Classification of skin lesions using an ensemble of deep neural networks

Balazs Harangi, *Member, IEEE*, Agnes Baran and Andras Hajdu *Senior Member, IEEE*

Abstract— **Skin cancer is among the deadliest variants of cancer if not recognized and treated in time. This work focuses on the identification of this disease using an ensemble of state-of-the-art deep learning approaches. More specifically, we propose the aggregation of robust convolutional neural networks (CNNs) into one neural net architecture, where the final classification is achieved based on the weighted output of the member CNNs. Since our framework is realized within a single neural net architecture, all the parameters of the member CNNs and the weights applied in the fusion can be determined by backpropagation routinely applied for such tasks. The presented ensemble consists of the CNNs AlexNet, VGGNet, GoogLeNet, all of which have been won in subsequent years the most prominent worldwide image classification challenge ImageNet. For an objective evaluation of our approach, we have tested its performance on the official test database of the IEEE International Symposium on Biomedical Imaging (ISBI) 2017 challenge on Skin Lesion Analysis Towards Melanoma Detection dedicated to skin cancer recognition. Our experimental studies show that the proposed approach is competitive in this field. Moreover, the ensemble-based approach outperformed all of its member CNNs.**

I. INTRODUCTION

Skin cancer – also known as malignant melanoma – is one of the deadliest form of cancer if not recognized in time. Since the pigmented areas/moles of the skin can be nicely observed by simple, non-invasive visual inspection (e.g. by a dermatoscope), the clinical protocols of its recognition also consider several visual features. Namely, perhaps the most classic clinical protocol is the one applying the ABCDE (Assymetry, Border, Color, Diameter, Evolution) rule [1] with a corresponding scoring. Additional features/textures are involved with other scoring schemas like the 7-points checklist [2].

As for the computer-aided support of this field, till the appearance of machine learning approaches considering convolutional neural networks (CNNs), the majority of the computer algorithms tried to recognize the features of the clinical protocols and to apply similar scoring. However, the current spreading of CNN-based approaches in various image processing tasks, like detection, localization, segmentation and classification [3]–[5], has also reached this domain with predicting a machine dominance in recognition accuracy against human experts [6].

In this work, we propose a deep learning-based approach for melanoma detection via the fusion of different individual CNN architectures that have already proven their efficiencies in pattern recognition scenarios. Namely, we compose an ensemble of the CNNs AlexNet [7], VGGNet [8] and GoogLeNet [9], each of which has won the largest publicly available image classification contest ImageNet [10] in different years between 2012 and 2016. As a novel contribution, we remove the final fully-connected and classification layers of these individual CNNs, and interconnect them with inserting a joint fully connected layer followed by the classic softmax/classification layers for the final prediction. In this way, we create a single network architecture from the three member CNNs, which can be then trained by backpropagation in the usual way for neural nets. As an additional refinement to improve decision accuracy, we assign different weights to the outputs of the member CNNs. These weights are also adjusted automatically within our framework – there is no need of manual tuning.

Our experimental evaluation was considered on a publicly available database recently released at the IEEE International Symposium on Biomedical Imaging (ISBI) 2017 challenge on Skin Lesion Analysis Towards Melanoma Detection dedicated to skin cancer recognition [11]. This challenge corresponds to a three-classes classification task to recognize nevus (healthy lesion), malignant melanoma (cancerous lesion), and seborrheic keratosis (age-related, non-cancerous lesion); for sample images see Figure 1. We have found that the ensemble of the CNNs has outperformed all the individual members, so it seems to be a competitive approach in this field, especially with noticing that our framework allows the inclusion of more members in the future.

II. METHODOLOGY

Deep learning-based approaches considering CNN architectures are undoubtedly the primary tools of computer-aided image classification tasks in these days. Their dominance can be dated back to 2012, when the first such CNN architecture AlexNet [7] remarkably improved the recognition accuracy on the challenge ImageNet [10]. In the forthcoming years, several more CNN architectures have been proposed, like GoogLeNet [9], ResNet [12], VGGNet [8], etc.

These architectures were trained on a huge dataset of 1.28 million natural images and are also publicly available. As a common guideline for their application to a specific domain with a limited training dataset, transfer learning [13] can be considered. In this work, we also follow this recommendation with using the weights and biases from the

*Research was supported in part by the projects GINOP-2.1.1-15-2015-00376 and EFOP-3.6.2-16-2017-00015 supported by the European Union and the State of Hungary, co-financed by the European Social Fund.

Balazs Harangi is with the Faculty of Informatics, University of Debrecen, POB 400, 4002, Debrecen, Hungary. (corresponding author to provide e-mail: harangi.balazs@inf.unideb.hu; phone, +36 52 512-900/75121).

Agnes Baran and Andras Hajdu are with the Faculty of Informatics, University of Debrecen, POB 400, 4002, Debrecen, Hungary (e-mail: baran.agnes@inf.unideb.hu, hajdu.andras@inf.unideb.hu).

978-1-5386-3646-6/18/$31.00 ©2018 IEEE

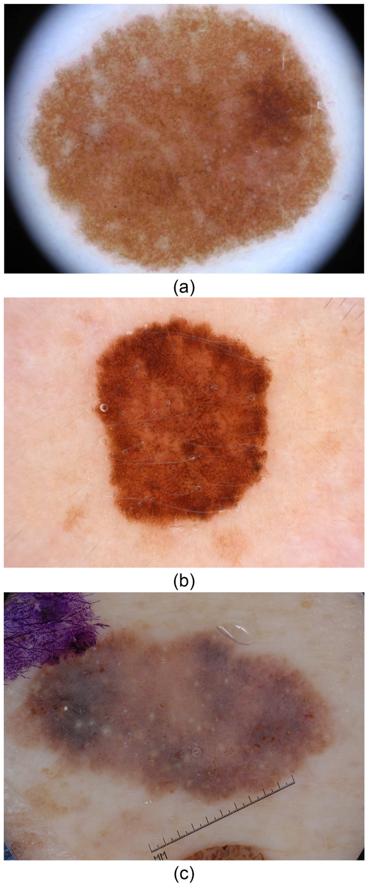

Fig. 1. Different types of skin lesions to be classified: (a) nevus; (b) melanoma; (c) seborrheic keratosis.

pre-trained models with fine-tuning only the semantically deeper features by adding skin lesion data.

A. Members of the ensemble

Our first choice, the CNN AlexNet [7] is composed from 5 convolutional layers some of which are followed by max-pooling ones. Additionally, for final classification 3 more fully-connected layers and a softmax one are considered.

As the second member of our ensemble, we have selected the CNN VGGNet described in [8]. This neural network considers 16-19 layers in depth with applying relatively small dimensional spatial filters for convolution. In this study, we consider VGG16, which variant is built up from 13 convolutional layers; filter size was selected to be 3×3 pixels. Altogether, 5 max-pooling layers are applied before some of the convolutional ones using a mask of 2×2 pixels and stride 2. As final ones, 3 fully-connected layers are applied after a stack of convolutional layers.

Finally, we have selected GoogLeNet for our third member proposed by Szegedi *et al.* in [9]. GoogLeNet consists of 22 convolutional layers including 9 Inception modules. An Inception module contains convolutional kernels of three different sizes (5×5, 3×3, and 1×1 pixels) and a pooling one of size 3×3.

B. The ensemble of the networks

The architecture of the ensemble is shown in Figure 2, where the abbreviations conv/relu/pool/drop/fc stand for the convolutional/rectified linear unit/max pooling/dropout/fully-connected layers, respectively. Besides the layers of the member CNNs, at the bottom of the figure the layers for aggregation and final classification can be seen. The graph also demonstrates that the main contribution of this work is to propose a single neural network architecture, which allows to train all the member CNNs simultaneously to build up an efficient ensemble of them. Since our framework is modular and can be also extended with more members, in the forthcoming formal description we will refer to the members as CNN_1, CNN_2, and CNN_3.

For a proper formulation of our classification problem, let the possible object classes nevus, melanoma, seborrheic keratosis be denoted by C_N, C_M, and C_{SK}, respectively. For training, we consider the data set $X = \{x^{(1)}, \ldots, x^{(M)}\}$ of cardinality M supplied with the corresponding labeling (ground truth) information $\{y^{(1)}, \ldots, y^{(M)}\}$. That is, $x^{(i)}$ is the i-th training image labeled by $y^{(i)} \in \mathbb{R}^3$ with $y^{(i)} = (1, 0, 0)$, $y^{(i)} = (0, 1, 0)$, or $y^{(i)} = (0, 0, 1)$ regarding whether $x^{(i)}$ belongs to C_N, C_M, or C_{SK}, respectively.

To interconnect the ensemble member CNNs, first we remove their original final softmax/classification layers. Then, according to our three-classes classification problem we replace their original 1000-dimensional final fully connected layers with three-dimensional ones. Thus, the outputs of the members CNN_1, CNN_2, CNN_3 for the i-th training sample $x^{(i)}$ are determined as the vectors $\widehat{y}_1^{(i)}$, $\widehat{y}_2^{(i)}$, $\widehat{y}_3^{(i)} \in \mathbb{R}^3$, respectively. These outputs are then aggregated in the fully connected layer fc_AVG in our ensemble architecture via

$$\overline{y}^{(i)} = \sum_{j=1}^{3} A_j \widehat{y}_j^{(i)}, \tag{1}$$

where A_1, A_2, and A_3 are matrices of size 3×3. Notice that, a special weighting is applied in (1) instead of a classic simple weighted average of the outputs, since our experimental results indicated improvement in classification, when the predictions of the member CNNs were allowed to be weighted differently for the three classes. The layer fc_AVG is also followed by usual softmax/classification ones to normalize the result of aggregation for classification. As for notation, we will write $\overline{y}_{SM}^{(i)}$ for this normalized variant of $\overline{y}^{(i)}$. Hence, the overall loss function considering the whole training dataset can be given as the mean squared error (MSE) of the prediction and the manual labeling

$$\text{MSE} = \frac{1}{2M} \sum_{i=1}^{M} \left(\overline{y}_{SM}^{(i)} - y^{(i)} \right)^2. \tag{2}$$

During the training phase, backpropagation is applied to minimize the loss given in (2) via adjusting all the parameters of the member CNNs simultaneously together with the weights considered for aggregation in the matrices A_1, A_2, and A_3. As for preparations, the weighing matrices

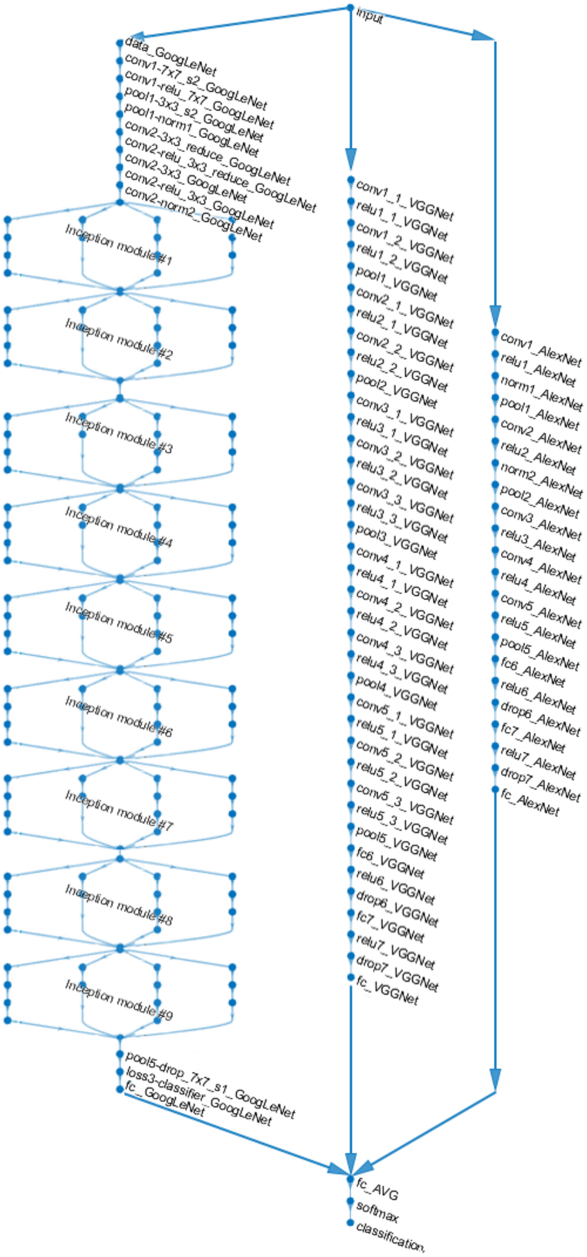

Fig. 2. Architecture of the proposed ensemble of three convolutional neural networks. The ends of the layer names are the name of the original architecture.

are initialized as

$$A_1 = A_2 = A_3 = \begin{pmatrix} 1/3 & 0 & 0 \\ 0 & 1/3 & 0 \\ 0 & 0 & 1/3 \end{pmatrix}, \qquad (3)$$

while the parameters of the members are set to be same as in the pre-trained CNN models before the first forward propagation step.

III. EXPERIMENTAL RESULTS

A. Training and test data sets

For experimental and also comparative analyses, we have considered the dataset made publicly available during the ISBI 2017 challenge. The training part of this set includes 2,000 skin lesion images organized into three classes: C_N (nevus, 1,372), C_M (malignant melanoma, 374), C_{SK} (seborrheic keratosis, 254). The test set of the challenge contained 393 nevus, 90 melanoma, and 117 seborrheic keratosis cases, respectively.

As we have mentioned earlier, transfer learning is a possible way to remedy the possible lack of training data. As another recommendation for the same purpose, we have considered common augmentation techniques including cropping, horizontal flipping and rotating. In this way, we have obtained a training dataset consisting of 14,300 images. As for the skin lesion classes, the cardinality of the nevus subset was increased from 1,372 to 8,200, melanoma from 374 to 4,600, while seborrheic keratosis from 254 to 1,500. All of our training activities have been done on the training dataset, while the testing part was considered for performance evaluation completely independently.

B. Performance evaluation

To evaluate our ensemble-based approach, we have considered the common error measures accuracy (ACC) and area-under-receiver operating characteristic curve (AUC). Notice that, accuracy is a single value corresponding to the 0.5 confidence level on the ROC curve, so AUC gives a bit more detailed description. These measures are calculated from the true/false-positive/negative cases of the classification process, which can be applied to binary classification scenarios. To convert our originally three-classes task to a binary problem, we have followed a one-vs-all approach. That is, we have considered the following three binary classification tasks, and evaluated the performance accordingly: NEV – C_N vs. $C_M \cup C_{SK}$; MEL – C_M vs. $C_N \cup C_{SK}$; SK – C_{SK} vs. $C_N \cup C_M$. The same measurement protocol has been considered also at the ISBI 2017 challenge.

Table I summarizes our quantitative results for the error measures ACC/AUC and three classification scenarios. For a comprehensive evaluation, the average performances (AVG) calculated for the three scenarios are also enclosed. It can be observed that the proposed ensemble-based system outperformed all the individual member CNNs, when they were trained independently, and also the networks composed as an ensemble of any arbitrary two members.

For a better insight to compare the performances of the ensemble and of its members, we present their corresponding ROC curves in Figure 3 calculated for the average (AVG) behaviors on the three skin lesion classes.

For implementation we have used the deep learning toolbox of MATLAB® 2017b (provided by MATHWORKS®), since it has a user friendly developers' environment and optimized for both CPU and GPU usages. Training have been performed using an NVIDIA TITAN X GPU card with 7 TFlops of single precision, 336.5 GB/s of memory bandwidth, 3,072 CUDA cores, and 12 GB memory. The parameters of the architectures were found by a stochastic gradient descent algorithm in 77 epochs. Training times have been found to be 47.5 hours for AlexNet, 147.5 hours for

TABLE I

ACCURACY AND AUC RESULTS ON THE TEST SET.

	AVG_AUC	NEV_AUC	MEL_AUC	SK_AUC	AVG_ACC	NEV_ACC	MEL_ACC	SK_ACC
AlexNet	0.826	0.815	0.786	0.877	0.822	0.772	0.823	0.872
VGGNet	0.813	0.824	0.766	0.849	0.793	0.753	0.802	0.825
GoogLeNet	0.831	0.833	0.802	0.859	0.826	0.772	0.825	0.880
AlexNet & VGGNet	0.821	0.808	0.780	0.874	0.799	0.742	0.810	0.845
GoogLeNet & AlexNet	0.832	0.809	0.799	0.890	0.807	0.745	0.818	0.857
GoogLeNet & VGGNet	0.828	0.814	0.802	0.868	0.812	0.747	0.830	0.858
GoogLeNet & AlexNet & VGGNet	**0.848**	**0.837**	**0.836**	**0.869**	0.838	0.797	0.857	0.862

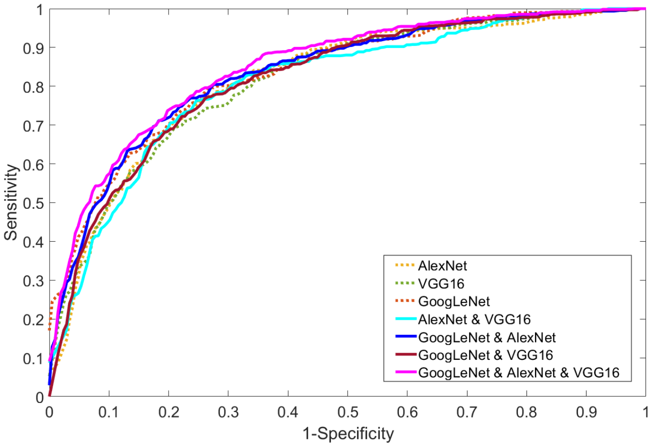

Fig. 3. ROC curves of the proposed ensemble and of its members (AlexNet, VGGNet, GoogLeNet).

VGGNet, 95 hours for GoogLeNet, and 287.5 hours for the ensemble of them.

IV. CONCLUSIONS

Ensemble-based approaches are usually applied to raise the accuracy of the individual members. In this work, we propose a method to organize different convolutional neural networks in a single architecture to be able to train them to co-operate more efficiently. Our investigations were motivated to address the recognition of skin cancer based on dermatoscopy images. Namely, we have tested our technique on a dataset made publicly available during a dedicated ISBI 2017 challenge. As for the experimental results, the proposed ensemble-based method is competitive with outperforming the classification accuracy of the individual deep learning-based ones.

Since both the necessary software and hardware tools are being developed rapidly, a natural way to improve further the performance of our approach is to include more CNN members. Regularization issues regarding the weighing parameters used in the aggregation of the ensemble members to avoid overfitting might also be addressed by adding some corresponding terms to the currently used loss function. As the combined architecture is rather complex, increasing the volume of the training data set can be expected to raise accuracy besides helping to fight with overfitting again.

REFERENCES

[1] "Abcdes of melanoma," https://www.melanoma.org/understand-melanoma/diagnosing-melanoma/detection-screening/abcdes-melanoma, accessed: 2018-01-26.

[2] F. M. Walter, A. T. Prevost, J. Vasconcelos, P. N. Hall, N. P. Burrows, H. C. Morris, A. L. Kinmonth, and J. D. Emery, "Using the 7-point checklist as a diagnostic aid for pigmented skin lesions in general practice: a diagnostic validation study," *Br J Gen Pract*, vol. 63, no. 610, pp. e345–e353, 2013.

[3] P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, and Y. LeCun, "Overfeat: Integrated recognition, localization and detection using convolutional networks," *CoRR*, vol. abs/1312.6229, 2013. [Online]. Available: http://arxiv.org/abs/1312.6229

[4] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 580–587.

[5] N. Zhang, J. Donahue, R. Girshick, and T. Darrell, "Part-based r-cnns for fine-grained category detection," in *European conference on computer vision*. Springer, 2014, pp. 834–849.

[6] 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, pp. 115–118, Jan. 2017. [Online]. Available: http://dx.doi.org/10.1038/nature21056

[7] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Commun. ACM*, vol. 60, no. 6, pp. 84–90, May 2017. [Online]. Available: http://doi.acm.org/10.1145/3065386

[8] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *CoRR*, vol. abs/1409.1556, 2014. [Online]. Available: http://arxiv.org/abs/1409.1556

[9] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. E. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," *CoRR*, vol. abs/1409.4842, 2014. [Online]. Available: http://arxiv.org/abs/1409.4842

[10] "Imagenet," http://www.image-net.org/, accessed: 2018-01-15.

[11] N. C. F. Codella, D. Gutman, M. E. Celebi, B. Helba, M. A. Marchetti, S. W. Dusza, A. Kalloo, K. Liopyris, N. K. 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)," *CoRR*, vol. abs/1710.05006, 2017. [Online]. Available: http://arxiv.org/abs/1710.05006

[12] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *CoRR*, vol. abs/1512.03385, 2015. [Online]. Available: http://arxiv.org/abs/1512.03385

[13] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson, "How transferable are features in deep neural networks?" in *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*, ser. NIPS'14. Cambridge, MA, USA: MIT Press, 2014, pp. 3320–3328. [Online]. Available: