BENCHMARKING THE EFFICIENCY OF DEEP LEARNING METHODS ON THE PROBLEM OF PREDICTING SUBJECTS' AGE BY CHEST RADIOGRAPHS

V. KOVALEV, PhD
Head of the Laboratory of Biomedical Images Analysis

V. LIAUCHUK
Research Assistant of the Laboratory of Biomedical Images Analysis

A. KALINOVSKY
Research Officer of the Laboratory of Biomedical Images Analysis

A. SHUKELOVICH
Junior Scientist of the Laboratory of Biomedical Images Analysis

Biomedical Image Analysis Department, United Institute of Informatics Problems, National Academy of Sciences of Belarus, Republic of Belarus
E-mail: vassili.kovalev@gmail.com

Abstract. This paper presents results that were obtained in comparative study of the efficiency of conventional and Deep Learning methods on the problem of predicting subjects' age by their chest radiographs. A large study group consisting of chest radiographs of 10 000 people was created by random sub-sampling of suitable subjects from the input image repository containing 1.8 million items. The age range was chosen to span from 21 to 70 years. The age prediction was performed by Convolutional Neural Networks AlexNet and GoogLeNet as well as using conventional methods based on Local Binary Patterns and extended co-occurrence matrices as image features followed by kNN, Random Forest, Linear Model, SVM, and Decision Trees classifiers. The conclusion was that the convolutional neural networks greatly outperform conventional methods. It was found that the lowest RMSE error achieved on the task of age prediction using convolutional networks is 5.77 years whereas conventional methods demonstrate on the same data much higher error value of 11.73 years.

The purpose. Recent achievements in biomedical image classification using Deep Learning methods and Convolutional Neural Networks (CNN) give well-grounded promises to become an effective tool in biomedical image analysis [1-4]. Several studies accomplished by authors on the use of CNNs for histology image classification in breast cancer diagnosis [5], lung segmentation [6] and lung lesion detection in computed tomography images of tuberculosis patients [7] confirm the applicability and power of Deep Learning methods in medical imaging domain.

In the context of a difficult choice of the most efficient machine leaning methods and software solutions for medical image analysis the primary goal of this study was to examine abilities of CNNs and to compare them to conventional methods on a large sample of chest radiographs acquired from as many as 10 000 people. The performance comparison was accomplished on the hard problem of predicting patient's age based on their chest X-ray images. Such an examination was performed using both machine learning modes including classification and regression.

Image data. A large database of natively digital chest radiographs containing about 1.8 million items resulted from pulmonary screening of population of a large city was used as input image data repository of this study. Subjects' age was measured in complete years with the precision of one year. A study group consisting of chest radiographs of 10 000 subjects was created by random sub-sampling of suitable subjects from the input image repository. The age range was chosen to span from 21 to 70 years.

In order to create a study group which is well balanced by both age and gender, for every year of life we selected 100 male and 100 female subjects what finally constituted a study group consisting of (100+100) * 50 years = 10 000 subjects. No attention has been given to the subjects' health status.

This particularly means that the created study group mostly represents healthy subjects. Nevertheless, it is still possible that a small fraction of people with certain lung abnormalities at their early stage as well as subjects with some anatomical deviations could be presented in the study group too.

Since the primary goal of this study was not the analysis of chest radiographs as such but benchmarking of Deep Learning methods, original images were preprocessed to avoid unnecessary variability of the image content and to reduce computational expenses. The preprocessing included visual quality assessment, normalization of intensity, and reformatting. The normalization of image intensity was done using commonly known technique of intensity quantiles.

Fig. 1. Examples of chest images used in this study

More specifically, a small fraction of 1% of minimal and maximal values of intensity histograms was saturated and the resultant intensity range was rescaled down to the 0-255. The image crop was performed by cutting off 25% of rows of original image size from the bottom and 5% from the other three sides. Finally, all the images were resized down to 256x256 pixels. Example images of subjects of different age and gender are presented in Fig. 1.

Experimentation outline. At the preliminary stage of preparing experimentation the input images were shuffled within every age year of each gender, i.e. within of each 100 male and 100 female subgroups of every complete year of life. Since there was sufficient amount of image data available, it was decided to subdivide the whole set of 10 000 images into the training and validation sets in the proportion of 70% to 30%. Thus, the training and the test sets consisted of 7000 and 3000 images

respectively. Once created, exactly the same training and validation image sets were used in all the experiments performed in this work. Such a technique guarantees that the results of different experiments are kept comparable in all over the study.

It should be noted that in this work we considered the subjects' age prediction results obtained on the validation image set only. This is because analysis of corresponding results achieved on the training set is typically performed for studying the issues related to the convergence characteristics, influence of certain imbalance or lack of objects of certain classes, solving the problem of overfitting, exploring the necessity of image data augmentation for a proper CNN training, etc. However, all these problems are either atypical for present work or lie outside of the scope of this paper.

Deep Learning methods. Two different approaches were used for prediction of subjects' age based on Deep Learning tools. The first approach makes use of CNNs for a direct age prediction either in regression or in classification mode. In case of classification CNN categorizes an image into one of 50 age classes each of which corresponds to 50 full age years in the range of 21-70. However, the characteristic feature of the first approach is that the final, fully-connected layer of CNN is used as a classifier.

The second approach predicts subjects' age in a similar way. The exception is only that here CNN is employed only for creation of image descriptor. The last pooling layer generated by CNN which contains 1024 elements is used as image descriptor. This layer precedes the fully-connected layer of CNN. The fully-connected layer can be viewed as an "internal" classifier of CNN and which is not used in the second approach. Instead, once created the image descriptor is extracted from CNN and supplied to an "external" classifier which could also be executed in regression and classification mode.

Combination of two training options either in regression or in classification mode and the usage of 6 different classifiers employed in this study resulted in 12 different CNN-related algorithms being examined. The list of classifiers includes internal CNN classifier (i.e., the fully-connected network) along with such external classifiers as kNN, Random Forests, Linear Model, SVM, and Binary Regression Decision Tree. These algorithms are enumerated in Tab. 1 and abbreviated for brevity. Note that the leading letter "D" stands for image descriptor created by CNN on the prediction stage.

Table 1. Twelve age prediction algorithms utilizing CNNs and their abbreviations.

No	Internal/External classifier of CNN	Algorithm abbreviation (executed in 2 modes)
1	Fully Connected Layer of CNN (internal)	CNN
2	kNN (descriptor-based, external)	D-kNN
3	Random Forests (descriptor-based, external)	D-RF
4	Linear Model (descriptor-based, external)	D-LM
5	Support Vector Machines (descriptor-based, external)	D-SVM
6	Binary Regression Decision Tree (descriptor-based, external)	D-DT

Training convolutional networks. Two convolutional networks AlexNet [8] and GoogLeNet [9] were trained under Linux operating system using the Caffe framework from Berkeley Learning and Vision Center which supports GPU acceleration via cuDNN to massively reduce training time [10]. The Caffe framework was chosen from the list of freely available Deep Learning frameworks [11] because of the following reasons:

– Recently, the Caffe framework is one of the best frameworks optimized for GPU-based computing using convolutional networks for image classification.

– The Caffe framework is supported by Nvidia company and it is integrated into the Deep

Learning GPU Training System (DIGITS) interface [12] which provides high level user interface.

– The Caffe framework is well supported by a large academic community which provides voluntary consulting, share pre-trained and trained CNNs for free, etc.

The training was performed on a personal computer equipped with recent Intel® Core™ i7 central processor and two GPU of Nvidia TITAN X type with 3072 CUDA Cores and 12 GB of GDDR5 onboard memory each. The network training parameters were set to the following values:

– Network architectures: AlexNet, GoogLeNet (the first version of Inception architecture from Google).

– Batch size: 32 (the minimum batch size to place network in GPU memory).

– Solver: SGD Caffe solver.

– Number of iterations: 13 140.

– Number of epochs: 60.

– Training set size: 7000 images, 256x256 pixels each.

Age prediction with the help of internal CNN classifier was performed using DIGITS interface. Several Python scripts using PyCaffe interface were written for extracting image descriptors produced by convolutional layer of CNN and inputting them into external classifiers.

The network training time varied from 30 to 60 minutes depending on such parameters as batch size, number of iterations and some other.

Results achieved with Deep Learning methods. The first series of experiments with AlexNet and GoogLeNet convolutional networks have revealed that GoogLeNet slightly but systematically outperforms AlexNet in the quality of subjects' age prediction by chest radiographs. In terms of the Root Mean Square Error (RMSE) of the deviation of predicted age from real one the prediction quality achieved by GoogLeNet was approximately for 0.3-0.8 years better comparing to the one provided by AlexNet. Thus, all the prediction results reported below were obtained with the help of GoogLeNet.

Results of all twelve experiments measured in RMSE error presented in a condensed form on the left panel of Fig. 2.

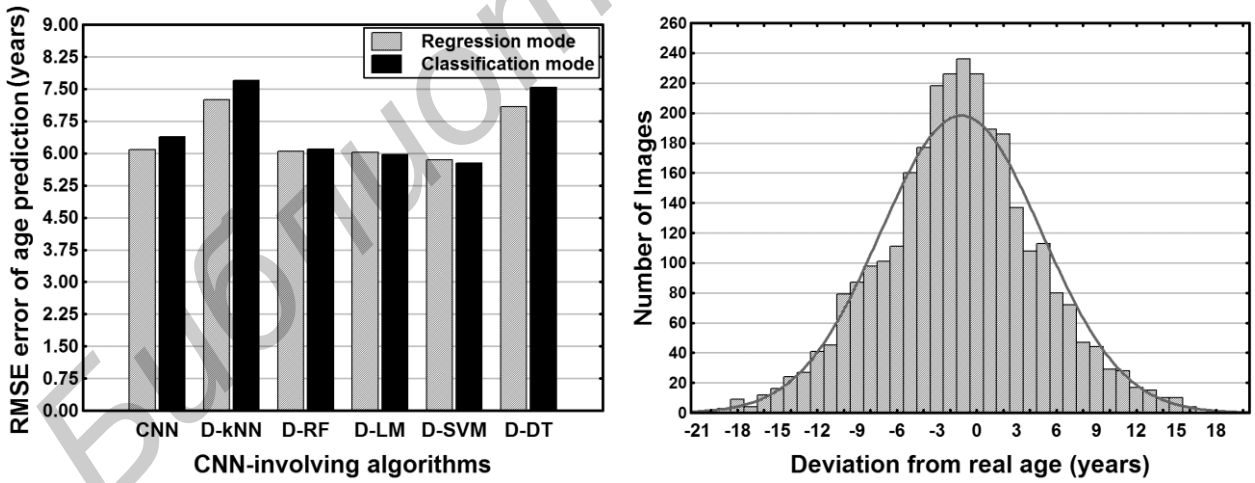

Fig. 2. Results of predicting age of 3000 subjects using convolutional network. Left panel: RMSE error for 12 different algorithms (the lower the better). Right plot: example histogram of residuals.

The right panel provides rather typical example of histogram of residuals in case of running GoogLeNet in regression mode with the native fully-connected layer as a classifier (see the first bar of the left plot).

Local Conclusions. Results of age prediction presented in Fig. 2 allows drawing the following local conclusions.

– Depending on the specific algorithm employing CNN the mean error of age prediction varies in the range from 5.77 years in the best case using of image descriptors created by GoogLeNet in classification mode which were inputted to the external SVM classifier up to 7.25 years of mean error for the same descriptors supplied to kNN classifier.

– The "direct" age prediction by CNN (i.e., without additional manipulation with extracting image descriptors and employing external classifiers) do not provide the best results. However, it is reasonably good with its RMSE value of 6.08 (see the first column of the left plot of Fig 2) compared to the best one of 5.77 achieved by CNN followed by SVM.

– Despite the best value obtained in case of running CNN in classification mode, there can be some tendency observed for better results being achieved when CNN is used in regression mode (see gray bars in Fig. 2). The reason behind could be purely technical such as classification for 50 age classes provide integer age output whereas regression predicts age with the precision of a fraction of year.

The histogram of residuals depicted on the right panel of Fig. 2 demonstrates relatively good fit to the Gaussian distribution what is suggestive for bias-free prediction model.

Conventional methods. Prediction of subjects' age based on chest X-ray images was done by implementation of a three-step procedure comprising of calculation of image descriptors, performing the Principal Component Analysis (PCA) and inputting resultant features into classifiers for age prediction. Below these steps are described in more details.

Step-1: Calculating image descriptors. Since chest images used in this study exhibit typical textural appearance, texture features were employed as image descriptors. Two kinds of texture features were used in order to obtain more extensive, reliable and trustful results. They include commonly known Local Binary Patterns (LBP, [13]) and extended multi-sort and multi-dimensional co-occurrence matrices introduced in [14].

In case of LBP rotation-invariant versions of both uniform and non-uniform types of binary patterns were examined. In case of co-occurrence image descriptors we used 2D version of six-dimensional co-occurrence matrices [15] abbreviated as IIGGAD which fuse intensity (denoted by I) gradient magnitude (G) and anisotropy (angle between gradient vectors A) image features for pixel pairs with inter-pixels distances ranged from 1 to D. It is easy to see that the classical intensity co-occurrence matrices IID with varying inter-pixel spacing can be viewed as a reduced version of the above general case. Next, the particular case of AD type gives us some rotation-invariant version of widely used Histogram of Oriented Gradients (HOG), etc. Technically, all reduced versions can be obtained from IIGGAD by summing up (collapsing) the unnecessary dimensions. It should be remembered also that dimensionality of extended co-occurrence matrices depends on the number of selected features characterizing the pixel pair and not related to image dimensionality (see [15] for more details).

Step-2: Principal Component Analysis. The above "raw" texture features (e.g., elements of co-occurrence matrices) can be very large and contain mutually-correlated elements. Performing PCA dramatically reduce feature space and resulted in uncorrelated principal components which contain essentially the same information because any original variable can be presented as a linear combination of principal components.

Step-3: Age prediction. The principal components obtained on the previous step were considered as image features. Similar to the neural networks case considered in previous sections they were inputted into the same kNN, Random Forests, Linear Model, SVM, Binary Regression Decision Tree classifiers and executed in regression mode to predict subjects' age. In all the occasions the control parameters were kept same to make comparison of results obtained by conventional methods and CNNs straightforward.

Results achieved with conventional methods. Preliminary experiments. In context of this particular study it should be emphasized that in the case of using convolutional network there were al-

most no control parameters notably influencing the quality of image descriptors produced by convolutional layers. However, this is not the case with LBP and extended co-occurrence features. Thus, in order to avoid unnecessary favoritism towards newly immerged Deep Learning tools there were a number of experiments performed for tuning control parameters of LBP and extended co-occurrence image descriptors.

A total of 18 variants of rotation-invariant LBP descriptors were examined including 9 uniform and 9 non-uniform versions with the radius of local circular neighborhood of 1, 3 and 5 pixels and number of pixels compared to the central one equal to 8, 12 and 16. As a result it was found that depending on combination of these parameters the RMSE error varied in the range from 12.93 to 17.30 years.

Similar investigation was performed for extended co-occurrence matrices. A total of 32 variants of IIGGAD, AD, and GGD matrices were evaluated with the number of intensity bins equal to 8, 16, 24 and 32, gradient magnitude bins equal to 8 and 16, number of angle bins 12 and 16 as well as inter-pixel distances of 1, 3 and 5 pixels. Note that not all possible combinations of control parameters were tested. Also, it was found that contrary to a wide spread believe an increase of intensity resolution (number of intensity bins) towards 256 not necessarily increases quality of final results. Finally, the exploratory experiments with extended co-occurrence descriptors have revealed that RMSE error of age prediction varied between 13.12 and 16.84 years what is similar to the range obtained when using LBP.

Principal Component Analysis. Instead of selecting the most prominent variants from 18 particular LBP and 32 co-occurrences image descriptors described above they were merged into two corresponding data tables as subsets of variables and supplied to PCA. As a result we got two sets of image features obtained with the help of LBP and extended co-occurrence image descriptors. In all the occasions the output principal components were selected so that they explain 99% of variation of raw image descriptors of training image dataset. As a result, the number of selected principal components varied from several dozen up to one hundred.

It is important to note that the output principal components derived from raw LBP and extended co-occurrence image descriptors were also mutually correlated. This is not surprising because these two kinds of features describe quantitatively the content of the same image set. Finally, they were inputted together into the "second" PCA for obtaining a set of joint image features combining advantages of both LBP and extended co-occurrence image descriptors.

Final results. The results obtained based on LBP, extended co-occurrence and joint image features using 5 different classifiers are presented in Fig. 3. Similar to age prediction results obtained with convolutional networks, they measured in RMSE error. The right plot of Fig. 3 depicts histogram of residuals obtained based on the joint image descriptor using SVM classifier and illustrates proportions of different prediction errors.

Local Conclusions. Results of age prediction by the subjects' chest X-ray images using conventional methods which are presented in Fig. 3 allow to draw the following local conclusions.

– The use of joint image descriptors always provide better result comparing to LBP and extended co-occurrence alone (black bars in Fig. 3 are lower for all 5 classifiers).

– Formally, the best result with minimal error value of 11.73 years was achieved based on joint image descriptors using SVM classifier. However, this is only for 0.04 and 0.15 years better than prediction provided with the help of Linear Model and Random Forest respectively.

– The LBP descriptors alone (see gray bars) slightly outperform extended co-occurrence matrices. For instance, in case of Random Forest classifier the error value of LBP is 12.39 against 12.61 years achieved by co-occurrence what corresponds to subtle difference of 0.22 years. The highest difference of errors in age prediction of 0.60 years between these two descriptors is observed in case of kNN (16.24 vs. 16.84) whereas with Decision Tree classifier LBP and co-occurrence descriptors demonstrate same error of 15.50 years.

– Random Forest, Linear Model and SVM classifiers doing always better than kNN and

Decision Trees.

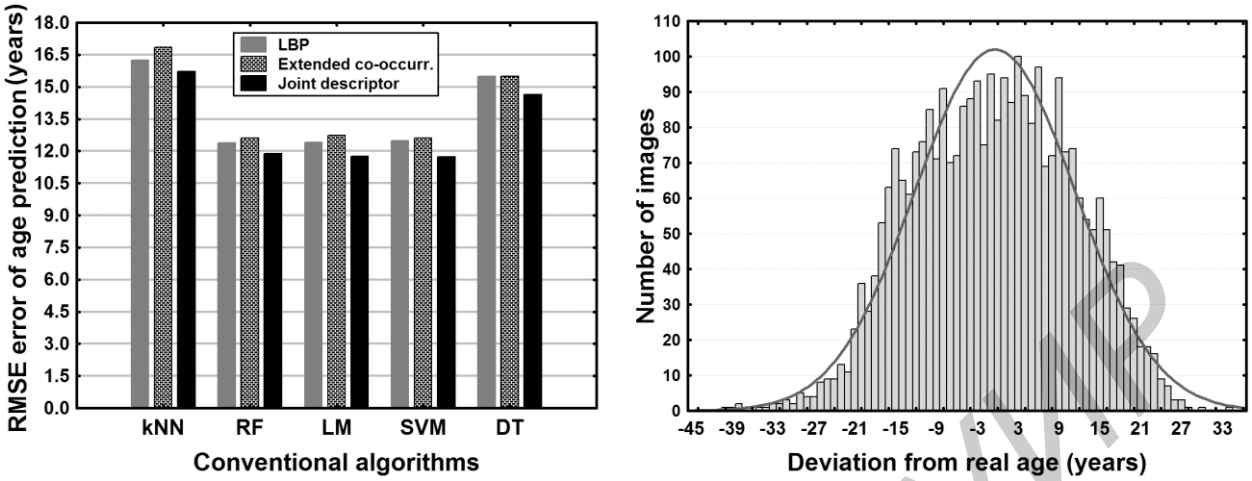

Fig. 3. Results of predicting age of 3000 subjects using conventional algorithms. Left panel: RMSE error obtained using 3 types of image descriptors and 5 classifiers (the lower the better). Right plot: example histogram of residuals.

As can be seen from the histogram of residuals which was calculated as the difference between real and predicted age, there is a tendency for overestimating subjects' age when using conventional approaches (see the shift to negative values of histogram of Fig. 3).

Conclusion. Results obtained with this comparative study of the efficiency of conventional and Deep Learning methods on the problem of predicting subjects' age by their chest radiographs allow drawing the following conclusions.

(1) The convolutional neural networks greatly outperform conventional methods. The lowest RMSE error achieved on the task of age prediction using convolutional networks is 5.77 years whereas conventional methods demonstrate on the same data much higher error value of 11.73 years.

(2) The worst error value of 7.25 years obtained in 12 experiments with neural networks is still far better than the best result of 11.73 year error obtained in 15 experiments following conventional approach. In general, results obtained with convolutional network approximately twice as good comparing to the conventional methods examined in this study.

(3) Results produced by convolutional layers during the network training can be used as compact image features describing the image content.

Acknowledgements. This study was partly supported by the National Institute of Allergy and Infectious Diseases, National Institutes of Health, U.S. Department of Health and Human Services, USA through the CRDF project OISE-16-62631-1.

References

[1]. Ravi D., Wong C., Deligianni F., Berthelot M., Andreu-Perez J., Lo B., Yang G.Z. Deep Learning for Health Informatics, IEEE Journal on Health and Biomedical Informatics, vol. 21, No 1, 2017, pp. 4-21.

[2]. Deep Learning for Medical Image Analysis, Zhou S. K, Greenspan H., Shen D. (Eds), Academic Press, ISBN: 9780128104088, 2017, 458 p.

[3]. Litjens G., Kooi T., Bejnordi B.E., Setio A.A.A., Ciompi F., Ghafoorian M., van der Laak J.A.W.M. , van Ginneken B., Sánchez C.I. A Survey on Deep Learning in Medical Image Analysis, arXiv:1702.05747, 2017, 34 p.

[4]. Kovalev V., Kalinovsky A., and Kovalev S. Deep Learning with Theano, Torch, Caffe, TensorFlow, and deeplearning4j: which one is the best in speed and accuracy? In: XIII Int. Conf. on Pattern Recognition and Information Processing, 3-5 October, Minsk, Belarus State University, 2016, pp. 99-103.

[5]. Kovalev V., Kalinovsky A., Liauchuk V. Deep Learning in Big Image Data: Histology image classification for breast cancer diagnosis, In: Big Data and Advanced Analytics, Proc. 2nd International Conference, BSUIR, Minsk, June 2016, pp. 44-53.

[6]. Kalinovsky A. and Kovalev V. Lung image segmentation using Deep Learning methods and convolutional neural networks . In: XIII Int. Conf. on Pattern Recognition and Information Processing, 3-5 October, Minsk, Belarus State University, 2016, pp. 21-24.

[7]. Liauchuk V., Kovalev V., Kalinovsky A., Tarasau A., Gabrielian A., Rosenthal A. Examining the ability of convolutional neural networks to detect lesions in lung CT images. Journal of Computer Assisted Radiology and Surgery, CARS-2017 International Congress, Barcelona, 20-25 June 2016 (in press).

[8]. Krizhevsky A., Sutskever I., Hinton G. Imagenet classification with deep convolutional neural networks, In: Advances in neural information processing systems, 3-8 December, USA, 2012, pp. 1097-1105.

[9]. Szegedy, Christian, et al. Going deeper with convolutions. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 1-9.

[10].Jia Y., Shelhamer E., Donahue J., Karayev S., Long J., Girshick R., Darrell T. Caffe: Convolutional architecture for fast feature embedding. In: Proceedings of the ACM International Conference on Multimedia, ACM, 2014, pp. 675-678.

[11].Kovalev V., Kalinovsky A., and Kovalev S. Deep Learning with Theano, Torch, Caffe, TensorFlow, and deeplearning4j: which one is the best in speed and accuracy? In: XIII Int. Conf. on Pattern Recognition and Information Processing, 3-5 October, Minsk, Belarus State University, 2016, pp. 99-103.

[12].https://devblogs.nvidia.com/parallelforall/deep-learning-computer-vision-caffe-cudnn/ Last visited March 2017.

[13].Pietikäinen M., Hadid A., Zhao G., Ahonen T. Computer Vision Using Local Binary Patterns. Volume 40, Springer-Verlag, London, 2011, ISBN 978-0-85729-747-1, DOI 10.1007/978-0-85729-748-8.

[14].Kovalev V. and Petrou M. Multidimensional co-occurrence matrices for object recognition and matching, Graphical Models and Image Processing, vol. 58, No. 3, pp. 187-197, 1996.

[15].Kovalev V.A., Kruggel F., Gertz H.-J., and von Cramon D.Y. Three-dimensional texture analysis of MRI brain datasets, IEEE Transactions on Medical Imaging, vol. 20, No. 5, pp. 424-433, 2001.