End-to-end Bone Age Assessment with Residual Learning

Daniel Souza
Instituto de Informática - UFRGS
Porto Alegre, RS, Brazil
dsouza@inf.ufrgs.br

Manuel M. Oliveira
Instituto de Informática - UFRGS
Porto Alegre, RS, Brazil
oliveira@inf.ufrgs.br

***Abstract*—Bone age is a reliable metric for determining the level of biological maturity of children and adolescents. Its assessment is a crucial part of the diagnosis of a variety of pediatric syndromes that affect growth, such as endocrine disorders. The most commonly used method for bone age assessment (BAA) is still based on the comparison of the patient's hand and wrist radiograph to a bone age atlas. Such a method, however, takes considerable time, requires an expert rater, and suffers from high inter-rater variability. We present a deep-learning-based approach to estimate bone age from radiographs. It provides a fast, deterministic solution for bone-age assessment. We demonstrate the effectiveness of our method by using it to rate a set of 200 radiographs as part of a contest organized by the Radiological Society of North America. The results of this experiment have shown that our method's performance is similar to the one of a trained physician. Our system is available on-line, providing a free global service for doctors working in remote areas or in institutions with no BAA experts.**

I. Introduction

A child's *bone age*, or *skeletal age*, is an indicator of her/his level of biological and structural maturity, which may not agree with the child's chronological age. Both delayed and increased bone age can be symptoms of more serious pediatric disorders, hence the importance of bone age assessment (BAA). However, making good estimates of skeletal maturity is a complex and specialized task, requiring detailed examination of many related factors and an understanding of the processes associated with bone development [1]. Simply put, one has to analyze the growth and deposition of calcium in regions undergoing ossification [1], which is done through the inspection of radiographs of the hand and wrist. By convention, the left hand and wrist are used.

Currently, the most common ways of performing BAA are the Greulich-Pyle (GP) [2] and Tanner-Whitehouse (TW2) methods [3]. GP is an atlas-based solution, meaning that bone age is estimated by comparing the patient's radiograph with the most similar standard radiograph on a gender-specific atlas. Today, digital atlases, such as [1] and [4], make the evaluation process more convenient, but the actual assessment still depends on the rater's expertise [5]. TW2 is an improvement on the original Tanner-Whitehouse (TW1) method [6], and consists of analyzing twenty regions of interest (ROIs) in the hand and wrist and assigning a discrete stage of skeletal maturity (*e.g.*, pre-puberty, early-and-mid puberty, late-puberty, post-puberty) to each ROI. Each stage has an associated score, and a table is used to convert the sum of all scores into a bone age estimate. There is an inherent uncertainty in the estimates obtained with both methods. The GP standard radiographs are from a 1931-1942 study conducted with white upper-middle class American boys and girls, and does not take into account ethnic variability. Both GP and TW2 are time-consuming methods, require expert raters, and suffer from high inter-rater variability. Such variability may be critical when making decisions about the most appropriate therapy for each case [7].

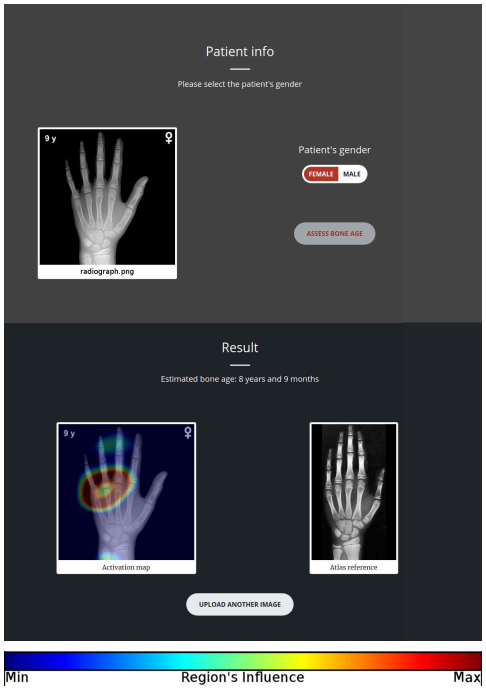


Fig. 1. Interface of the free on-line BAA service based on our technique. (top) Input radiograph and patient's gender. (bottom) BAA estimate, input radiograph with an overlayed colored activation map (color scale below), and most similar reference radiograph from the GP atlas. The interface can be easily configured for any language.

BAA stands as a natural application for machine-learning techniques. Deep *Convolutional Neural Networks* (CNNs) have matched or even surpassed human performance on several image-related tasks, and have quickly become the state-of-the-art for several medical image analysis tasks [8], including

2377-5416/18/$31.00 ©2018 IEEE
DOI 10.1109/SIBGRAPI.2018.00032

classification [9], segmentation [10], and enhancement [11].

We present an automated approach for bone age assessment based on CNNs. We train and validate the performance of our solution on a partially-public dataset containing over 12,500 radiographs from the Pediatric Bone Age Challenge [12], a competition for automating BAA, organized by the Radiological Society of North America (RSNA). We propose a deep, end-to-end solution using *residual learning*. To help doctors understand the decision process that led our CNN to each of its results, we overlay colored *gradient-weighted activation maps* [13] (Grad-CAMs) on the evaluated radiographs (see Figure 1). The accuracy obtained with our solution is comparable to the one produced by a trained radiologist, and superior or equal to state-of-the-art automated methods, while being easier to apply to different datasets and requiring less training time. Our system is available on-line, providing a free global service that is particularly relevant for doctors working in remote areas or in institutions with no BAA experts.

Figure 1 illustrates the interface of our on-line system, which can be configured for different languages. The input radiograph is shown on the top, where the user also informs the patient's gender. The bottom shows the BAA estimate, the input radiograph with an overlayed colored activation map, and the most similar reference radiograph from the GP atlas.

The **contributions** of our work include:

- An end-to-end solution for bone age assessment whose accuracy is similar to the one of expert radiologists (Section IV). Our approach uses a convolution neural network based on residual learning, and can be easily extended with new data;
- An analysis of the most important hand and wrist structures for bone age assessment performed by a CNN that handles BAA as a regression task (Section VI). The results of the analysis are presented for each evaluated radiograph as overlayed color maps indicating the weight of each structure for the bone age assessment;
- A free global on-line bone age assessment service for doctors working in remote areas or in institutions with no BAA experts.

II. Related Work

Several researchers have attempted to develop automatic image processing solutions for estimating skeletal maturity [14]–[16]. These techniques try to detect and measure features from the radiographs, but were unable to handle the high-variability observed in the development of the hand and wrist bones [1].

BoneXpert [17] is an automated BAA solution which has been approved for clinical use in Europe. It segments 15 bones in hand and wrist radiographs and uses the extracted shapes, intensities, and textural features to infer bone age using either GP or TW2 method. BoneXpert does not take carpal bones into account, which may negatively impact the BAA for young patients, for whom these bones have distinguished features.

Somkantha et al. [15] detect boundaries of carpal bones and extract 5 features from them. These features are used for regression using support vector machine (SVM). However, they use a small dataset consisting of 180 images of carpal bones extracted from a digital hand atlas. The used radiographs only cover children from 0 to 6 years old, which is a major limitation to their work.

The work of Spampinato et al. [18] is among the first deep-learning-based approaches to automate BAA. The authors report an average error of about 9.6 months. This is bigger than the error of other solutions, including ours (6.44 months).

Recently, an automated system for BAA using deep learning was proposed by Lee et al. [7]. Their approach consists of fine-tuning GoogLeNet [19] pre-trained on the ImageNet dataset. They use a pre-processing pipeline to segment ROIs and standardize the input radiographs. Unlike previous approaches, this one uses a bigger dataset that contains 8,325 radiographs. The technique casts BAA as a classification task, thus rounding all bone ages down. This limits their assessments to a 1-year granularity. The approach achieves a 57.32% and 61.40% accuracy for female and male patients, respectively.

Iglovikov et al. [20] proposed an automated system using deep learning developed concurrently with ours for the RSNA Bone Age Challenge, and used the same dataset as we did. The authors achieved a Mean Absolute Distance (MAD) of 4.97 months on the test dataset of the RSNA Bone Age Challenge. Although the accuracy of their solution is slightly higher than ours (6.44 months), their approach performs several pre-processing steps using additional CCNs and requires manual intervention at some point of the training process. Firstly, they segment the hand and wrist from the input radiographs using a U-Net [10]. Training such U-Net requires manual generation of segmentation masks, even though it is possible to automate this process to a certain level. They also translate and rotate the radiographs so that the hand bones have a desired position and orientation. This is done by identifying key points on the hand bones using an additional CNN based on the VGG architecture [21] with a regression output. This CNN also requires manual label generation.

Using three different CNNs and the need for manual intervention to create training sets, the average BAA estimates obtained by Iglovikov et al.'s solution was 45 days closer to the ground truth than ours. Given the inherent imprecision of the GP and TW2 methods, the accuracy of our solution can be considered comparable to Iglovikov et al.'s. Our end-to-end solution, however, can be easily applied to different datasets (*e.g.*, different ethnic groups) and requires less training time.

III. Dataset and RSNA challenge

For training our CNN, we used a dataset of png images provided by the Radiological Society of North America (RSNA) as part of RSNA's Pediatric Bone Age Challenge [12]. The dataset consists of hand and wrist radiographs with their respective bone ages. It presents a high variability when it comes to the radiographs, including different acquisition methods, and variations in brightness, contrast, resolution, and even aspect ratio. Figure 2 show 8 radiographs extracted from the training dataset. The top row shows images acquired

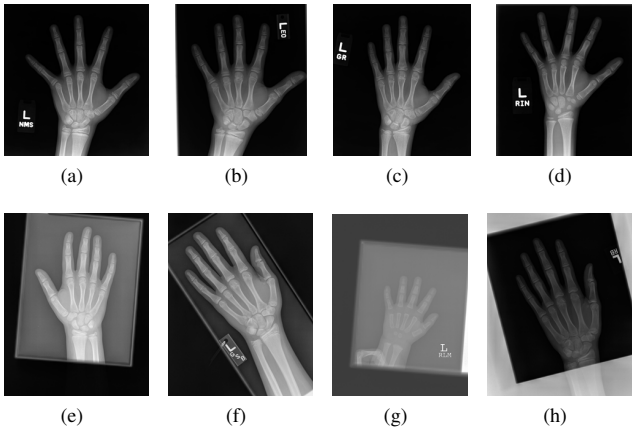


Fig. 2. 8 sample radiographs available on the RSNA training dataset. (top) Images acquired trough Computed Radiography (CR) or Digital Radiography (DR). (bottom) Images digitized from traditional film.

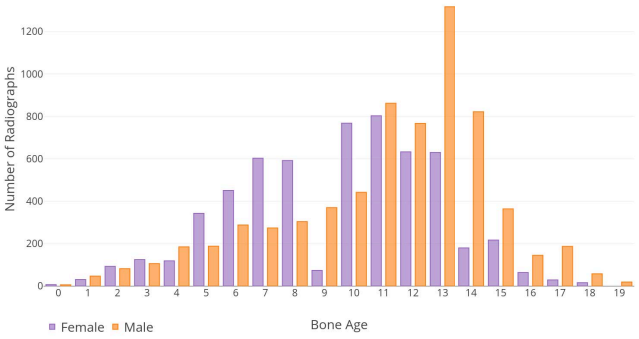


Fig. 3. Age and gender distribution in the RSNA training dataset

trough processes such as Computed Radiography (CR) or Digital Radiography (DR) [22], which usually result in clean images. The bottom row shows radiographs digitized from traditional film, which may not always result in proper images. Since BAA is performed to identify growth disorders, there is a natural unbalance in the age and gender distributions of patients, which is illustrated in Figure 3. Ideally, datasets used for training machine-learning solutions should be balanced. Thus, this might negatively impact the performance of our CNN for groups (*i.e.*, age and gender pairs) for which just a few samples were available for training.

In accordance with the challenge's rules, we used the *mean absolute distance* (MAD) [23] and *concordance correlation coefficient* (CCC) [24] metrics to evaluate the performance of our solution. Given a sample i, its corresponding ground truth value x_i, and the predicted bone age for that sample y_i, the MAD is obtained as:

$$MAD = \frac{\sum_{i=1}^{n} |x_i - y_i|}{n}. \quad (1)$$

CCC is used to measure the agreement between two continuous variables, in our case, the predicted bone age and the ground truth. Given $\bar{x}$ and $\bar{y}$, respectively, the mean for the ground truth and for the predicted bone age, the variances s_x^2 and s_y^2, and covariance s_{xy} for a dataset of size N, CCC is defined as:

$$CCC = \frac{2s_{xy}}{s_x^2 + s_y^2 + (\bar{x} - \bar{y})^2}, \quad (2)$$

where

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i \quad \text{and} \quad \bar{y} = \frac{1}{N}\sum_{i=1}^{N} y_i, \quad (3)$$

$$s_x^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2 \quad \text{and} \quad s_y^2 = \frac{1}{N}\sum_{i=1}^{N}(y_i - \bar{y})^2, \quad (4)$$

and

$$s_{xy} = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y}). \quad (5)$$

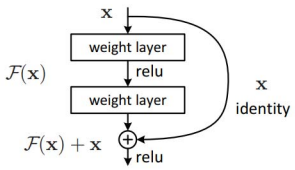


Fig. 4. Definition of a generic building block. Source: [25]

IV. Automated BAA with Residual Learning

Automated BAA comes as a solution to the rater variability in traditional methods. It can also increase productivity and assist less experienced radiologists. We use deep CNNs to automate BAA. One major problem with training deep neural networks is that by increasing the network depth, its accuracy tends to saturate, quickly degrading. Surprisingly, this is not due to over-fitting. Simply adding layers leads to higher training error [25]. *Residual networks* address this problem using *skip connections*, which allow a network to learn from input/output of previous layers. This is illustrated in Figure 4, where the identity connection x causes the resulting mapping to become $F(x)+x$. This is desirable since the previous layers might have degraded the value of $F(x)$.

Residual networks can go deeper and learn to represent more complex models without saturating accuracy. We have proposed several custom residual network architectures that consist, partly, of an ensemble of *custom blocks*. Custom blocks are composed of a *convolutional layer* followed by an *activation layer*, which is then followed by a number of *building blocks* ranging from 1 to 3 (Figure 5 - top). Each building block has 2 convolutional layers, with an activation layer in-between them, followed by an addition operator and another activation layer. Finally, our custom block has a *pooling step*. Each custom block can be followed by a similar wider block. By progressively increasing the number of kernels used in deeper layers (see Table I), our network can learn more from its deeper layers than from its earlier ones. Then, we flatten the output of our custom block ensemble, and concatenate to it the patient's gender, a one-hot encoded vector

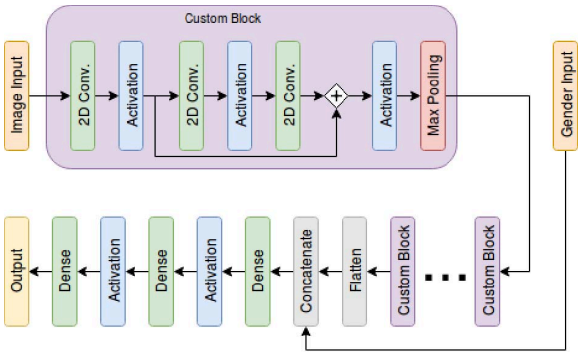


Fig. 5. Representation of our candidate architectures (Table I): A to D use 3 custom blocks with a single building block each, while architectures E to H use 5 custom blocks with 3 building blocks each.

where each gender is treated as a different category. Finally, we append dense layers with activations in-between to obtain the bone age (Figure 5). Since radiographs from patients with the same age, but different genders, are largely different, we chose to provide the gender as input to the dense part of our neural network.

We have implemented our solution using Keras [26] on top of Tensorflow [27]. Defining complex architectures in Keras is as simple as "stacking" layers. Tensorflow is open-source, has great performance, and is constantly improved by Google.

A. Candidate Architectures

We have compared a total of 8 custom architectures, from A to H, as seen on Table I. We have also compared the performance of the Inception-V3 architecture [28] when fine-tuned on our data. Our custom architectures are progressively more complex (from A to H). For all architectures, we normalize the input images in the training dataset so that the intensities of each such image have (approximately) zero mean and variance one. This is a standard process for training neural networks.

We also pre-process the training and test datasets by resizing the images and performing *feature scaling*. Feature scaling transforms the original pixel values, represented as 8-bit integers, to float values in the [0.0,1.0] interval, where 0.0 and 1.0 correspond to the minimum and maximum values, respectively. Both steps are necessary due to the high variability in the dataset. We resize the images to 256×256 pixels, so that we can fit a batch with 32 images on the GPU memory. Although image downscaling implies discarding some information, it also reduces the amount of weights to be learned, thus accelerating convergence. We trained our CNN using an NVIDIA GeForce GTX 1080 Ti GPU, with 11 GB of memory. Radiographs have different dimensions (height and width), and the aspect ratio of each radiograph should be preserved to avoid feature distortion (*e.g.*, the relative size of the epiphyses of the phalanges with respect to the size of adjacent metaphyses). Therefore, we downscale the radiographs so that its largest dimension contains 256 pixels. The smaller dimension is then padded with zeros to 256 pixels. Alternatives to resizing the images include:

- Dividing the input image in ROIs, and training on each ROI individually. BAA could then be performed by using the ensemble of individual evaluations. However, this is a complex approach that may not outperform the use of image downscaling;
- Using a Bulk Synchronization Parallel (BSP) model [29]. This approach consists in splitting the image into smaller patches that are fed to the CNN separately. Then, a padding and normalization technique merges the patches into a single image. This is also a more complex approach, which can be explored as future work.

We experimented with several techniques and hyper-parameters (see Table I) that have a direct impact on our network's performance:

Regularization consists in trading-off flexibility with model complexity to avoid overfitting. We used two regularization methods: Dropout and L2 Regularization. *Dropout* consists in randomly disregarding the output of some neurons during the training phase. This gives neurons that were dependent on their neighbors the ability to learn new features and helps generalize the learning process [30]. *L2 regularization* applies a penalty that is progressively higher as our neural network gets deeper. Deeper convolutional layers learn more complex features, which may not generalize well and should have a penalty applied to their weights [31].

Normalization standardizes the inputs to hidden layers. We use either *batch normalization* [32] or *instance normalization* [33]. Batch normalization consists in subtracting the batch's mean from the input of a hidden layer, and then dividing the resulting value by the batch's standard deviation. Instead of performing normalization within a batch, instance normalization does it within a single sample. Both processes have several benefits, such as accelerating convergence and adding regularization to our network.

Data Augmentation makes the model more robust to variations in the input data, as well as virtually increase the size of our dataset. We perform data augmentation on our training dataset by randomly applying transformations that cause no impact on the assessment of the bone age. These include rotations, uniform scaling, horizontal and vertical jitter, and horizontal flip.

B. Training and Evaluating the Proposed Architectures

We trained all architectures described in Table I, and the fine-tuned Inception-V3 for at least 50 epochs. The progression of the validation MAD for the trainings of the 4 best performing architectures (B, D, F, and H) and for Inception-V3 are shown in Figure 6. Furthermore, we trained architectures G, H, and Inception-V3 for additional 100 epochs, for we believed that they, due to their complexity, would be able to learn more if trained longer. The final validation MAD values as well as the corresponding numbers of trained epochs for architectures A to H are shown at the bottom of Table I.

TABLE I
CANDIDATE ARCHITECTURES

	Experiment							
	A	B	C	D	E	F	G	H
Conv Layers (CL)	9	9	9	9	35	35	35	35
Custom Blocks (CB)	3	3	3	3	5	5	5	5
Building Blocks per CB	1	1	1	1	3	3	3	3
CL per CB	3	3	3	3	7	7	7	7
Filters per CB	16, 32, 64	16, 32, 64	16, 32, 64	16, 32, 64	16, 32, 64, 128, 256	16, 32, 64, 128, 256	16, 32, 64, 128, 256	16, 32, 64, 128, 256
Filter size	3×3	3×3	3×3	3×3	3×3	5×5	5×5	5×5
L2 regularization	–	–	–	–	–	–	–	0.0001
Dropout	–	–	–	–	–	–	0.15	0.15
Normalization	–	–	batch	instance	instance	instance	instance	instance
Data augmentation	–	yes	yes	yes	yes	yes	yes	yes
Final validation MAD	20.47	10.47	16.61	12	16.98	10.73	20.33	9.74
# of trained epochs	50	50	50	50	50	50	150	150

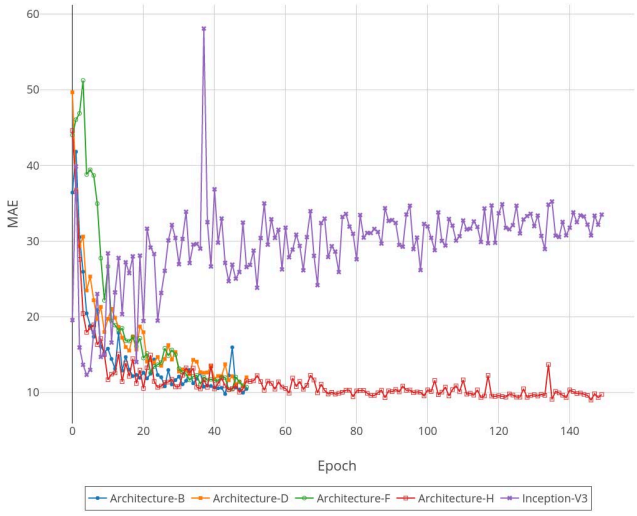


Fig. 6. MAD evolution by epoch for the best architectures and Inception-V3.

For all proposed architectures, as well as for Inception-V3, we used the Adaptive Moment Estimation (Adam) optimizer [34]. Adam is able to adapt the learning rate to the parameters, granting faster and better convergence. We randomly split the training dataset, leaving 80% of the images for training and 20% for validation. During training, the batches were reshuffled at each epoch. After analyzing the results of the eight architectures shown in Table I, *we selected architecture H for deployment of our system.*

V. RESULTS AND DICUSSION

We evaluated the accuracy of our network by performing BAA for the 200 radiographs that constitute the test dataset of the RSNA challenge. Some of these images are shown in Figure 7. Note the variability in brightness, contrast, orientation, scale, and even the presence of a flipped radiograph that appears like a right hand (top row, third image from left to right). The ground-truth bone age for these radiographs was only known by the organizers of the challenge.

The average time for estimating the bone age for each radiograph was approximately $35\ ms$. The complete set of estimates was submitted on-line to the RSNA contest, and our results achieved a MAD of 6.44 months and a CCC of 0.97. Such CCC value demonstrates *a substantial level of agreement between our predictions and the ground-truth* [35].

The Digital Hand Atlas [36], prepared by researchers from the University of Southern California, is the most comprehensive dataset of hand radiographs with bone age estimates produced by two raters. For this dataset, the inter-rater variability can be summarized by an overall RMSE of 0.59 years (0.57 years for males, and 0.54 years for females), and 0.66 years for children between 5 and 18 years old. A more recent study [37] reports an inter-rater variability of 0.51 ± 0.44 years with the use of the GP method. Our results obtained a MAD of 0.536 years (6.44 months), which is in accordance with results produced by expert radiologists.

For comparison, the average error in the estimates obtained with the recent deep-learning-based system described in [7] is 0.82 years for males and 0.93 years for females. The commercial software BoneXpert has an average error of 0.72 years [38]. Our results surpasses these systems.

The best results for the RSNA challenge were obtained by the system described by Iglovikov et al. [20], which obtained a MAD of 4.97 months (0.414 years), a difference of only about 45 days with respect to our estimates. As already mentioned, unlike Iglovikov et al.'s system, ours is an end-to-end solution.

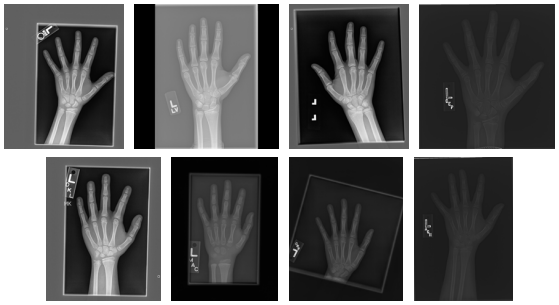


Fig. 7. 8 samples of radiographs available on the test dataset. Source: RSNA

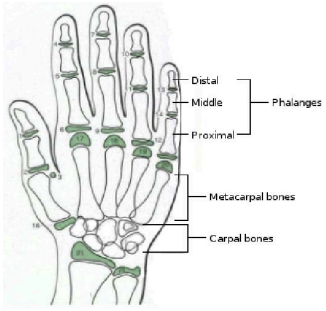


Fig. 8. Diagram of hand and wrist bones. Adapted from [41].

We developed a website [39] where users can upload radiographs and receive the corresponding bone age assessment computed by our network (Figure 1). This website's front-end was developed using Angular – an open-source front-end web application framework – and its back-end was implemented with Django – a high-level Python web framework. Our server handles the uploading of images, the serving of static content, and the actual BAA.

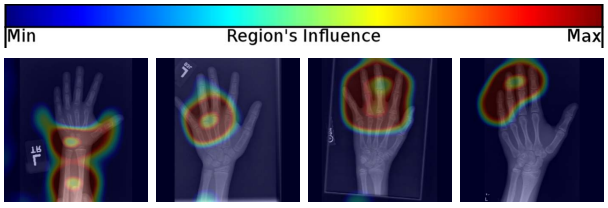
VI. Highlighting the Relevant Features

CNNs, particularly end-to-end systems as ours, may excel in accuracy, but often lack in terms of explaining how their results were obtained. In order to better understand the behaviour of our CNN and its results, we use a technique known as Gradient-weighted Class Activation Mapping [13] (Grad-CAM), available in the Keras Visualization Toolkit [40].

We use Grad-CAM with the gradients relative to the upper convolutional layers of our network to produce a map highlighting the structures that most influenced the BAA estimate. The toolkit underwent a series of changes to accept our input, since our network expects both the gender and the radiograph of the patient, and KerasVIS only takes images. We also had to adapt it to work with single-channel images. Our Grad-CAM uses a color map to represent the importance of a region to the BAA, with warmer colors representing stronger influence.

Figure 9 show Grad-CAM for radiographs of different stages of skeletal maturity: pre-puberty, early and mid-puberty, late-puberty, and post-puberty – toddlers were left out on purpose, for there are few images for this age group in the training dataset, as seen in Figure 3. We compared the regions highlighted by our Grad-CAMs to the regions taken into account by radiologists while performing BAA and noticed that they are not necessarily similar. Here are the observed differences for each stage of skeletal maturity:

- *Pre-puberty*: In this stage, radiologists primarily compare the size of the *epiphyses* of the phalanges to the size of adjacent *metaphyses* (see Figure 8). Our CNN takes a different approach, deeming the ulna, radius, carpal and metacarpal bones as the most important structures for BAA during this stage;
- *Early and mid-puberty*: radiologists primarily analyze the size of the epiphyses in the distal and middle phalanges. Our CNN, again, takes a different approach, treating the metacarpal bones and proximal phalanges as the most distinctive structures;
- *Late-puberty*: radiologists primarily consider the degree of epiphyseal fusion of the distal phalanges. Our CNN finds the distal phalanges to be the most relevant structures for BAA during this stage, but also finds the middle and proximal phalanges, and metacarpal bones to be important;
- *Post-puberty*: radiologists primarily analyze the degree of epiphyseal fusion of the ulna and radius bones. Our CNN takes a different approach, deeming the phalanges alone to be the most important structures for BAA in this stage.



(b) Pre-puberty (c) Early and Mid-puberty (d) Late-puberty (e) Post-puberty

Fig. 9. Importance of bone structures for estimating skeletal maturity according to our CNN.

The fact that the regions considered most relevant by our CNN do not match the regions analyzed by radiologists might suggest ways for improving our understanding about the bone maturing process, as well as for improving traditional BAA methods.

VII. Conclusion and Future Work

We presented an automated approach for bone age assessment based on residual learning. We trained and validated our CNN on a partially-public dataset containing over 12,500 radiographs from the Pediatric Bone Age Challenge [12], organized by the Radiological Society of North America (RSNA). We evaluated the accuracy of our network by performing BAA for 200 radiographs from the test dataset of the RSNA challenge. Our results achieved a MAD of 6.44 months and a CCC of 0.97, indicating a substantial level of agreement between our predictions and the ground-truth.

The accuracy of our solution is similar to the one obtained by expert radiologists and superior to previous automated

systems. For the RSNA challenge dataset, our system obtained a MAD of only 45 days bigger than the winner system [20]. Note, however, that among radiologists the inter-rater variability corresponds to 0.51 $\pm$ 0.44 years [37], making the performance of both systems comparable.

We have discussed the most relevant hand and wrist structures identified by our CNN for BAA, and have compared them to the features observed by expert radiologists. A free bone age assessment service based on our system is available on-line and can be a valuable resource for doctors working in remote areas or in institutions with no BAA experts.

We would like to evaluate the performance of our algorithm by integrating it to a Picture Archiving and Communication System (PACS), a technology used for storing, retrieving, visualizing, and sharing medical images. Such systems are commonly used in hospitals, and would allow us to integrate our algorithm to the radiologists' work-flow.

ACKNOWLEDGMENTS

Manuel M. Oliveira acknowledges CNPq grants 306196/2014-0, 423673/2016-5.

REFERENCES

[1] V. Gilsanz and O. Ratib, *Hand Bone Age: A Digital Atlas of Skeletal Maturity*. Springer Berlin Heidelberg, 2005.

[2] W. W. Greulich and S. I. Pyle, *Radiographic Atlas of Skeletal Development of the Hand and Wrist*. Stanford University Press, 1959.

[3] J. Tanner, R. Whitehouse, N. Cameron, and W. e. a. Marshall, *Assessment of skeletal maturity and prediction of adult height (TW2 method)*. Academic Press London, 1975.

[4] C. Gaskin, S. Kahn, J. Bertozzi, and P. Bunch, *Skeletal Development of the Hand and Wrist: A Radiographic Atlas and Digital Bone Age Companion*. Oxford University Press, USA, 2011.

[5] P. M. Bunch, T. A. Altes, J. McIlhenny, J. Patrie, and C. M. Gaskin, "Skeletal development of the hand and wrist: digital bone age companion-a suitable alternative to the greulich and pyle atlas for bone age assessment?" *Skeletal radiology*, vol. 46, no. 6, pp. 785–793, 2017.

[6] J. Tanner and R. WhitehouseandM, "A new system for estimating skeletal maturity from the hand and wrist, with standards derived from a study of 2600 healthy british children." *International Children's Centre, Paris*.

[7] H. Lee, S. Tajmir, J. Lee, and M. e. a. Zissen, "Fully-automated deep learning system for bone age assessment," *Journal of Digital Imaging*, pp. 1–15, 2017.

[8] G. J. S. Litjens, T. Kooi, B. E. Bejnordi, and A. A. A. e. a. Setio, "A survey on deep learning in medical image analysis," *CoRR*, vol. abs/1702.05747, 2017.

[9] A. Esteva, B. Kuprel, and R. A. e. a. Novoa, "Dermatologist-level classification of skin cancer with deep neural networks," *Nature*, vol. 542, pp. 115–, Jan. 2017.

[10] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," *CoRR*, vol. abs/1505.04597, 2015.

[11] R. Li, W. Zhang, and H. I. e. a. Suk, "Deep learning based imaging data completion for improved brain disease diagnosis." *Medical image computing and computer-assisted intervention*, vol. 17, no. Pt 3, pp. 305–312, 2014.

[12] "RSNA pediatric boneage challenge," 2018, [Online; accessed 30-May-2018]. [Online]. Available: http://rsnachallenges.cloudapp.net/competitions/4

[13] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization," *ArXiv e-prints*, Oct. 2016.

[14] A. Zhang, A. Gertych, and B. J Liu, "Automatic bone age assessment for young children from newborn to 7-year-old using carpal bones," vol. 31, pp. 299–310, 06 2007.

[15] K. Somkantha, N. Theera-Umpon, and S. Auephanwiriyakul, "Bone age assessment in young children using automatic carpal bone feature extraction and support vector regression," *Journal of Digital Imaging*, vol. 24, pp. 1044–1058, 2011.

[16] J. Seok, B. Hyun, J. Kasa-Vubu, and A. Girard, "Automated classification system for bone age x-ray images," in *2012 IEEE Intern. Conference on Systems, Man, and Cybernetics (SMC)*, Oct 2012, pp. 208–213.

[17] H. H. Thodberg, S. Kreiborg, A. Juul, and K. D. Pedersen, "The bonexpert method for automated determination of skeletal maturity," *IEEE Trans. on Medical Imaging*, vol. 28, no. 1, pp. 52–66, Jan 2009.

[18] C. Spampinato, S. Palazzo, and D. e. a. Giordano, "Deep learning for automated skeletal bone age assessment in x-ray images," *Medical image analysis*, vol. 36, pp. 41–51, 2017.

[19] C. Szegedy, W. Liu, Y. Jia, and P. e. a. Sermanet, "Going deeper with convolutions," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015, pp. 1–9.

[20] V. Iglovikov, A. Rakhlin, A. A. Kalinin, and A. Shvets, "Pediatric bone age assessment using deep convolutional neural networks," *CoRR*, vol. abs/1712.05053, 2017.

[21] 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

[22] M. Krner, C. H. Weber, S. Wirth, K. Pfeifer, M. F. Reiser, and M. Treitl, "Advances in digital radiography: Physical principles and system overview," *RadioGraphics*, vol. 27, no. 3, pp. 675–686, 2007, pMID: 17495286.

[23] C. Sammut and G. I. Webb, Eds., *Mean Absolute Error*. Boston, MA: Springer US, 2010, pp. 652–652.

[24] L. I. Lin, "A concordance correlation coefficient to evaluate reproducibility," *Biometrics*, vol. 45, no. 1, pp. 255–268, 1989.

[25] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," *ArXiv e-prints*, Dec. 2015.

[26] F. Chollet *et al.*, "Keras," https://github.com/fchollet/keras, 2015.

[27] M. Abadi, A. Agarwal, and P. B. et al., "TensorFlow: Large-scale machine learning on heterogeneous systems," 2015, software available from tensorflow.org. [Online]. Available: https://www.tensorflow.org/

[28] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the Inception Architecture for Computer Vision," *ArXiv e-prints*, Dec. 2015.

[29] S. Wu, M. Zhang, G. Chen, and K. Chen, "A new approach to compute cnns for extremely large images," in *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, ser. CIKM '17. ACM, 2017, pp. 39–48.

[30] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *J. of Machine Learning Research*, vol. 15, pp. 1929–1958, 2014.

[31] C. Cortes, M. Mohri, and A. Rostamizadeh, "L2 regularization for learning kernels," *CoRR*, vol. abs/1205.2653, 2012. [Online]. Available: http://arxiv.org/abs/1205.2653

[32] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," *CoRR*, vol. abs/1502.03167, 2015. [Online]. Available: http://arxiv.org/abs/1502.03167

[33] D. Ulyanov, A. Vedaldi, and V. S. Lempitsky, "Instance normalization: The missing ingredient for fast stylization," *CoRR*, vol. abs/1607.08022, 2016. [Online]. Available: http://arxiv.org/abs/1607.08022

[34] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *CoRR*, vol. abs/1412.6980, 2014. [Online]. Available: http://arxiv.org/abs/1412.6980

[35] G. McBride, "A proposal for strength-of-agreement criteria for lin's concordance correlation coefficient," pp. 2005–2062, 01 2005.

[36] F. Cao, H. Huang, and E. e. a. Pietka, "Digital hand atlas for web-based bone age assessment: System design and implementation," vol. 24, pp. 297–307, 04 2000.

[37] S. Young Kim, Y. Joung Oh, and J. e. a. Yeon Shin, "Comparison of the greulich-pyle and tanner whitehouse (tw3) methods in bone age assessment," vol. 13, pp. 50–55, 06 2008.

[38] "The bonexpert product," https://www.bonexpert.com/products/the-bonexpert-product.

[39] I. Health, "Bone age assessment," https://iarahealth.com/boneage/, 2018.

[40] R. Kotikalapudi and contributors, "keras-vis," https://github.com/raghakot/keras-vis, 2017.

[41] I. D. Academy, "Skeletal maturity indicator," https://www.slideshare.net/indiandentalacademy/skeletal-maturity-indicator-2, 2016.