. The main aspect of this solution is the introduction of the random choice between two sublayers, which are distinguished by the activation function at each level of the signal flow.

In each processing step, there are 2 options of activation: either ReLU or softplus. So ReLU may be selected in the first layer level and softplus in the second one. Such a small change in the procedure of image processing has shown significant improvement in the generalization ability of CNN. The presented strategy is strongly recommended in the construction of CNN architecture when a very small number of learning samples is available. This trick is universal and can be applied in different forms of deep learning systems.

Although the work shows the application of two of the most popular activation functions (ReLU and softplus), the approach is open to other forms of activation, for example, sigmoidal or the recently introduced idea of the so-called scaled polynomial constant unit activation function [21]. Especially the last form of activation is interesting since the shape of the function can be significantly changed by a few hyperparameters. In further research, we will apply a larger number of activation functions, hopefully leading to a further increase in the generalization capability of deep networks.

The numerical experiments performed on the small image database of melanoma and non-melanoma cases have proved the better efficiency of this approach compared to the classical CNN structures. The proposed random CNN architecture has shown higher values in quality measures (accuracy, sensitivity, specificity, precision, or area under the ROC curve) in class recognition of test samples that did not participate in the learning process. A significant improvement was also observed in the value of the area under the ROC curve. In our opinion, the randomness introduced into the network structure is an efficient form of regularization in deep learning, especially in the case of a very small population of learning samples.

Abbreviations

CNN: Convolutional neural network; ReLU: Rectified Linear Unit; ACC: Accuracy; ROC: Receiver operating characteristic; AUC: Area under roc curve; SENS: Sensitivity; SPEC: Specificity; PREC: Precision; F1: Harmonic mean of precision and recall; ROI: Region of interest.

Acknowledgements

The research was supported by the Centre for Priority Research Area Artificial Intelligence and Robotics of Warsaw University of Technology within the Excellence Initiative: Research University (IDUB) program.

Authors' contributions

BS: conceptualization, methodology, formal analysis, software. SO: writing—original draft preparation, validation, supervision. GG: formal analysis. JK: formal analysis. MS: preparation of medical material. IL: preparation of medical material. All authors read and approved the final manuscript.

Funding

There was no funding for the research reported.

Availability of data and materials

The datasets used and/or analyzed during the current study are available from the corresponding author on reasonable request.

Declarations

Competing interests

There are no financial competing interests.

Author details
[1]Faculty of Applied Informatics and Mathematics, Warsaw University of Life Sciences, 166 Nowoursynowska Street, 02-787 Warsaw, Poland. [2]Faculty of Electrical Engineering, Warsaw University of Technology and Faculty of Electronic Engineering, Military University of Technology, 75 Koszykowa Street, 00-662 Warsaw, Poland. [3]Dermatologic Clinic Military Institute of Medicine, Central Clinical Hospital, Ministry of Defense, Warsaw, Poland. [4]Maria-Sklodowska Curie National Research Institute of Oncology, Warsaw, Poland. [5]Faculty of Electronic and Information Technology, Warsaw University of Technology, 75 Koszykowa Street, 00-662 Warsaw, Poland.

Received: 26 February 2021 Accepted: 12 January 2022
Published online: 08 February 2022

References

1. T. Poggio, Q. Liao, Theory I: deep networks, the curse of dimensionality, bulletin of the polish academy of sciences. Tech Sci **66**(6), 761–773 (2018)
2. Q. Zheng, M. Yang, J. Yang, Q. Zhang, X. Zhang, Improvement of generalization ability of deep CNN via implicit regularization in two-stage training process. IEEE Access **6**, 15844–15869 (2018)
3. P. Zhou, J. Feng, Understanding generalization, optimization performance of deep CNNs, in Proceedings of the 35th International Conference on Machine Learning, Stockholm, Sweden, PMLR 80 1–19 (2018)
4. C. Yan, B. Gong, Y. Wei, Y. Gao, Deep multi-view enhancement hashing for image retrieval. IEEE Trans Pattern Anal Mach Intell **43**(4), 1445–1451 (2020). https://doi.org/10.1109/TPAMI.2020.2975798
5. C. Yan, Z. Li, Y. Zhang, Y. Liu, X. Ji, Y. Zhang, Depth image denoising using nuclear norm and learning graph model. ACM Trans Multimedia Comput Commun Appl (2020). https://doi.org/10.1145/3404374
6. V. Vapnik, *Statistical learning theory* (Wiley, New York, 1998)
7. P.N. Tan, M. Steinbach, V. Kumar, *Introduction to data mining* (Pearson Education Inc., Boston, 2013)
8. Y. Bengio, Y. LeCun, G. Hinton, Deep learning. Nature **521**, 436–444 (2015)
9. J. Brownlee, Deep learning for natural language processing. Develop deep learning models for your natural language problems, Ebook, (2018).
10. I. Goodfellow, Y. Bengio, A. Courville, *Deep learning* (MIT Press, Massachusetts, 2016)
11. I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, Generative adversarial nets. Advances in Neural Information Processing Systems, arXiv: 1406.2661, 1–9 (2014).
12. S. Guan, M. Loew, Using generative adversarial networks and transfer learning for breast cancer detection by convolutional neural networks, Proc. SPIE 10954. Medical Imaging 2019: Imaging Informatics for Healthcare, Research, and Applications. (2019). https://doi.org/10.1117/12.2512671
13. H. Ren, J. Su, H. Lu, Evaluating generalization ability of CNN and capsule networks for image classification via top-2 classification, arXiv: 1901.10112v2 [cs.CV], 1–18 (2019)
14. J. Kurek, B. Swiderski, S. Osowski, M. Kruk, W. Barhoumi, Deep learning versus classical neural approach to mammogram recognition. Bull Polish Acad Sci Tech Sci **66**(6), 831–840 (2018)
15. L. Kuncheva, *Combining pattern classifiers: methods and algorithms* (Wiley, New York, 2004)
16. M. Kruk, B. Świderski, S. Osowski, J. Kurek, M. Słowińska, I. Walecka, Melanoma recognition using extended set of descriptors and classifiers. Eurasip Jo Image Video Proc **43**, 1–10 (2015). https://doi.org/10.1186/s13640-015-0099-9
17. B. Neyshabur, S. Bhojanapalli, D. McAllester, N. Srebro, Exploring generalization in deep learning, NIPS 2017, arXiv: 1706.08947, 1–19 (2017)
18. C. Zhang, S. Bengio, M. Hardt, B. Recht, O. Vinyals, Understanding deep learning requires rethinking generalization, International Conference Learning Representations (ICLR) 2017, arXiv: 1611.03530v2 [cs.LG], 1–15 (2017)
19. Matlab user manual, MathWorks, Inc. Natick, USA, 2017
20. P. Kingma, M. Welling, An introduction to variational autoencoders. Found Trends Mach Learn **12**, 307–392 (2019)
21. J. Kiselak, Y. Lu, J. Svihra, P. Szepe, M. Stehlik, SPOCU: scaled polynomial constant unit activation function. Neural Comput. Appl. **33**, 3385–3401 (2021). https://doi.org/10.1007/s00521-020-05182-1

Publisher's Note