Medical Image Analysis 36 (2017) 41–51

Contents lists available at ScienceDirect

Medical Image Analysis

journal homepage: www.elsevier.com/locate/media

Deep learning for automated skeletal bone age assessment in X-ray images☆

C. Spampinato[a,*], S. Palazzo[a], D. Giordano[a], M. Aldinucci[b], R. Leonardi[c]

[a] Pattern Recognition and Computer Vision (PeRCeiVe) Lab, Department of Electrical, Electronics and Computer Engineering, University of Catania, Viale Andrea Doria, 6 - 95125 - Catania, Italy
[b] Computer Science Department, University of Torino, Corso Svizzera, 185 - 10149 - Torino, Italy
[c] Department of Orthodontics, University of Catania, Via Santa Sofia, 78 - 95125 - Catania, Italy

ARTICLE INFO

Article history:
Received 2 March 2016
Revised 10 October 2016
Accepted 12 October 2016
Available online 29 October 2016

Keywords:
Convolutional neural networks
Deep learning for medical images
Tanner–Whitehouse
Greulich and Pyle

ABSTRACT

Skeletal bone age assessment is a common clinical practice to investigate endocrinology, genetic and growth disorders in children. It is generally performed by radiological examination of the left hand by using either the Greulich and Pyle (G&P) method or the Tanner–Whitehouse (TW) one. However, both clinical procedures show several limitations, from the examination effort of radiologists to (most importantly) significant intra- and inter-operator variability. To address these problems, several automated approaches (especially relying on the TW method) have been proposed; nevertheless, none of them has been proved able to generalize to different races, age ranges and genders.

In this paper, we propose and test several deep learning approaches to assess skeletal bone age automatically; the results showed an average discrepancy between manual and automatic evaluation of about 0.8 years, which is state-of-the-art performance. Furthermore, this is the first automated skeletal bone age assessment work tested on a public dataset and for all age ranges, races and genders, for which the source code is available, thus representing an exhaustive baseline for future research in the field.

Beside the specific application scenario, this paper aims at providing answers to more general questions about deep learning on medical images: from the comparison between deep-learned features and manually-crafted ones, to the usage of deep-learning methods trained on general imagery for medical problems, to how to train a CNN with few images.

1. Introduction

Skeletal bone age assessment is a procedure used in pediatric radiology for both diagnostic and therapeutic investigations (White, 1963) of endocrinology problems (Carty, 2002), children growth and genetic disorders (Poznanski et al., 1978). It is usually performed by radiological examination of the left hand, because of the discriminant nature of bone ossification stages of the non-dominant hand, and then compared to chronological age: a discrepancy between the two values indicates abnormalities. The analysis of left-hand X-ray images is widely used for the evaluation of bone maturity due to simplicity, minimum radiation exposure, and the availability of multiple ossification centers. Although there is no standard clinical procedure, two clinical methods are mostly employed: (1) Greulich and Pyle (1959) (G&P) and (2) Tanner–Whitehouse (TW) (Carty, 2002). The G&P method is the approach used by 76% of radiologists (because of its simplicity and speed) and is based on the comparison between the whole X-ray scan and a reference atlas. Nevertheless, it suffers greatly from intra- and inter-observer variability (inter-observer differences range from 0.07 to 1.25 years and intra-observer differences from 0.11 to 0.89 year; see Berst et al., 2001). TW-based methods, TW2 and TW3 (Carty, 2002; particularly used in the U.S.), analyze specific bones, instead of the whole hand as in the G&P method, whose standard maturity varies according to age population, race and gender. In particular, TW methods take into account a set of specific regions of interest (ROIs) divided into epiphysis/metaphysis ROIs (EMROIs) and carpal ROIs (CROIs), as in Fig. 1. The development of each ROI is divided into discrete stages, and each stage is given a letter (A,B,C,D, ..., I) corresponding to a numerical score which varies according to race and sex. By adding the scores of all ROIs, an overall bone maturity score is achieved. Though being less used because of the time needed to perform the analysis, TW methods

☆ For additional information about paper and authors, see http://perceive.dieei.unict.it.

* Corresponding author.
E-mail addresses: cspampin@dieei.unict.it (C. Spampinato), simone.palazzo@dieei.unict.it (S. Palazzo), dgiordan@dieei.unict.it (D. Giordano), aldinuc@di.unito.it (M. Aldinucci), rleonardi@unict.it (R. Leonardi).

http://dx.doi.org/10.1016/j.media.2016.10.010
1361-8415/

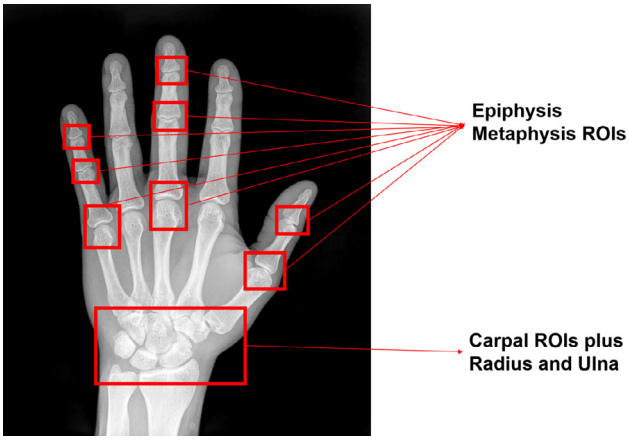

Fig. 1. Regions of Interest (ROIs) used in the Tanner Whitehouse method: Epiphysis/Metaphysis ROIs and Carpal ROIs.

yield a more accurate estimation than the G&P method (King et al., 1994) and their modular structure make them suitable for automation (see Section 2). These methods, like most of those for automatic medical image analysis, try to replicate a clinical approach relying mainly on domain expert feedback. In such cases, the main arising question is: are the visual features (from a computer vision perspective) identified by domain experts or used in the clinical practices suitable to build automated methods?

To answer this question, in this paper we investigate several deep-learning approaches to perform automatic skeletal bone age assessment and compare the automatically-learned deep features to the ones employed in the TW clinical methods. Also, most of the existing automated methods either do not release their code or are tested on non-public X-ray datasets, making those results unreproducible and a systematic comparison not possible. In order to provide a proper and comprehensive baseline (inexistent so far) for automated skeletal bone age assessment, we tested our methods on a public X-ray scan dataset, covering all age ranges, genders and races, and make our source code available.

Beside the contributions to the research on automated skeletal bone age assessment, this work also aims to investigate general questions related to deep-learning in the medical imaging field such as: (1) What are the generalization capabilities of deep-learning approaches, trained on general imagery, for medical image analysis tasks? (2) Training deep-learning methods usually needs large datasets, often not available in the medical domain: how to perform CNN training with a small amount of images? (3) How much do deep-learned features differ from those employed in the clinical practices? (4) Can deep-learned features support the development of new clinical investigation methods?

The remainder of the paper is organized as follows: Section 2 presents a critical analysis of automated skeletal bone age assessment methods as well as of deep-learning ones in relation to the above questions. Section 3 describes our deep-learning approaches for skeletal bone age assessment. Section 4 shows the performance evaluation of the tested methods and a comparison to the state-of-the-art, while Section 5 presents the conclusions and future directions.

2. Related work

The goal of this paper is to perform automatic skeletal bone age assessment (BAA) using deep-learning methods. Thus, we will first review existing automated bone age assessment methods analyzing their advantages and limitations and then deep-learning–based approaches for medical images according to the questions raised in Section 1. The majority of the automated bone maturity assessment methods using left-hand X-ray scans builds on the TW methods since they are more prone, given their modular structure, to automation than G&P one. Automated BAA approaches reproducing the TW method can be mainly classified based on whether they use image processing or knowledge-based techniques and a thorough review can be found at Mansourvar et al. (2013). Most of the image processing–based ones date back to the 2000s; Pietka et al. (2001); 2003) propose an EMROI segmentation method by an ad-hoc phalangeal distance extraction and TW stage assignment is carried out by a fuzzy classifier. Deformable models have been also largely adopted for EMROI (Davis et al., 2012; Lin et al., 2012; Giordano et al., 2007) and CROI (Hsieh et al., 2007a; Adeshina et al., 2014; Zhang et al., 2007) segmentation. EMROIs and CROIs have been also used together for accurate and robust bone age assessment (Hsieh et al., 2007; Giordano et al., 2010; Seok et al., 2016).

Knowledge-based approaches, mainly relying on decision rules (Seok et al., 2016) or fuzzy logic (Gertych et al., 2007; Aja-Fernandez et al., 2004) or Bayesian networks (Mahmoodi et al., 2000), represent the alternative approach for automated bone maturity assessment. However, model initialization and generalization hampered the achievement of good-enough results for all these methods. While most of the above methods are based on the TW method, recently, Thodberg et al. (2009) proposed *BoneXpert*, which performs automatic age assessment through a unified model of TW and G&P methods. However, BoneXpert needs high-quality X-ray scans to obtain reliable results; in fact, it rejects images with poor quality or abnormal bone structure, for which cases the analysis needs to be manual.

Table 1 report performance of some state of the methods, in terms of either mean absolute error (MAE) or mean square error (MSE), as well as the employed dataset (if publicly available) and its size, the age range and race they were devised for. Unfortunately, most of the above methods were tested on private X-ray datasets (except for Gertych et al. (2007), which released the dataset used in our work) or do not provide source code, thus their results are not reproducible or usable as baselines.

Despite some methods yield very accurate results, all the existing methods suffer from two main limitations:

- Most of the above methods operate only with X-ray scans of caucasian subjects younger that 10 years, when bones are not yet fused, thus easier than in older ages where bones (especially, the carpal ones) overlap.
- All of them assess bone age by extracting features from the bones (either EMROIs or CROIs or both of them) commonly adopted by the TW or G&P clinical methods, thus constraining low-level (i.e., machine learning and computer vision) methods to use high-level (i.e., coming directly from human knowledge) visual descriptors. This semantic gap usually limits the generalization capabilities of the devised solutions, in particular when the visual descriptors are complex to extract as in the case of mature bones.

The deep-learning approaches presented in this paper, instead, aim at overcoming these problems by learning visual features, regardless of age ranges and races, that may facilitate the assessment process. To the best of our knowledge, deep-learning or Convolutional Neural Networks (CNNs) have not been applied yet for automated skeletal bone age assessment (except for a recent work — FingerNet, in Lee et al., 2015 — that, however, does not perform any age assessment but only finger joint detection in radiographs) while shallow neural networks or support vector machines have been adopted for image segmentation and age assessment (Kashif et al., 2016; Mansourvar et al., 2015; Bocchi et al., 2003; Tristan-Vega and Arribas, 2008; Zhang et al., 2007; Liu et al., 2008). In Mansourvar et al. (2015) bone age regression is performed

Table 1
Performance in terms of either MAE (mean absolute error) or MSE (mean squared error) of state-of-the-art methods. *NS* stands for not specified, while *Cau* for Caucasian race. *Performance taken from Giordano et al. (2010). **Performance taken from Giordano et al. (2016).

Method	Dataset	images	Age	Race	MAE	MSE
Giordano et al. (2016)	Private	360	0–6	Cau	0.39	–
Kashif et al. (2016)	Private	1100	0–18	All	0.60	–
Seok et al. (2016)	Private	135	NS	NS	–	0.19
Mansourvar et al. (2015)	Private	1100	0–18	All	–	0.22
Giordano et al. (2010)	Private	106	0–10	Cau	0.75	–
Thodberg et al. (2009)	Private	1559	7–17	NS	–	0.42 (G&P)
					–	0.80 (TW2)
Gertych et al. (2007)	Public	1400	0–18	All	2.15	–
Hsieh et al. (2007)*	Private	106	0–10	Cau	1.41	–
Pietka et al. (2003)**	Private	360	0–6	Cau	1.93	–
Pietka et al. (2001)**	Private	360	0–6	Cau	2.41	–

by means of extreme learning machines over a large dataset of 1100 X-ray images, achieving promising results. In particular, the approach is based on content image retrieval concepts, and feature extraction is carried out by means of principal component analysis. In Zhang et al. (2007); Liu et al. (2008), bone segmentation is performed by particle swarm optimization and bone age regression by neural networks reaching an average matching between automatic and manual evaluations over 95%. Despite the claimed performance, these methods show generalization limitations, as the employed features are mainly low-level ones not able to describe complex structures as bones. On the contrary, deep-learning solutions (in particular, Convolutional Neural Networks), have been successfully used in a multitude of other medical imaging applications in past and recent years. One of the first applications of CNNs on medical images is Lo et al. (1995) for both classification of lung nodules in chest X-ray scans and detection of microcalcifications in mammograms. A three-layer (two convolutional and one fully-connected) CNN was trained on small image datasets (e.g., for lung nodules classification the authors used 55 chest X-ray scans — 25 positives and 30 negatives), suitably augmented, achieving high classification performance. The automatically high-level learned features showed characteristic morphological variations of the investigated body structures. Similar methods were proposed in Sahiner et al. (1996); Lo et al. (2002), for distinguishing between healthy and cancerous tissues in mammographies (the learned visual features were not reported). After a break of about two decades, CNNs were re-discovered (mainly because of the recent progress in computing hardware) for automated visual analysis and, of course, applied to medical imaging tasks. In Ciresan et al. (2012), a Deep Neural Network is applied on stacks of electron microscopy images of brain slices for pixel-wise classification of neuronal membrane tissue as well as for image pre-processing (i.e., foveation and non-uniform sampling). In Malon and Cosatto (2013), manually-designed nuclear features combined to learned deep features extracted by CNNs were used with good results for mitotic figure recognition in regions of interest of hematoxylin and eosin stained tissue. This work also shows how CNNs are able to handle the variety of appearances of mitotic figures and decrease sensitivity to manually-crafted features. Roth et al. (2015) proved that training CNNs provides better chances of capturing discriminative features for anatomy-specific classification in CT images than when using hand-crafted features.

CNNs have been also employed in a kind of "bagging" configuration as in Roth et al. (2014); Ciompi et al. (2015), where random representations of VOI are used to train a CNN for lymph node detection. CNN outputs are then averaged to compute the final classification probability for each considered VOI.

The usage of off-the-shelf CNN features has been also recently investigated (Wolterink et al., 2015; van Ginneken et al., 2015). For example, in van Ginneken et al. (2015), CNNs trained on general imagery are employed for classification of pulmonary nodules in computed tomography scans yielding accurate classification results. Analogously, Ciompi et al. (2015) use OverFeat (a pre-trained convolutional neural network) for morphology feature extraction in automatic classification of pulmonary nodules in CT images.

The effects of deep learning design and training techniques for medical images is another important issue. Havaei et al. (2015) explore several CNN architectures (e.g., two-pathway, cascade architectures, etc.) and different training schemes, such as Maxout (Goodfellow et al., 2013) hidden units and Dropout (Srivastava et al., 2014) regularization, for brain segmentation in MRI images. The results showed that cascaded CNNs perform sensibly worse than one- or two-pathway–CNN and regularization enhances the segmentation results.

In conclusion, the current state of the art shows how deep learning for medical imaging is an active area and, in this paper, we intend to contribute it through a "deep" analysis of CNNs applied to a classic X-ray image analysis problem.

3. Convolutional neural networks for automated skeletal bone age assessment

In this section, we describe the deep-learning approaches and configurations we employed to perform automated skeletal bone age assessment: (1) using off-the-shelf features extracted from CNNs trained on general imagery; (2) fine-tuning pre-trained CNN models; and (3) building an ad-hoc CNN to take into account the peculiarities (e.g., including nonrigid hand/bone deformation) of the tackled X-ray images. The first approach does not require any model adaptation and exploits the global description capability of the last layers of a trained CNN to capture common yet distinctive low- and middle-level visual patterns (Razavian et al., 2014). The second approach aims at suiting an existing model (trained on a different dataset) to the task at hand, skipping most of kernel learning but finely adapting the whole network to an unseen kind of images. The last approach, instead, aims at building from scratch a CNN designed ad-hoc and training it with task-specific images. All the three approaches share a similar architecture (see Fig. 2) which is made up of two consecutive networks: (1) a *convolutional network* with a variable number of convolutional layers (depending on the employed models) aiming at extracting low and middle-level visual features, and (2) a *regression network*, which consists of a variable set of fully-connected layers followed by a one-neuron output layer providing bone age estimate.

3.1. Off-the-shelf CNN feature extraction

A common approach which has been proved to work successfully (Razavian et al., 2014) consists in using a pre-trained (on another dataset) CNN as a feature extractor, by providing an input image to the network and reading the output vector of a fully-connected layer, which can then be used to train a simpler

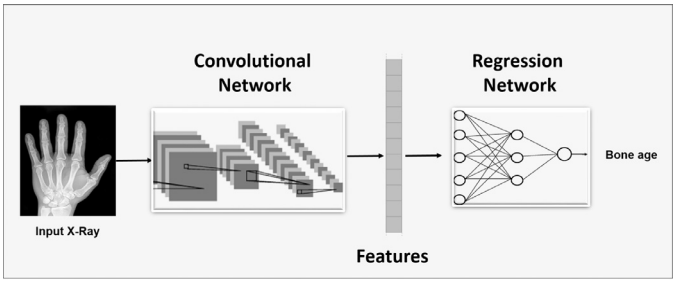

Fig. 2. General architecture of deep learning methods for bone age assessment. It consists of a) a **convolutional network** consisting of an arbitrary number of convolutional layers (that can be derived from pre-trained CNN models or can be designed from scratch) for feature extraction; and b) a **regression network** consisting of a set of fully connected layers (generally one or two) and a linear scalar output layer providing the bone age estimate.

classifier, e.g., MLP or SVM. This approach is especially useful if the dataset on which the network was trained is similar to the target dataset, as the patterns learned by the convolution kernels are likely to be equally discriminative. Although this is not our case, since X-ray images are markedly different from the real world images commonly employed for training general-purpose CNNs (e.g., the ImageNet dataset), it is a convenient alternative since it only requires to train the regression network, which takes as input, features extracted from one of the deepest layers of off-the-shelf CNN models. More specifically, in this scenario, the convolution network of our model architecture consists of a pre-trained CNN from which the final layers are removed so that the output of a fully-connected layer is exposed, while the regression network is made up of one or two layers with ReLU nonlinearity, followed by a linear scalar output layer. Of course, during training, only the regression network's weights are updated, while the convolution network is simply used as a feature extractor. As off-the-shelf CNNs we considered three common pre-trained CNNs as feature extractors: OverFeat (Sermanet et al., 2014), GoogLeNet (Szegedy et al., 2015) and OxfordNet (Simonyan and Zisserman, 2014).

In Section 4 the performance of different settings of each CNN model at different layer depths is reported to identify the most suitable off-the-shelf model for the problem at hand. This experiment also allowed us to identify the best-performing architecture for the regression network, used in the next experiment.

3.2. Fine-tuning CNNs trained on general imagery

Another common alternative for re-using existing models (computationally expensive to train from scratch) is to use a pre-trained model as initial conditions and "fine-tune" it on the target dataset. This technique greatly speeds up training and helps prevent overfitting, as the starting solution may be already close to a good local minimum and unlikely to be "moved" too far. Analogously to the previous case, we fine-tuned OverFeat, GoogLeNet and OxfordNet on our X-ray image dataset. The corresponding model used for bone age estimation was, therefore, designed by using a pre-trained convolutional model (without the softmax classification layer) as convolutional network of our model architecture, followed by best-performing regression network identified in the off-the-shelf feature extraction experimental scenario.

3.3. BoNet: an ad-hoc CNN for skeletal bone age assessment

Our last model consists of a CNN — called *BoNet* — trained from scratch on the X-ray scan dataset. The advantage of training a new CNN over fine-tuning an existing one lies in the possibility to tweak the network architecture to the type of images at hand: optimizing the network for grayscale images, reducing the number of layers, letting the network learn specific filters instead of adapting more generic ones.

We tested several network architecture for *BoNet*, then chose the one which achieved the best accuracy as the final model for evaluation (see Section 4.5). The generic layout shared by all architectures is as follows:

- As convolutional network, we employed a sequence of the following modules:
  - A pre-trained convolutional layer obtained as a grayscale version (through channel-wise averaging) of OverFeat's first convolutional layer. We chose to adapt the first-layer kernels as they encode common application-independent low-level visual patterns, while reducing the risk of overfit due to the small number of available X-ray images
  - A variable number of convolutional layers: unlike the previous layer, these are initialized randomly and trained on dataset images only.
  - An optional *deformation layer* (Jaderberg et al., 2015), which learns an adaptive geometric transformation to apply to input images (or, equivalently, to feature maps) in order to provide invariance to affine warping, thus accounting for nonrigid object deformation. This new layer learns (in an end-to-end fashion) to compute the parameters of an affine transformation which scales, translates, rotates and crops an input image in order to tackle object nonrigid deformations. In detail, the input image is first processed by a *localization network* (typically, a small CNN) which computes a 6-dimensional parameter vector $\Theta = \{\theta_1, \cdots, \theta_6\}$, defining the transformation to apply. Then, a grid of points $\{(x_i^s, y_i^s)\}$ is sampled from the input image, and the grid points' new locations $\{(x_i^t, y_i^t)\}$ are computed by applying the transformation defined by Θ:

$$\begin{pmatrix} x_i^t \\ y_i^t \end{pmatrix} = \begin{bmatrix} \theta_1 & \theta_2 & \theta_3 \\ \theta_4 & \theta_5 & \theta_6 \end{bmatrix} \begin{pmatrix} x_i^s \\ y_i^s \\ 1 \end{pmatrix} \quad (1)$$

    This transformation causes the initial grid to localize a warped crop of the original image. A *sampler* module then resamples the points in the crop to produce a set of output maps matching the input's size, thus making the module transparent to subsequent layers. It is important to understand that the transformation parameters Θ are computed for each input image: the module learns to generate such parameters according to the characteristics of each input. This layer can be optionally inserted either before the first convolutional layer — thus acting as a pre-processing layer directly on the input image — or after one of the deeper convolutional layers (as recently done in Johnson et al. (2015)) — in order to operate on deformations of higher-level hand features. Fig. 3 shows an example of transformations applied by such module when used directly on input images: the resulting images tend to be more evenly stretched and cropped.
- As regression network, we adopted a single fully connected layer (with number of neurons decided experimentally) followed by a linear scalar layer which outputs the estimated skeletal age for the input image.

Each "convolutional layer" includes a ReLU nonlinearity and a max-pooling layer. Different tested network architectures vary by the presence/position of the deformation layer, the number of convolutional layers, and the number of feature maps. According to the results shown in Section 4.5 the architecture of the best performing CNN, shown in Fig. 4, consists of five convolutional layers, a deformation layer located after the fourth convolutional layer, a

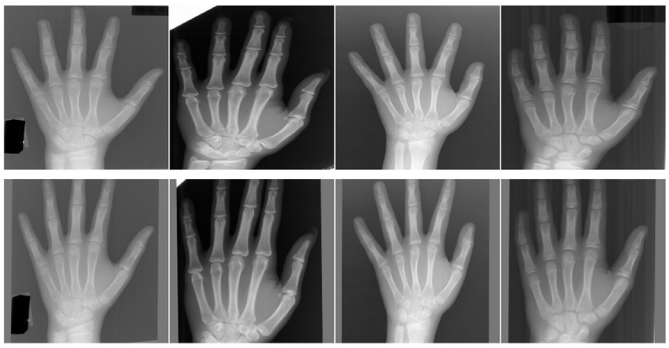

Fig. 3. Example outputs from the deformation layer when applied before convolutional layers. Top row: original images; bottom row: resulting images, more evenly stretched and cropped.

2048-neuron fully-connected layer followed by the single-neuron layer providing the estimate.

4. Experimental results

In this section, we first describe the dataset and metrics used for performance analysis and then report the results obtained by the tested models compared to the state of the art; finally, a comparative analysis between deep-learned and hand-crafted features is also presented.

4.1. Dataset

The assessment of the correctness and the accuracy of the CNN-based methods, presented in the previous section, was carried out on the Digital Hand Atlas Database System[1] (Gertych et al., 2007), a public and comprehensive X-ray dataset for automated skeletal bone age benchmarking. The dataset contains 1391 X-ray left-hand scans of children of age up to 18 years old, divided by gender and race. Each X-ray scan comes with two bone age values, provided by two expert radiologists. The distribution of images among these categories is shown in Table 2.

Table 2
Distribution of the dataset images by gender, race and age. Key for race abbreviations: *A.*: asian; *B.*: black; *C.*: caucasian; *H.*: hispanic. *Two of the images in the female/black category in the 5 years age group were not readable. The numbers reported in the table do not consider them.

	Gender and race							
	Male				Female			
Age	A.	B.	C.	H.	A.	B.	C.	H.
0	2	5	3	4	1	4	3	1
1	5	5	5	5	5	5	5	5
2	5	5	5	5	5	5	5	5
3	5	5	5	5	5	5	5	5
4	5	5	5	5	5	5	5	5
5	9	9	10	9	8	9*	7	10
6	6	7	8	9	6	9	7	10
7	7	9	9	10	7	9	8	10
8	5	10	10	10	9	11	9	9
9	7	10	7	10	7	9	8	10
10	14	15	11	12	15	12	12	14
11	15	15	14	14	12	10	13	15
12	15	15	13	15	14	15	15	15
13	15	15	12	15	15	15	13	15
14	12	14	10	14	13	12	11	14
15	10	10	10	10	10	10	10	10
16	10	10	10	10	10	10	10	10
17	10	10	10	10	10	10	10	10
18	10	10	10	10	10	10	10	10
Totals	**168**	**184**	**167**	**182**	**167**	**175**	**166**	**183**
		700				**691**		
				1391				

4.2. Training details

Images in the dataset were rescaled so that the smallest dimension was 256 pixels, keeping aspect ratio, and normalized to have zero mean and unitary standard deviation.

[1] http://www.ipilab.org/BAAweb/

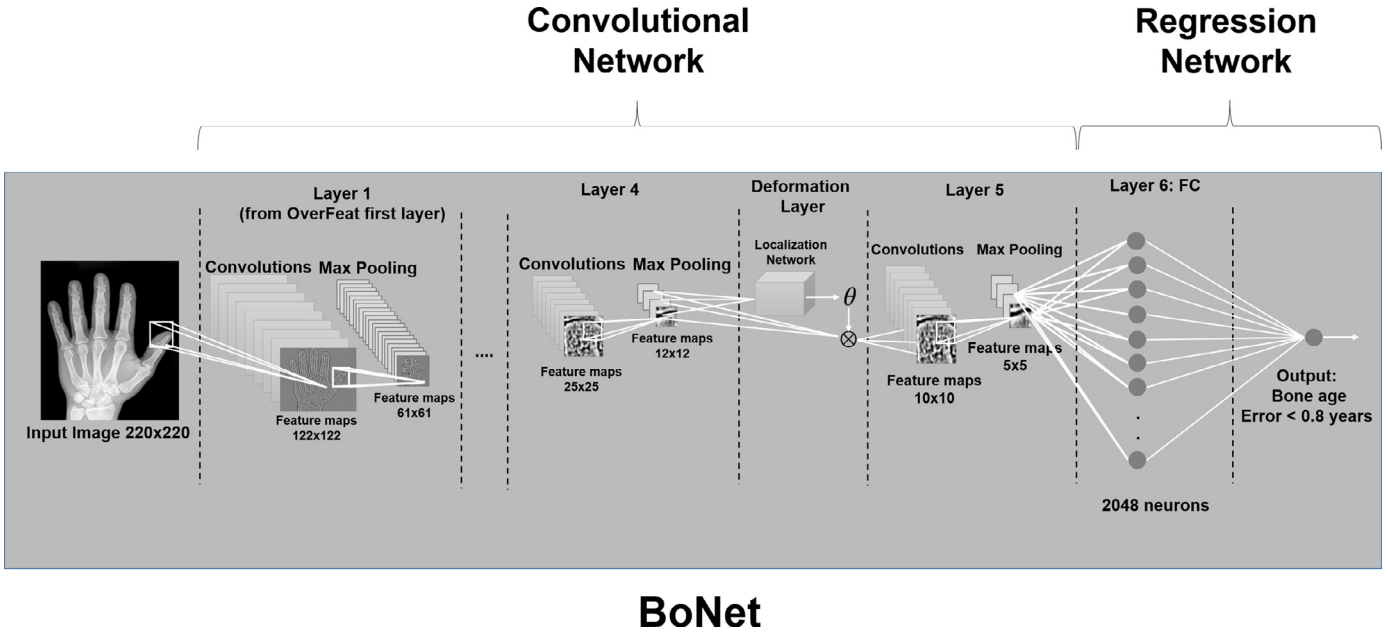

Fig. 4. Overview of BoNet Architecture: it consists of five convolutional and pooling layers, aiming at the extracting low and middle-level visual features, and one deformation layer facing bone nonrigid deformation and two fully connected layers for bone age regression.

Table 3
Mean Absolute Error (MAE) achieved using three different "off-the-shelf" CNN models for feature extraction at different depth of layers and with different regression settings. For each CNN model, we underline the best results over all the considered layer depths. OVFeat stands for OverFeat, GNet for GoogLeNet and OxNet for OxfordNet.

CNN	Reading 1				Reading 2			
	Regression network				Regression network			
	128	128+128	256	256+128	128	128+128	256	256+128
OvFeat								
FC1	1.26	1.22	1.26	1.25	1.25	1.23	1.24	1.26
FC2	1.38	1.34	1.38	1.35	1.36	1.33	1.37	1.38
GNet								
FC	1.18	1.16	1.19	1.18	1.17	1.15	1.18	1.18
OxNet								
FC1	1.30	1.31	1.43	1.29	1.37	1.33	1.45	1.36
FC2	1.38	1.42	1.38	1.47	1.39	1.44	1.38	1.42

During training, data augmentation was performed by extracting 10 uniformly-spaced crops from each input image, with the size of the crop depending on the CNN's expected input size: 221 × 221 for OverFeat, 299 × 299 for GoogLeNet, and 224 × 224 for OxfordNet and *BoNet*. During validation and test, since convolutional layers can process arbitrary-sized input, and fully-connected layers can be implemented as 1 × 1 convolutions, we provided a whole rescaled image as input to the CNN, obtaining as output a 2D map of age estimations corresponding to different regions of the input image, which we averaged to compute the overall age estimation.

All models were trained using mini-batch stochastic gradient descent on an MSE loss function, with batch size set to 4, learning rate initially set to 0.002 and decreased with decay 0.0002 (applied at each mini-batch), and momentum factor set to 0.9. We used the average of the two expert readings as target for models' training. Training was carried out for 150 epochs over the augmented dataset.

4.3. Evaluation criteria

For all methods under investigation, we evaluated the performance over the whole Digital Hand Atlas Database using 5-fold cross validation: for each fold we computed the mean absolute error (MAE) between the two manual readings of each X-ray scan and the corresponding estimated age, given as model output.

In the final evaluation between the best performing models based on pre-trained networks and the finalized *BoNet* architecture, we provide a detailed analysis of the accuracy of the trained models on subgroups of the dataset by age, race and sex. Note that the results computed for each subgroup are obtained using the models trained on the whole dataset.

In the tables, best results are underlined for better readability.

4.4. Performance of pre-trained convolutional neural networks

In this section we report the results achieved by our deep learning approaches. First, we compared the performance of the three off-the-shelf CNNs, namely, OverFeat, GoogLeNet, and OxfordNet (at different depth of layers) when used as feature extractors (as part of the convolutional network shown in Fig. 2) and combined to four regression networks composed by, respectively, a single fully-connected layer with 128 neurons — "128" —, two fully-connected layers with 128 neurons each — "128+128" —, one fully-connected layer with 256 neurons — "256" —, and two fully-connected layers with 256 and 128 neurons — "256+128", all followed by a single neuron layer providing bone age estimate. The results are shown in Table 3 and provide some initial interesting considerations: (1) performance at the first (i.e., FC1 for OverFeat and OxfordNet) fully-connected layers are higher than at deeper layers (i.e., FC2 in OverFeat and OxfordNet); (2) using two 128-neuron layers for the regressor yielded generally the best performance compared to the other regression settings; (3) off-the-shelf CNNs were already able to provide a satisfactory accuracy, especially if compared to state-of-the-art methods (see Table 1).

Table 4
Comparison, in terms of Mean Absolute Error (MAE), between off-the-shelf (**OTS**) OverFeat, OxfordNet and GoogLeNet and their fine-tuning (**FT**) on the Digital Hand Atlas Database.

CNN model	Reading 1		Reading 2		Average		Gain
	FT	OTS	FT	OTS	FT	OTS	
OverFeat	1.00	1.22	0.94	1.23	0.97	1.22	21%
GoogLeNet	0.86	1.16	0.79	1.15	0.82	1.15	28%
OxfordNet	0.88	1.31	0.79	1.33	0.83	1.32	37%

For our fine-tuning experiments, we plugged our best-performing regressor (two fully connected layers with 128 neurons each) at the end of each of the three considered pre-trained CNN models; the achieved results when fine-tuning over the Digital Hand Atlas dataset are given in Table 4, and also compared to the best performance obtained by the off-the-shelf CNN models according to Table 3. Fine-tuning GoogLeNet achieved the best performance (average MAE over the two readings was 0.82) enhancing the performance of about 28% with respect to the off-the-shelf model.

4.5. Performance of BoNet

In order to define the final architecture of *BoNet* and to compare its performance to pre-trained CNNs as well as to state-of-the-art methods, we performed several tests using different configurations defined by the following parameters:

- Number of convolutional layers.
- Number of feature maps per convolutional layer.
- Presence and position of the deformation layer. In particular, for a given number of convolutional layers and for the best performing configuration (as number of feature maps), we tested two options for deformation layer position: (a) before all convolutional layers, operating directly on the input X-rays images (configurations indicated with a "Yes" in Table 5) in order to face nonrigid hand deformation, and (b) before the last convolutional layer (indicated with "Yes: *X*" in Table 5, i.e,. after the *X*th convolutional layer) to address nonrigid deformation of smaller regions (e.g., bones). For models with fewer than four convolutional layers, we only tested the deformation layer applied directly on input images, since the networks were

Table 5
Performance of different network configurations for BoNet. The "Deform." column specifies the position of the deformation layer: "No" indicates that the layer is not present; "Yes" indicates that the deformation layer is the first layer in the model and operates directly on the input image; "Yes: *X*" indicates that the deformation layer has been inserted after a specific convolutional layer (typically, the second-to-last). The presence and position of deformation layer was tested for the best performing configuration (as number of feature maps) for a given number of convolutional layers. "R" stands for "Reading".

			MAE		
Deform.	Conv.	# feat. maps	R1	R2	Average
No	2	96, 1024	1.31	0.90	1.10
No	2	96, 2048	1.19	0.87	1.03
Yes	2	96, 2048	2.11	1.10	1.60
No	3	96, 512, 1024	1.10	0.83	0.96
No	3	96, 2048, 1024	1.09	0.84	0.96
No	3	96, 2048, 2048	0.93	0.90	0.91
Yes	3	96, 2048, 2048	0.97	0.89	0.93
No	4	96, 512, 1024, 1024	1.17	0.86	1.01
No	4	96, 1024, 1024, 1024	1.08	0.84	0.96
No	4	96, 2048, 2048, 2048	0.94	0.78	0.86
Yes	4	96, 2048, 2048, 2048	0.92	0.90	0.91
Yes: 3	4	96, 2048, 2048, 2048	0.88	0.80	0.84
No	5	96, 512, 1024, 1024, 1024	1.01	0.82	0.91
No	5	96, 2048, 1024, 1024, 1024	0.91	0.81	0.86
No	5	96, 2048, 2048, 2048, 2048	0.95	0.80	0.87
Yes	5	96, 2048, 2048, 2048, 2048	0.92	0.89	0.91
Yes: 4	5	96, 2048, 1024, 1024, 1024	0.80	0.79	0.79
No	6	96, 512, 1024, 1024, 1024, 1024	1.22	1.17	1.19
No	6	96, 1024, 1024, 1024, 2048, 2048	1.01	1.14	1.07
No	6	96, 2048, 2048, 2048, 2048, 2048	1.07	1.05	1.06
Yes	6	96, 2048, 2048, 2048, 2048, 2048	1.55	1.06	1.30
Yes: 5	6	96, 2048, 2048, 2048, 2048, 2048	0.96	0.98	0.97

too shallow and positioning it at the penultimate layer was unlikely to affect the performance.

For all experiments where it was employed, our deformation layer scaled down the input images by a factor of two and processed them through the localization network consisting of three convolutional layers with 20 feature maps each (with 5 × 5 kernels) to estimate the deformation parameters. All layer's weights are learned from scratch.

In all models, the first convolutional layer employed 7 × 7 kernels; the second one 5 × 5 kernels; and the following layers (if present) 3 × 3 kernels. After the cascade of convolutional layers, the regression network was made up of a single 2048-neuron fully-connected layer followed by the single-neuron regressed estimate.

Table 5 shows the results for several tested configurations. The best CNN architecture consisted of five convolutional layers, one deformation layer after the fourth convolutional layer, and the aforementioned regression network. This version was chosen as the finalized *BoNet* model.

While satisfactory results can be achieved even with a smaller number of convolutional layers (e.g., four layers reached an average MAE between the two readings of 0.86), it is interesting to notice that the best results are obtained when using a number of convolutional layers (five) equal to OverFeat — although the latter has an additional fully-connected layer. This suggests, conversely to recent findings Shin et al. (2016), that a combination of a partial initialization of low-level kernels and from-scratch training of high-level kernels can be more effective the fine-tuning existing networks.

As for the deformation layer, there was a performance increase (on average 7% in networks with more than three convolutional layers) when inserting it deeper in the network, while the performance sensibly worsened when it was added at the beginning. This may be due to: (a) operating on the original image has the effect of "smoothing out" some structural differences (as shown in Fig. 3) which may be discriminative for the subsequent analysis; (b) bone age assessment strictly depends on specific region of interests (ROIs) (identified at the deeper layers, see Fig. 6) rather than on the whole image; thus, it is more important to handle the nonrigid deformation (due to different races, gender, etc.) of such ROIs than of the hands' (which are less subject to nonrigid transformations in this specific application).

We finally compared the performance of the final *BoNet* and the fine-tuned GoogLeNet and OxfordNet (we excluded OverFeat since, among the tested CNN models, it performed the worst) over subgroups of the dataset, defined according to age, sex and race. For this evaluation, we split the dataset into the 0–9 and 10–18 age ranges, and computed MAE for each gender, race and age group. The split into the two age ranges is important to understand methods' performance since bone age assessment is usually easier for young children (indeed most of the state-of-the-art methods operate only on younger patients as shown in Section 2) and more complex for adolescents, as bones fuse together and growth becomes slower. The resulting MAE values are shown in Table 6: it may appear surprising that in many cases the models yielded a higher accuracy on the 10–18 age range than the 0–9 one, where the visual differences between bone formation is supposed to be more evident. However, this is easily explained as a dataset bias, since the number of images in the 10–18 age range is larger (see Table 2).

Overall, *BoNet* outperformed, on average, all the other CNN-based solutions, reaching an average MAE across all races, genders and age ranges of 0.79. Furthermore, the results suggest that with gray-level X-ray images, many convolutional layers — as in the case of GoogLeNet — are not strictly necessary. However, CNNs pre-trained on general imagery performed fairly well (much better than state-of-the-art automated methods — see next subsection) on Digital Hand Atlas Database with performance comparable to those achieved by custom CNNs. In addition, the performance of CNN-based approaches did not vary too much with age ranges, genders and races as, instead, in state-of-the-art methods.

Table 6
Comparison, in terms of Mean Absolute Error (MAE), between BoNet and fine-tuned OverFeat, OxfordNet and GoogLeNet on age/race/sex subgroups of the Digital Hand Atlas Database.

Group	Reading 1			Reading 2		
	BoNet	GoogLeNet	OxfordNet	BoNet	GoogLeNet	OxfordNet
M-all-0-18	0,77	0,94	0,87	0,77	0,82	0,86
M-all-0-9	0,91	1,04	0,90	0,81	0,81	0,86
M-all-10-18	0,68	0,87	0,85	0,74	0,83	0,86
M-asi-0-18	0,64	0,74	0,80	0,51	0,54	0,70
M-asi-0-9	0,44	0,71	0,82	0,35	0,57	0,69
M-asi-10-18	0,74	0,76	0,79	0,59	0,53	0,70
M-blk-0-18	0,73	0,98	0,84	0,70	0,90	0,83
M-blk-0-9	0,89	1,17	0,97	0,66	0,81	0,91
M-blk-10-18	0,63	0,86	0,75	0,73	0,96	0,77
M-cau-0-18	0,84	0,97	0,81	0,95	0,76	0,91
M-cau-0-9	1,11	0,95	0,65	1,16	0,54	0,81
M-cau-10-18	0,66	0,99	0,93	0,80	0,92	0,98
M-his-0-18	0,86	1,03	1,02	0,90	1,05	0,98
M-his-0-9	1,08	1,24	1,12	0,96	1,24	0,98
M-his-10-18	0,69	0,89	0,95	0,86	0,90	0,98
F-all-0-18	0,81	0,82	0,85	0,78	0,76	0,78
F-all-0-9	0,85	0,81	0,84	0,82	0,77	0,79
F-all-10-18	0,79	0,83	0,87	0,75	0,76	0,77
F-asi-0-18	0,85	0,81	0,94	0,84	0,80	0,82
F-asi-0-9	0,91	0,94	0,80	0,88	0,98	0,70
F-asi-10-18	0,81	0,73	1,03	0,81	0,69	0,89
F-blk-0-18	0,94	0,92	0,98	0,73	0,80	0,81
F-blk-0-9	0,87	0,67	0,90	0,71	0,48	0,78
F-blk-10-18	0,98	1,09	1,03	0,74	1,01	0,83
F-cau-0-18	0,66	0,78	0,68	0,67	0,75	0,69
F-cau-0-9	0,56	0,71	0,65	0,68	0,77	0,71
F-cau-10-18	0,72	0,83	0,70	0,65	0,74	0,68
F-his-0-18	0,79	0,78	0,81	0,88	0,71	0,78
F-his-0-9	1,02	0,93	0,98	0,99	0,89	0,94
F-his-10-18	0,63	0,68	0,71	0,80	0,60	0,68

Table 7
Comparison of performance in terms of MAE between our methods and state-of-the-art ones over the Digital Hand Atlas Database.

Method	Reading		Average
	1	2	
Giordano et al. (2016)	1.92	1.73	1.82
Giordano et al. (2010)	2.46	2.38	2.42
Gertych et al. (2007)	2.60	1.70	2.10
Hsieh et al. (2007)	2.78	2.37	2.57
Fine-tuned OxfordNet	0.88	0.79	0.83
Fine-tuned GoogLeNet	0.86	0.79	0.82
BoNet	0.80	0.79	0.79

4.6. Comparison with the state of the art

In the last two decades, many automated skeletal bone age assessment methods, mainly based on the Tanner–Whitehouse procedure, have been proposed with accuracies (in terms of MAE) varying from 0.37 to 2.63 years (see Table 1). Nevertheless, all of these methods are either tested on private datasets or their source code is not available, thus the claimed results are not reproducible. Despite, at a first glance of Table 1, the accuracy of the tested deep learning approaches may seem lower of some methods, when we performed a comparison[2] over the whole Digital Hand Atlas Database, we observed that both *BoNet* and fine-tuning pre-trained CNN models outperformed significantly them, as shown in Table 7. Most importantly, *BoNet* was able to perform effectively and with high accuracy bone age assessment for all races, genders and age ranges from 0 to 18 years.

4.7. Hand-crafted features vs deep-learned features

Most of the existing methods, especially those based on the TW clinical methods, operate on local image patches (or regions of interest) while our deep-learning methods process images as a whole and do not need any pre-processing to extract some specific regions of interest. Nevertheless, training a CNN on specific images means identifying specific low- and middle-level features that, given the encouraging performance of *BoNet*, can be also useful for clinical procedures.

Fig. 5 shows a comparison between the regions of interest employed by clinicians when performing the TW methods, and the ones corresponding to the most active neurons at different layers (visualized according to Zeiler and Fergus (2014)). It can be noted that some of the regions of interest matched (e.g., the ones on first, third and fifth finger), while others did not (e.g., the ones on second and forth finger). Also, while TW methods highlight the importance of all carpal ROIs (Hsieh et al., 2007a; Adeshina et al., 2014; Zhang et al., 2007), the ROIs corresponding to the most active neurons (in all the considered images) showed that only tiny parts of carpal bones were used. Radius and ulna (especially their reciprocal distance), instead, seemed to be significant parameters for bone age assessment in all cases. When we analyzed all deep-learned features (not only those corresponding to the most active neurons), we observed that all TW ROIs were indeed learned, although most of them (as in the discussed carpal bones case) were not particularly significant (i.e., the corresponding neurons were the less activated in all the images) for the final performance. This suggests that some of the features currently employed by

[2] For this comparison, we used only our previous implementations of Giordano et al. (2016), Giordano et al. (2010) and Gertych et al. (2007) and Gertych et al. (2007), which was already tested on the same dataset. We did not test other methods since the source code was not available and we did not want to introduce any implementation bias in the evaluation.

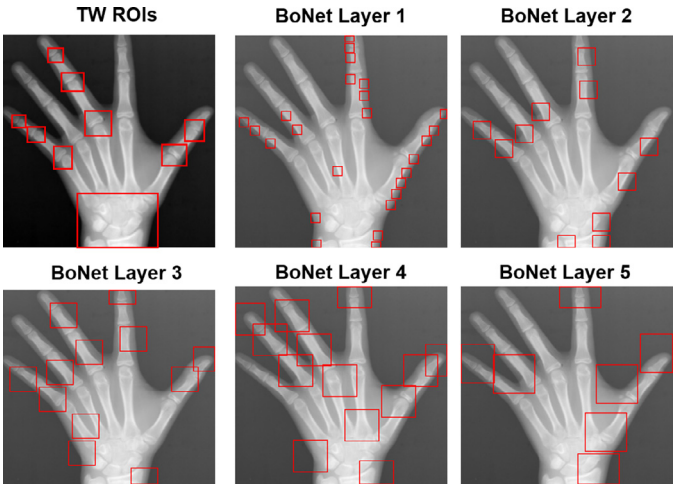

Fig. 5. ROIs employed in the TW methods vs ROIs corresponding to the BoNet learned features (shown in Fig. 6).

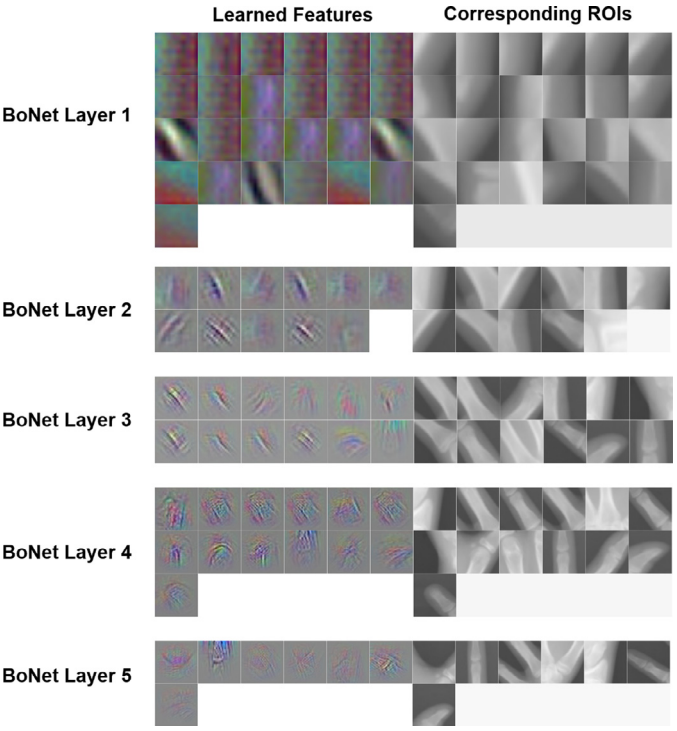

Fig. 6. Deep Learned Features for each BoNet layer and corresponding ROIs.

clinicians might not be necessary, while others should be taken into account. These findings, whose require, of course, a deeper investigation, may have a great impact on the clinical procedures that have been used on the last three/four decades.

4.8. Implementation details

All models and evaluations described in this work were implemented in Torch.[3] Pre-trained models[4] or public implementations[5] were employed where applicable. Our experiments were run on a machine with a 3 GHz CPU and 16 GB RAM, equipped with an Nvidia Titan X GPU. Training times ranged from 240 s per epoch (two convolutional layers) to 1730 s per epoch (six convolutional layers with deformation layer). Test times, instead, were about 5–7 ms for all models considering network forward times only, i.e., excluding data loading and image pre-processing time, which are generally longer than the actual processing in a realistic hospital setting.

4.9. Available resources

The Digital Hand Atlas Database System is available at http://www.ipilab.org/BAAweb/ , while the *BoNet* source code, the code for visualizing deep learned features (both written in Torch) as well as the deep learned features are available at http://perceive.dieei.unict.it. The Overfeat, GoogLeNet and OxfordNet CNNs can be found, respectively, in Sermanet et al. (2014), Szegedy et al. (2015) and Simonyan and Zisserman (2014).

5. Discussion

In this paper we have investigated the application of deep learning to medical images, and in particular for automated skeletal bone age assessment using X-ray images. We have tested several existing pre-trained convolutional neural networks (OverFeat, GoogLeNet and OxfordNet) on a dataset of about 1400 X-ray images and proved that deep learning solutions, even trained on general imagery, are able to cope effectively with all possible cases of automated skeletal bone age assessment (off-the-shelf GoogLeNet achieved an average MAE of 1.15, comparable to state-of-the-art performance). Fine-tuning pre-trained CNN models over the Digital Hand Atlas dataset resulted in an average performance gain of about 30%, outperforming methods that exploit low-level visual descriptors based on clinical procedures. We also designed and trained from scratch several a custom CNN — *BoNet* —, which proved to be the most effective and robust solution in assessing bone age across races, age ranges and gender. In particular, *BoNet* consists of five convolutional layers, one deformation layer before the last convolutional layer to face nonrigid object deformation, one 2048-fully connected layer followed by a single output neuron.

The final considerations that can be drawn from testing several deep learning models for automated bone age assessment are:

- *Effective training a CNN from scratch with a limited number of images is possible* by employing hybrid configurations with the first layer (since it encodes low-level and general visual features) initialized from a pre-trained network, and the following ones trained from scratch using application-specific images to learn discriminative visual features;
- *Using as many convolutional layers as possible does not necessarily mean high performance*. Indeed, in our case, the best performance was achieved when employing only five convolutional layers. Furthermore, when testing off-the-shelf CNN features, better performance were obtained at the least deep fully-connected layers in all the adopted models (see Table 3). This, in our opinion, uncovers an important aspect of CNN in the medical imaging domain, i.e., despite deep learned features can be usually transferred and reused, their level of aggregation (i.e., layers-depth) is strongly domain-dependent. This also explains why, in our case, we found that *BoNet* outperformed fine-tuned CNN models while recent studies (Shin et al., 2016) (tested on thoraco-abdominal lymph node detection and interstitial lung disease classification) found out the opposite.
- *Exploiting a layer able to cope with nonrigid object deformation enhances significantly performance*. Enriching our CNN with a

[3] http://torch.ch/

[4] OverFeat: https://github.com/jhjin/overfeat-torch; OxfordNet: https://gist.github.com/ksimonyan/3785162f95cd2d5fee77; GoogLeNet: https://github.com/Moodstocks/inception-v3.torch.

[5] Spatial Transformer Networks: https://github.com/qassemoquab/stnbhwd.

module able to learn (in an end-to-end fashion) geometric transformations to tackle nonrigid deformation led to increased performance.
- *Deep-learned features highlighted that some of currently-employed hand-crafted features might not be necessary*. We observed that all hand-crafted features were automatically learned by *BoNet*. However, most of them (especially carpal ones) did not influence the final performance.

Beside the above findings, to the best of our knowledge, this is the first work for automated skeletal bone assessment tested on a public dataset for all possible cases and whose source code is publicly released, thus serving as a proper baseline for future research in the field.

Acknowledgment

This work has been partially supported by the NVidia GPU Research Center at University of Torino in Italy.

References

Adeshina, S., Lindner, C., Cootes, T., 2014. Automatic segmentation of carpal area bones with random forest regression voting for estimating skeletal maturity in infants. In: Electronics, Computer and Computation (ICECCO), 2014 11th International Conference on, pp. 1–4.
Aja-Fernandez, S., De Luis-Garcia, R., Martin-Fernandez, M.A., Alberola-Lopez, C., 2004. A computational tw3 classifier for skeletal maturity assessment. a computing with words approach. J. Biomed. Inform. 37, 99–107.
Berst, M.J., Dolan, L., Bogdanowicz, M.M., Stevens, M.A., Chow, S., Brandser, E.A., 2001. Effect of knowledge of chronologic age on the variability of pediatric bone age determined using the Greulich and Pyle standards. AJR Am. J. Roentgenol 176, 507–510.
Bocchi, L., Ferrara, F., Nicoletti, I., Valli, G., 2003. An artificial neural network architecture for skeletal age assessment. In: Image Processing, 2003. ICIP 2003. Proceedings. 2003 International Conference on, Vol. 1. I–1077–80 vol.1
Carty, H., 2002. Assessment of skeletal maturity and prediction of adult height (tw3 method). Journal of Bone and Joint Surgery, British Volume 84-B, 310–311. 3rd edition. edited by j. m. tanner, m. j. r. healy, h. goldstein and n. cameron. pp 110. london, etc: W. b. saunders, 2001. isbn: 0-7020-2511-9. 69.95.
Ciompi, F., de Hoop, B., van Riel, S.J., Chung, K., Scholten, E.T., Oudkerk, M., de Jong, P.A., Prokop, M., van Ginneken, B., 2015. Automatic classification of pulmonary peri-fissural nodules in computed tomography using an ensemble of 2d views and a convolutional neural network out-of-the-box. Med. Image Anal. (pp.–)
Ciresan, D., Giusti, A., Gambardella, L.M., Schmidhuber, J., 2012. Deep neural networks segment neuronal membranes in electron microscopy images. In: Advances in neural information processing systems, pp. 2843–2851.
Davis, L., Theobald, B.-J., Bagnall, A., 2012. Automated bone age assessment using feature extraction. In: Yin, H., Costa, J., Barreto, G. (Eds.), Intelligent Data Engineering and Automated Learning - IDEAL 2012. In: Lecture Notes in Computer Science, Vol. 7435. Springer Berlin Heidelberg, pp. 43–51.
Gertych, A., Zhang, A., Sayre, J., Pospiech-Kurkowska, S., Huang, H.K., 2007. Bone age assessment of children using a digital hand atlas. Comput. Med. Imaging Graph 31, 322–331.
van Ginneken, B., Setio, A.A., Jacobs, C., Ciompi, F., 2015. Off-the-shelf convolutional neural network features for pulmonary nodule detection in computed tomography scans. In: Biomedical Imaging (ISBI), 2015 IEEE 12th International Symposium on, pp. 286–289.
Giordano, D., Kavasidis, I., Spampinato, C., 2016. Modeling skeletal bone development with hidden markov models. Comput. Methods Programs Biomed. 124, 138–147.
Giordano, D., Leonardi, R., Maiorana, F., Scarciofalo, G., Spampinato, C., 2007. Epiphysis and metaphysis extraction and classification by adaptive thresholding and dog filtering for automated skeletal bone age analysis. Conf. Proc. IEEE Eng. Med. Biol. Soc. 2007, 6552–6557.
Giordano, D., Spampinato, C., Scarciofalo, G., Leonardi, R., 2010. An automatic system for skeletal bone age measurement by robust processing of carpal and epiphysial/metaphysial bones. Instrumentation and Measurement, IEEE Transactions on 59, 2539–2553.
Goodfellow, I.J., Warde-Farley, D., Mirza, M., Courville, A.C., Bengio, Y., 2013. Maxout networks. In: Proceedings of the 30th International Conference on Machine Learning, ICML 2013, Atlanta, GA, USA, 16–21 June 2013, pp. 1319–1327.
Greulich, W.W., Pyle, S.I., 1959. Radiographic atlas of skeletal development of the hand and wrist. Am. J. Med. Sci. 238, 393.
Havaei, M., Davy, A., Warde-Farley, D., Biard, A., Courville, A.C., Bengio, Y., Pal, C., Jodoin, P., Larochelle, H., 2015. Brain tumor segmentation with deep neural networks. CoRR abs/1505.03540.
Hsieh, C.W., Jong, T.L., Chou, Y.H., Tiu, C.M., 2007a. Computerized geometric features of carpal bone for bone age estimation. Chin. Med. J. 120, 767–770.
Hsieh, C.-W., Jong, T.-L., Tiu, C.-M., 2007. Bone age estimation based on phalanx information with fuzzy constrain of carpals. Med. Biol. Eng. Comput. 45, 283–295.
Jaderberg, M., Simonyan, K., Zisserman, A., Kavukcuoglu, K., 2015. Spatial transformer networks. CoRR abs/1506.02025. URL: http://arxiv.org/abs/1506.02025
Johnson, J., Karpathy, A., Li, F., 2015. Densecap: fully convolutional localization networks for dense captioning. CoRR abs/1511.07571. URL: http://arxiv.org/abs/1511.07571
Kashif, M., Deserno, T.M., Haak, D., Jonas, S., 2016. Feature description with sift, surf, brief, brisk, or freak? a general question answered for bone age assessment. Comput. Biol. Med. 68, 67–75.
King, D.G., Steventon, D.M., O'Sullivan, M.P., Cook, A.M., Hornsby, V.P., Jefferson, I.G., King, P.R., 1994. Reproducibility of bone ages when performed by radiology registrars: an audit of Tanner and Whitehouse II versus Greulich and Pyle methods. Br. J Radiol 67, 848–851.
Lee, S., Choi, M., soo Choi, H., Park, M.S., Yoon, S., 2015. Fingernet: deep learning-based robust finger joint detection from radiographs. In: Biomedical Circuits and Systems Conference (BioCAS), 2015 IEEE, pp. 1–4.
Lin, H.-H., Shu, S.-G., Lin, Y.-H., Yu, S.-S., 2012. Bone age cluster assessment and feature clustering analysis based on phalangeal image rough segmentation. Pattern Recognit. 45, 322–332.
Liu, J., Qi, J., Liu, Z., Ning, Q., Luo, X., 2008. Automatic bone age assessment based on intelligent algorithms and comparison with TW3 method. Comput. Med. Imaging Graph 32, 678–684.
Lo, S.-C.B., Chan, H.-P., Lin, J.-S., Li, H., Freedman, M.T., Mun, S.K., 1995. Artificial convolution neural network for medical image pattern recognition. Neural Networks 8, 1201–1214.
Lo, S.-C.B., Li, H., Wang, Y., Kinnard, L., Freedman, M.T., 2002. A multiple circular path convolution neural network system for detection of mammographic masses. Medical Imaging, IEEE Trans. 21, 150–158.
Mahmoodi, S., Sharif, B.S., Chester, E.G., Owen, J.P., Lee, R., 2000. Skeletal growth estimation using radiographic image processing and analysis. IEEE Trans. Inf. Technol. Biomed. 4, 292–297.
Malon, C.D., Cosatto, E., 2013. Classification of mitotic figures with convolutional neural networks and seeded blob features. J. Pathol. Inform. 4.
Mansourvar, M., Ismail, M.A., Herawan, T., Raj, R.G., Kareem, S.A., Nasaruddin, F.H., 2013. Automated bone age assessment: motivation, taxonomies, and challenges. Comput. Math. Methods Med. 2013, 391626.
Mansourvar, M., Shamshirband, S., Raj, R.G., Gunalan, R., Mazinani, I., 2015. An automated system for skeletal maturity assessment by extreme learning machines. PLoS ONE 10, e0138493.
Pietka, E., Gertych, A., Pospiech, S., Cao, F., Huang, H.K., Gilsanz, V., 2001. Computer-assisted bone age assessment: image preprocessing and epiphyseal/metaphyseal roi extraction. IEEE Trans. Med. Imaging 20, 715–729.
Pietka, E., Pospiech-Kurkowska, S., Gertych, A., Cao, F., 2003. Integration of computer assisted bone age assessment with clinical pacs. Comput Med Imaging Graph 27, 217–228.
Poznanski, A.K., Hernandez, R.J., Guire, K.E., Bereza, U.L., Garn, S.M., 1978. Carpal length in children–a useful measurement in the diagnosis of rheumatoid arthritis and some concenital malformation syndromes. Radiology 129, 661–668.
Razavian, A.S., Azizpour, H., Sullivan, J., Carlsson, S., 2014. Cnn features off-the-shelf: an astounding baseline for recognition. In: Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition Workshops CVPRW '14. IEEE Computer Society, Washington, DC, USA, pp. 512–519.
Roth, H., Lu, L., Seff, A., Cherry, K., Hoffman, J., Wang, S., Liu, J., Turkbey, E., Summers, R., 2014. A new 2.5d representation for lymph node detection using random sets of deep convolutional neural network observations. In: Golland, P., Hata, N., Barillot, C., Hornegger, J., Howe, R. (Eds.), Medical Image Computing and Computer-Assisted Intervention âÈ¯ÃÀÃ»È¯ÃÀÃ» MICCAI 2014. In: Lecture Notes in Computer Science, Vol. 8673. Springer International Publishing, pp. 520–527.
Roth, H.R., Lee, C.T., Shin, H., Seff, A., Kim, L., Yao, J., Lu, L., Summers, R.M., 2015. Anatomy-specific classification of medical images using deep convolutional nets. CoRR abs/1504.04003.
Sahiner, B., Chan, H.-P., Petrick, N., Wei, D., Helvie, M., Adler, D.D., Goodsitt, M.M., et al., 1996. Classification of mass and normal breast tissue: a convolution neural network classifier with spatial domain and texture images. Medical Imaging, IEEE Trans. 15, 598–610.
Seok, J., Kasa-Vubu, J., DiPietro, M., Girard, A., 2016. Expert system for automated bone age determination. Expert Syst. Appl. 50, 75–88.
Sermanet, P., Eigen, D., Zhang, X., Mathieu, M., Fergus, R., LeCun, Y., 2014. Overfeat: Integrated recognition, localization and detection using convolutional networks. In: International Conference on Learning Representations (ICLR 2014). CBLS.
Shin, H., Roth, H.R., Gao, M., Lu, L., Xu, Z., Nogues, I., Yao, J., Mollura, D.J., Summers, R.M., 2016. Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning. IEEE Trans. on Med. Imaging.
Simonyan, K., Zisserman, A., 2014. Very deep convolutional networks for large-scale image recognition. CoRR. Abs/1409.1556.
Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., Salakhutdinov, R., 2014. Dropout: a simple way to prevent neural networks from overfitting. J. Mach. Learn. Res. 15, 1929–1958.
Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A., 2015. Going deeper with convolutions. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

Thodberg, H.H., Kreiborg, S., Juul, A., Pedersen, K.D., 2009. The bonexpert method for automated determination of skeletal maturity. IEEE Trans. Med. Imaging 28, 52–66.

Tristan-Vega, A., Arribas, J.I., 2008. A radius and ulna tw3 bone age assessment system. IEEE Trans. Biomed. Eng. 55, 1463–1476.

White, H., 1963. Radiography of infants and children. JAMA 185, 223.

Wolterink, J., Leiner, T., Viergever, M., IÅ !'gum, I., 2015. Automatic coronary calcium scoring in cardiac ct angiography using convolutional neural networks. In: Navab, N., Hornegger, J., Wells, W., Frangi, A. (Eds.), Medical Image Computing and Computer-Assisted Intervention MICCAI 2015. In: Lecture Notes in Computer Science, Vol. 9349. Springer International Publishing, pp. 589–596.

Zeiler, M.D., Fergus, R., 2014. Computer vision – eccv 2014: 13th european conference, zurich, switzerland, september 6-12, 2014, proceedings, part i. In: Chapter Visualizing and Understanding Convolutional Networks. Springer International Publishing, Cham, pp. 818–833.

Zhang, A., Gertych, A., Liu, B.J., 2007. Automatic bone age assessment for young children from newborn to 7-year-old using carpal bones. Comput. Med. Imaging Graph 31, 299–310.