



Bone age estimation using deep learning and hand X-ray images

Jang Hyung Lee¹ · Young Jae Kim¹ · Kwang Gi Kim¹

Received: 8 August 2019 / Revised: 4 February 2020 / Accepted: 13 February 2020 / Published online: 11 March 2020
© Korean Society of Medical and Biological Engineering 2020

Abstract

Bones during growth period undergo substantial changes in shape and size. X-ray imaging has been routinely used for bone growth diagnosis purpose. Hand has been the part of choice for X-ray imaging due to its high bone parts count and relatively low radiation requirement. Traditionally, bone age estimation has been performed by referencing atlases of images of hand bone regions where aging-related metamorphoses are most conspicuous. Tanner and Whitehouse' and Greulich and Pyle's are some well known ones. The process entails manual comparison of subject's hand region images against a set of corresponding images in the atlases. It is desired to estimate bone age from hand images in an automated manner, which would facilitate more efficient estimation in terms of time and labor cost and enables quantitative and objective assessments. Deep learning method has proved to be a viable approach in a number of application domains. It is also gaining wider grounds in medical image analysis. A cascaded structure of layers can be trained to mimic the image-based cognitive and inference processes of human and other higher organisms. We employed a set of well known deep learning network architectures. In the current study, 3000 images were manually curated to mark feature points on hands. They were used as reference points in removing unnecessary image regions and to retain regions of interest (ROI) relevant to age estimation. Different ROI's were defined and used—that of rather small area mostly made up of carpal and metacarpal bones and that includes most of phalanges in addition. Irrelevant intensity variation across cropped images was minimized by applying histogram equalization. In consideration of the established gender difference in growth rates, separate gender models were built. Certain age range image data are far scarcer and exhibit rather large excursion in morphology from other age ranges—e.g. infancy and very early childhood. Many studies excluded them and addressed only elder subjects in later developmental stages. Considering infant age group's diagnosis demand is just as valid as elder groups', we included entire age ranges for our study. A number of different deep learning architectures were trained with varying region of interest definitions. Smallest mean absolute difference error was 8.890 months for a test set of 400 images. This study was preliminary, and in the future, we plan to investigate alternative approaches not taken in the present study.

Keywords Deep learning · Hand bone · X-ray · Bone age · Greulich and Pyle atlas · Tanner and Whitehouse atlas

Jang Hyung Lee and Young Jae Kim have contributed equally to this work as first authors.

✉ Kwang Gi Kim
kimkg@gachon.ac.kr

Jang Hyung Lee
leejh161@naver.com

Young Jae Kim
youngjae@gachon.ac.kr

¹ Department of Biomedical Engineering, School of Medicine, Gachon University, 410-769, Incheon 21565, Korea

1 Introduction

Bones undergo substantial shape changes throughout lifetime. The changes are particularly the strongest during the growth period. Hand is joined to lower arm via radius and ulna bones and is comprised of carpal, metacarpal and phalange bones, which together count thirty bone parts (Fig. 1). The high bone count in a relatively small spatial volume and rather small radiation requirement makes hand an ideal body part to take X-ray images from. This property was exploited by Greulich and Pyle [1]. Specifically, they created an atlas of regions of hands in X-ray that are chronologically most distinguishing through the course of aging. Tanner and

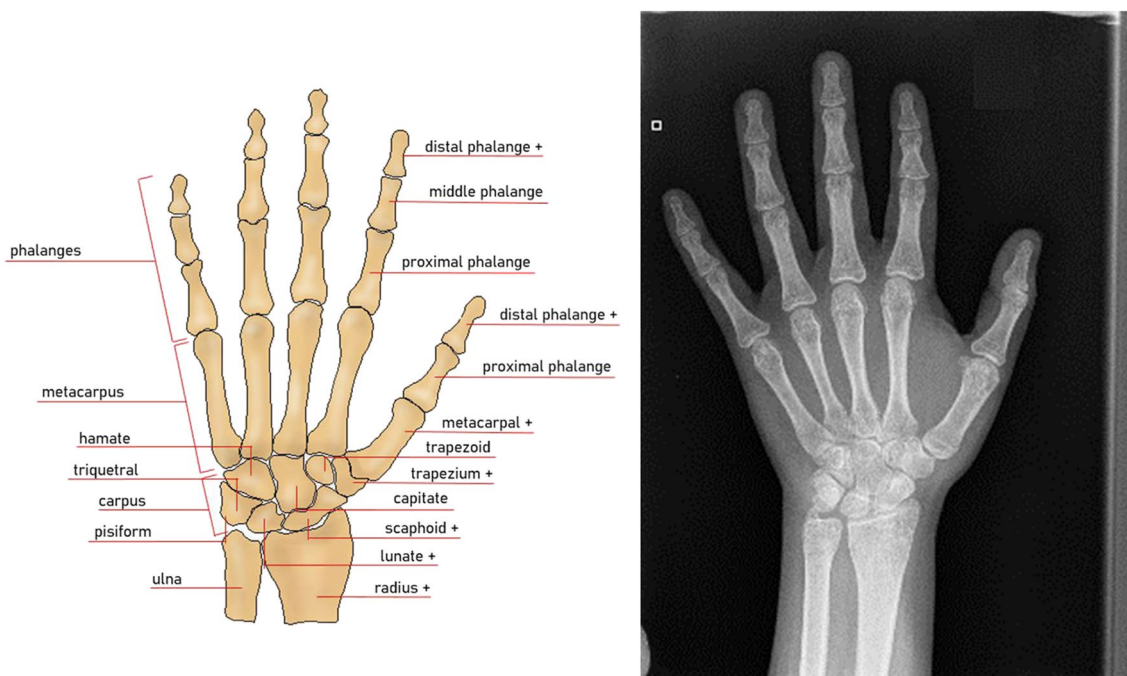


Fig. 1 Skeletal anatomy [3] and an X-ray image of hand [4]

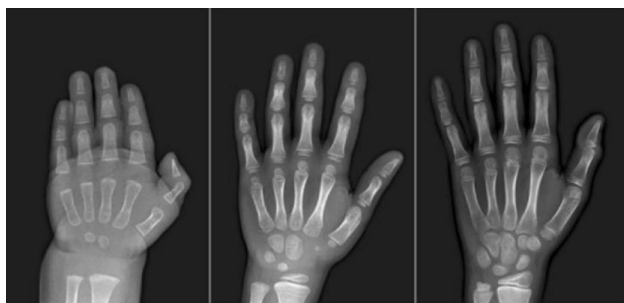


Fig. 2 Hand bones at different ages

Whitehouse later compiled a more extensive atlas of images characterizing age-wise morphological changes of bones [2].

Example images of hand bones at different ages are shown in Fig. 2.

Size change is the most well perceived aging-induced bone morphology change. During infancy, diaphyses make up most part of bone matter. Another notable shape change from aging process is ossification: especially, the cartilage in epiphyses regions calcify to form hard bones.

While most people follow regular growth path, some people might follow substantially delayed or advanced growth paths. People in infant to juvenile periods with relatively low growth status compared with their age peers may deserve medical diagnosis and special treatment for growth stimulation or removal of pathological causes such as endocrine disorders. In order to make an objective assessment

for normality of growth, diagnosis has to be as accurate as possible.

Atlas based method requires arduous human labor to compare patient's images against reference atlas images [5, 6]. Also the process is subjective - different people judging differently depending on their personal level of experiences and their training path specific views. Being based on visual similarity-cued individual decisions, the backing ground for diagnostic results is also hard to quantify. Possibly lower reproducibility is another disadvantage of atlas based manual estimation.

Machine learning methods provide for a way to automatically carry out image based inferences by forming a mapping from input to output. Computerized and automated processing and provision for quantitative measures of similarity, etc. are a merit afforded by the machine learning method. Minimal human intervention reduces labor cost and improves time efficiency in clinical environments. BoneXpert introduced in 2009 is a machine learning based bone age estimation system [7].

Deep learning technique is finding increasing use cases in diverse medical information processing and analysis fields [8, 9]. In particular, medical image analysis has been an apt field to leverage image-related cognitive and inferential efficacy of deep learning approaches towards. Typically, a model is formed which maps an input to a desired form of output. For the problem of hand bone age estimation, X-ray image corresponds to input and estimated age to the output. Neural network is made up of a large number of

nodes and edges, and a proper weighting for the edges have to be determined. A set of numerical values characterizing input and output relationship is formed via training process. Input image goes through a chain of processing stages to be mapped to desired kind of output. Traditional machine learning approaches typically assumed a set of human defined feature representations which may not always be optimal for problems being tackled.

From training process perspective, deep learning provides developers and researchers with the benefit of end to end training where raw input data together with desired output labels are presented and neural network is brought to a configuration reasonably close to optimal, with minimal human intervention.

Deep learning usually employs neural networks of high level of learning capacity. A layer is made up of a tiled array of processing units. Training process forwards inputs through network to produce output. It is then compared with correct desired output, difference of which is error signal. The error is backward propagated to adjust edge weights connecting network's constituent nodes [10]. Some issues emerge with deep learning process. One is the huge imbalance between the number of edges and the typical amount of available data to be used during the training phase. This will result in an overfitting where model is highly tuned to the specifics of training data and fails to generalize well to data instances that are not represented in the training dataset. Some measures such as dropout layer is often employed to combat the imbalance and overfitting issue [11]. Dropout layer typically is placed in between the fully connected layers. It dynamically switches on and off the edges connecting the two layers and helps reduce coadaptation phenomenon known to occur where nodes and weights are trained to similar configuration when the dropout layer is not in use. It is a utility layer effective during the training phase, and during the test and deployment phases, the dropout layer is removed.

1.1 Problem formulation and data preparation

Given an input hand X-ray image, we seek an automatic way to infer bone age. Since the age to be estimated is a continuous variable, a regression model is a reasonable way of mapping out the input–output relationship.

i.e. $\hat{a} = M(I)$

\hat{a} = estimated age and M is a regression model, I is input image of width, w and height, h , and i is in $I \in N^{wh}$, where N is natural number.

The present days' neural network architectures all assume fixed size input images. For example, CaffeNet [12, 13] assumes a fixed size of 227×227 pixels, and the

googlenet, 224×224 pixels. Then the above expression reduces to $\hat{a} = M(I), I \in N^{224 \times 224}$ for the case of Googlenet.

An anatomy of hand made up of carpal, metacarpal and phalange bones is shown in Fig. 1. A dataset is needed to train a model. We used the data from the RSNA challenge [4]. More than 12 thousand image files in png (portable network graphics) format are provided. For the training set, age and gender information is associated with each data instance. For the validation set, age information is provided yet gender information is missing. Test set has neither age nor gender annotations. Images are in gray-scale and ranges from 169 to 3639 kilobytes in file size, and dimension ranges between 594 and 2970 pixels a side. Due to partial or full lack of critical age and gender information from the validation and test sets, we solely used the training set in our study—i.e., it was partitioned into training and test sets in a mutually exclusive manner so that an entry in one set is excluded from the other set. Most of widely used deep learning networks assume three channel RGB color inputs. All X-ray images in the training set were single channel black and white images. Hence they were converted to three channel image data by duplicating their single channel data.

Hands assume quite substantial variability in angle with respect to border lines within the image frame. Also image intensity range is quite wide that hand regions appear in from nearly white to dark gray tints over data set (Fig. 3). Data in the entire age range from 0 months to 228 months were included for both training and testing in our study - no age ranges were excluded. We randomly selected 1400 images each of male and female cohorts to total 2800 images. For each gender, one thousand images were used to form training data and the remaining data were used to form testing data.

2 Methods

A number of factors determine the appearance of hand bone X-ray images. Values of individual pixels comprising an image are determined by the settings used during X-ray taking, relative orientation of hand with respect to image frame, overall intensity level of hand region in addition to the age-related morphological characteristics. What matter most from the pattern recognition perspective are the patterns that are formed by the spatial arrangement of intensity values assumed by pixels. On the other hand, absolute levels of intensity values may directly impact the output of neural network. Hence it was desired to minimize the intensity variation across image data instances that are irrelevant to age-specific morphological characteristics and changes.

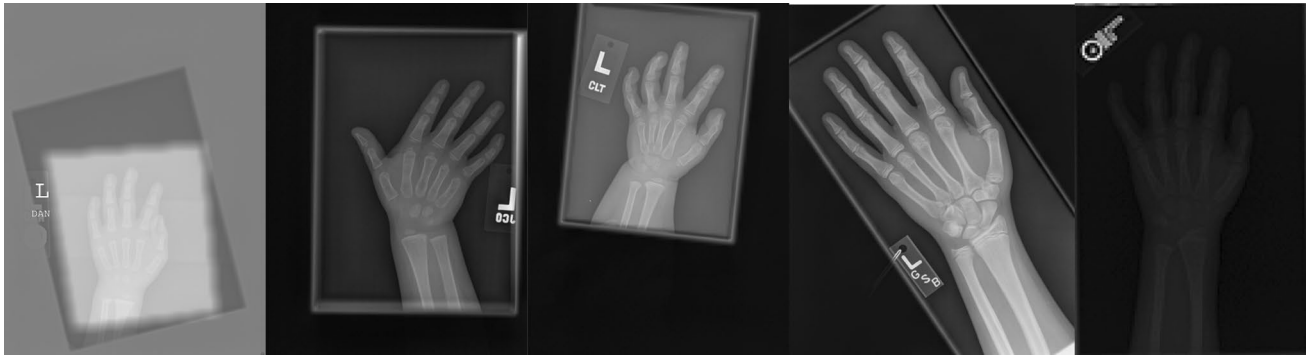


Fig. 3 Variability of hand X-ray images

2.1 Region of interest reduction

Setting an appropriate region to retain to use for inference may help reduce the appearance variations that are irrelevant to age. Background portion found in X-ray image data are irrelevant to age estimation. Hence they can be safely removed without any negative implication on the age estimation performance. Typically, similar pattern of ossification is found throughout the hand. Some appearance information may be regarded redundant which may then be excluded from age estimation process. The criteria and tradeoff involved are redundancy and uniqueness of information and whether an aspect of image is relevant

towards age estimation. In terms of degree of freedom in orientation and spatial extension of occupation, fingers exhibit a wider range of variability than the palm. In an attempt to minimize finger pose-induced image variation, a set of points on phalanges, radius and ulna were first defined. Specifically, they were located towards the proximal ends of phalanges and on the distal ends of radius and ulna (Fig. 4). A rectangle of largest area is defined which is aligned to the image border lines and has the reference points on its border or in its interior. A corresponding image is shown in Figs. 4 and 5 (henceforth we refer to this as ROI scheme 1, ROI₁). While the variability due to finger

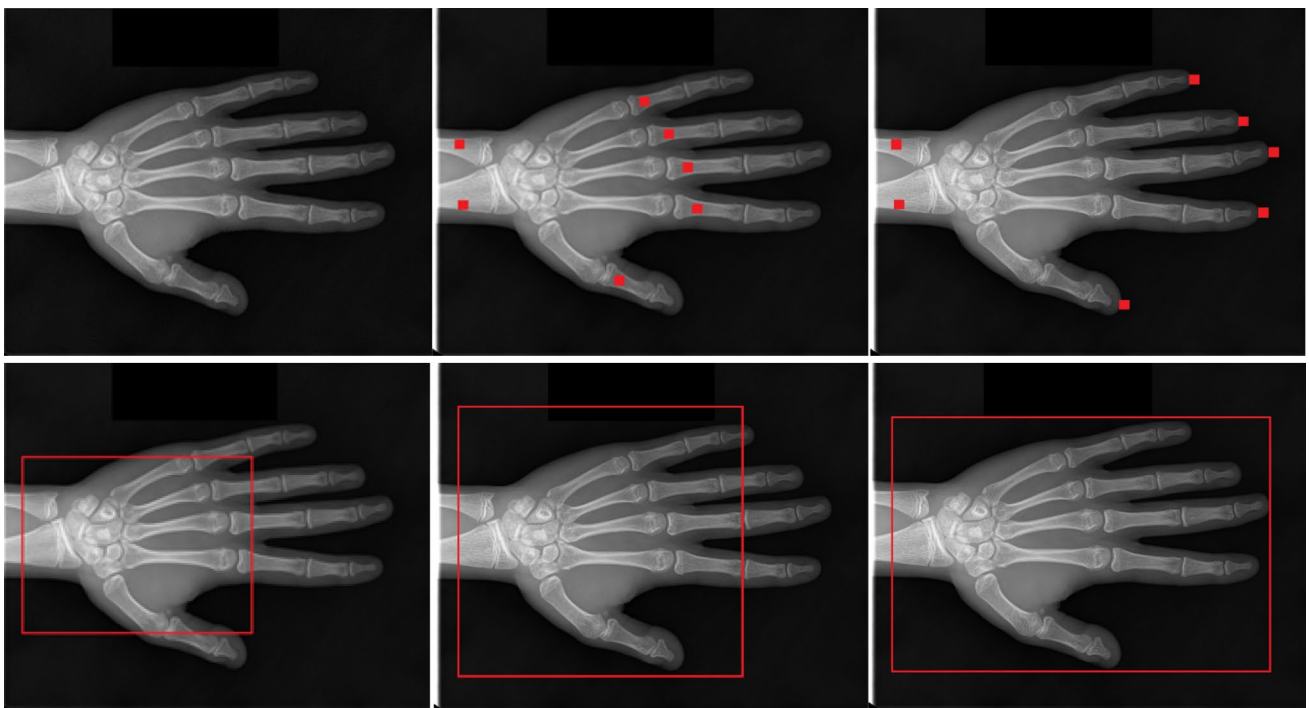


Fig. 4 An input image [4] and feature points marked thereon (upper row), and ROI1, 2 and 3 defined (bottom row)



Fig. 5 Hand images obtained by applying different ROI schemes

poses is greatly reduced, the age-related morphological information from the inter-phalangeal regions is lost.

Alternatively, a larger part of phalangeal bones may be included in an ROI so as to retain and reflect more of phalangeal specific age-related morphological information in the training process. The borderline coordinates in ROI₁ were expanded by 30% in all four directions except for wrist to result in a larger ROI (to be referred to as ROI₂). Entire proximal phalanges are included and all or parts of mid- and distal phalangeal bones are included.

With yet another ROI scheme, ROI₃, all phalanges, carpal and metacarpal bones were included. Overall, we implemented three settings for regions of interest, which we'll compare in terms of resulting estimation accuracy. Some images processed according to the ROI reduction schemes are shown in Fig. 5.

2.2 Intensity variation reduction

Also image intensity has a wide range of fluctuation in the RSNA training dataset, possibly due to the multi institutional nature of the data gathered by different medical agencies. The unwanted intensity variation in hand regions may be removed via a number of ways. Histogram equalization is well known to be a method of choice for equalizing intensity distributions across a set of images, which was adopted and applied to the ROI-reduced images obtained from the previous step. Figure 6 illustrates some images after the application of ROI reduction and histogram equalization.

Overall image processing flow was as shown in Fig. 7.

2.3 Neural network architecture

A large number of neural network architectures were proposed. In addition to the kinds of layers employed, they are distinguished in terms of the number of constituent layers and the topology of inter-connections between them. In this article, we employed a set of well known architectures and compared their performances—Caffenet, Googlenet and Resnet [14] [15]. The deep learning tool suite *Caffe* [12, 13] was used for conducting this study.

2.4 Performance measures

We used mean absolute difference and concordance correlation coefficient measures to quantify the performance levels of algorithms. Given a pair of true age and estimated age, the error, *e*, is defined as

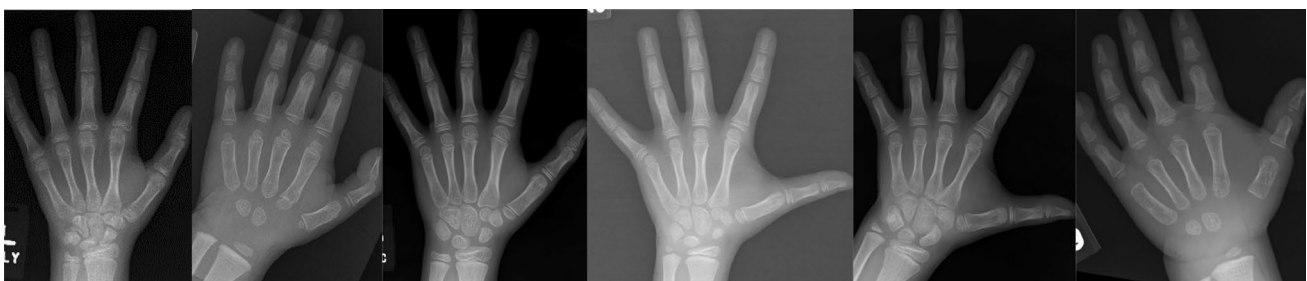


Fig. 6 Cropped and histogram equalized images

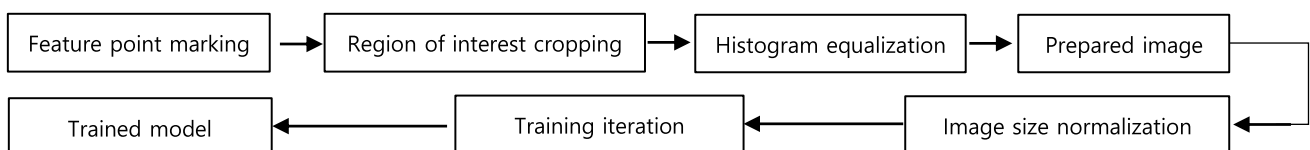


Fig. 7 Overall processing flow—preprocessing and deep learning stages

$e = \hat{a} - a$ and absolute error as $e = |\hat{a} - a|$

The sum of absolute errors is $\text{sum}_e = \sum_{k=1}^n |\hat{a} - a|$, and mean absolute difference (MAD) error is $\text{MAD} = \sum_{k=1}^n |\hat{a} - a|/n$. Often in clinical applications, it is needed or critical to have a low level of fluctuations in performance. Concordance correlation coefficient (CCC) [16] quantifies the stability in performance, which is defined as:

$$\rho_c = \frac{2\rho\sigma_x\sigma_y}{\sigma_x^2 + \sigma_y^2 + (\mu_x - \mu_y)^2}$$

Here ρ is Pearson's correlation coefficient, σ is standard deviation and μ is mean.

2.5 Neural network: architecture and weight setting

Deep learning neural network is typically built up by cascading a series of layers each performing distinct kinds of operations. An example architecture is illustrated in Fig. 8. Convolutional layer consists of two dimensional array of tiles which perform numerical convolution operations on inputs which are processed results from previous stages. Max pooling layer typically follows a convolution layer and performs maximum response value sampling. Fully connected layer has exhaustive edge connections to neurons from its preceding layer. Layers in later stages in the network get tuned to more specific patterns of image during the course of training and exhibits higher response specificity. The release version of CaffeNet and many others assume classification tasks. The comprising layers of networks were changed to suit to the present regression task.

2.6 Training iteration

Training deep learning network aims at finding a suitable performance level as close as possible to global optimum. A number of parameters are involved in the training process. Typically, deep learning process entails the determination of a large number of weight values for edges. On the other hand, the amount of data available for training are far less—often counting tens to hundreds. This imbalance may lead to overfitting of model to training data where model is highly tuned to supplied training data and fails to generalize to unseen data.

It is often believed the primitives that compose image characteristics such as edges, corners and blobs are rather general to recur over different image instances. Then specificity of an image instance is determined by the combination and spatial arrangement of the primitives. The layers which typically are placed in the later part of neural network including convolutional layers and fully connected layers will be suitable for finding such arrangement.

Imagenet [17] is an image database which includes more than one million image instances in more than a thousand image categories. This order of training data set size may facilitate a more proper determination of edge weights than a smaller size training data set. Pre-trained weight configuration files on ImageNet data are available for most of widely used deep learning architectures, which were obtained after a large number of training iterations. Hence an approach to remedying the data size imbalance issue was to take the pretrained weights as a starting configuration and apply problem-specific data to fine tune the weights towards specific learning goal.

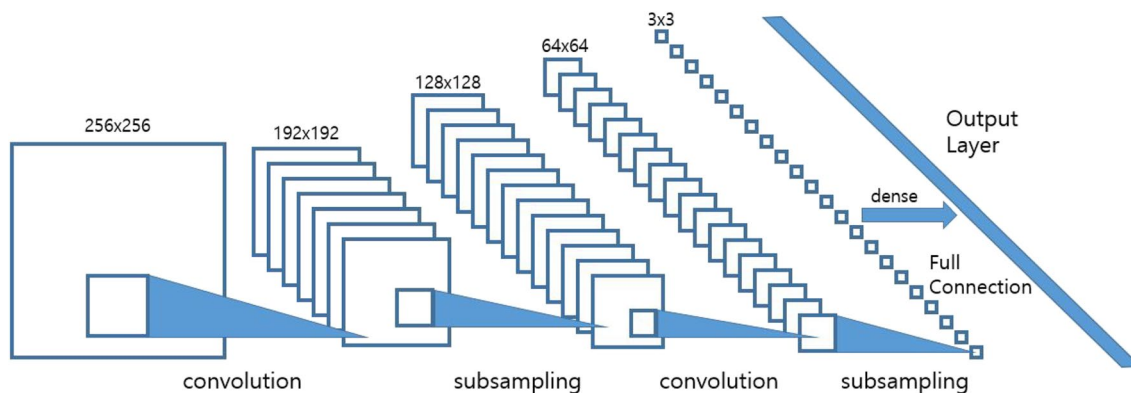


Fig. 8 An illustration of deep learning neural network architecture

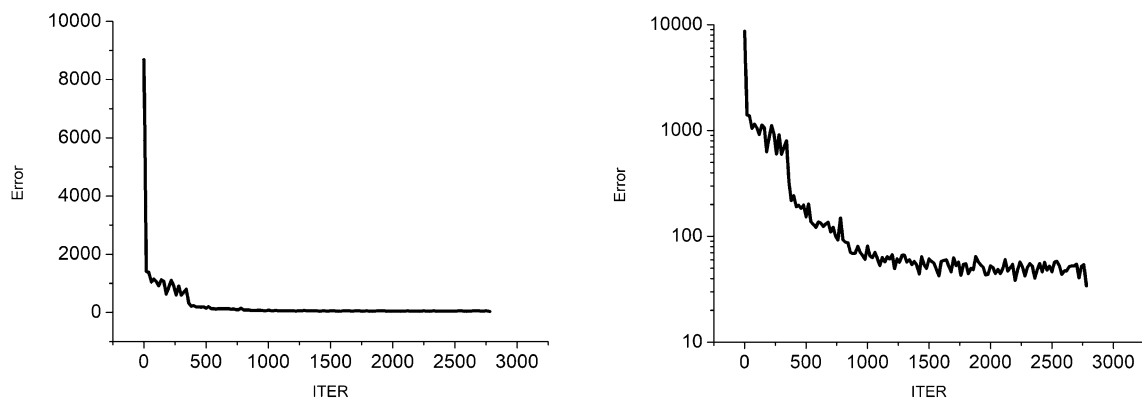


Fig. 9 Loss error versus iteration in linear and in log scales

Table 1 Error levels of gender agnostic and gender-aware models

	MAD (male, female)
Caffenet, gender-agnostic	16.933
Caffenet, gender-aware	12.649 (12.272, 13.026)
GoogleNet, gender-agnostic	12.642
GoogleNet, gender-aware	10.380 (10.186, 10.573)

MAD mean absolute difference

3 Results

The RSNA dataset comprises of more than 12 thousand images of hand bone X-rays. In following, 2000 randomly selected images were used for training and 800 images for testing.

3.1 Gender agnostic models and gender aware models

The established maturity rate difference due to gender difference implies a level of growth is attained at distinct ages by male and female cohorts on average [18, 19]. Hence needed are models which reflect subjects' gender. Figure 9 indicates decrease in errors in linear and log scales from a session of training iterations. There was substantial difference in performance between models that distinguish gender and do not (Table 1). For Googlenet, the mean MAD of gender aware models was 10.380 while the MAD of gender agnostic models was 12.642. These results clearly show gender aware models fare better than gender agnostic models which is in accordance with well known gender-wise growth rate difference. Overall higher level of MAD errors is seen for female cohort. It may be hypothesized that relatively higher growth rate of female cohort causes

Table 2 Performances by different network architectures

	MAD (female)	CCC
Caffenet	12.272	0.904
GoogleNet	8.890	0.941
Resnet	15.366	0.855

greater deviation from the nominal growth trajectory for individual subjects.

3.2 Network architectures

Since the introduction of the Alexnet [20], numerous architectures have been proposed. In addition to the sequential arrangement of the distinct kinds of layers, they may be distinguished one from another by the depth and the inter-connection between the layers. Caffenet has a rather shallow 8 layers deep architecture. On the other hand, Resnet used in here is made up of 50 layers far surpassing the layer count of other architectures. In this experiment, we compared different deep learning architectures—Caffenet, Resnet and Googlenet. Gender specific models were used in what follows (Table 2).

Overall googlenet showed the lowest error of 8.890 months. Caffenet showed a decent 12.272 months error, despite its simple network structure. Increased number of layers in general necessitates larger number of operations. Hence the adoption of more complex architectures may be justified only by commensurate increase in performance. In this regard, Caffenet may be regarded as a simple yet effective architecture. The performance of Resnet is somewhat counter-intuitive, considering the highest level of complexity of its architecture. A more carefully set training parameters may realize its full performance potential.

Table 3 Error levels obtained under different ROI schemes

	MAD (female)	CCC
ROI 1	8.890	0.944
ROI 2	9.699	0.931
ROI 3	10.209	0.922

Googlenet was used

3.3 ROI schemes

Hand in X-ray images may appear in a wide range of angular and locational displacements. Despite the learning capacity of deep learning networks, irrelevant appearance variability of objects may negatively impact the inferential capability and hence lower their overall performance. In particular, fingers typically assume wider variability in pose, contrary to the palm. In the third experiment, we tried to assess the tradeoff between the informativeness and age-unrelated variation of different hand regions by comparing performances obtained with ROI 1, 2 and 3. Overall, using the smallest ROI produced the best result (Table 3).

For note, image augmentation did not produce substantial performance difference from the case of no data augmentation (data not shown) [21].

4 Discussion

Traditional hand bone age atlases relied on data that were compiled decades ago which may not reflect the stature and growth rate changes over the many generations. Time and labor cost consideration rather discourages atlas based manual estimation. We assessed the trade off and performance implication from including different hand bone regions. Rather surprisingly, mostly relying on carpal and metacarpal bones while nearly excluding phalanges produced lowest level of estimation errors. This suggests entire hand region images do not have to be included in estimating bone ages and having palm regions intact may be just sufficient.

With a rather simple approach, a quite decent error level of 9.35 months was achieved. Males and females are known to follow different rates of growths, with boys lagging girls by 12–18 months in attaining growth milestones. Accordingly, models distinguishing genders were built. A more elaborate estimation method could incorporate subroutines such as region segmentation to further reduce age-unrelated image variation and feature point detection for image registration. Compared with full-fledged estimation method, our approach requires minimal development efforts and still produces a decent performance level to be competent in clinical application fields. Possibly, a large part of performance is contributed by the use of deep learning network

which enables efficient mapping of input image to estimation objective in a close to optimal manner.

It is well known infants show larger variance from nominal bone growth states than older ages, and clinical data in the range is far scarcer. Despite the difficulty and negative performance implication from dealing with infant data, we did not exercise data exclusion based on specific age range criteria and included entire age ranges data in our study. Further segregating data by age ranges and training models separately may improve overall estimation performance.

We plan to implement aforementioned routines to realize a fully streamlined end to end age estimation process and to improve performance. The aging process as manifested in hand bones is known to be rather nonlinear with particularly rapid morphological changes taking place in certain age intervals. Then there's chance a classification neural network may fare better than regression model for such highly nonlinear input–output response relation. Classification neural network may also be implemented and its performance compared with regression network in our future work.

The practice of bone age estimation typically purposes to assess whether a subject is in a normal growth regime and to diagnose any growth problems of diverse origins, etc. So having chronological age information of subject available is a clinically reasonable assumption and then it's plausible to build an estimation model with such information available. The available age information may be used to performance advantage. We also plan to implement such model in the future and assess its performance.

Acknowledgements This work was partly supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) (No. 2019-0-01750, Development of an optimal limb-compressing cardiovascular treatment using deep learning technique) and Gachon University Gil Medical Center (2018-5283), the GRRC program of Gyeonggi province (GRRC Gachon 2017-B01).

Compliance with ethical standards

Conflict of interest The authors declare that there is no conflict of interest regarding the publication of this paper.

Ethical statement The research is done following all the ethics guidelines provided by the Springer.

References

1. Greulich WW, Pyle SI. Radiographic atlas of skeletal development of the hand and wrist. 2nd ed. Stanford, CA: Stanford University Press; 1959.
2. Tanner JM, Whitehouse RH. Clinical longitudinal standards for height, weight, height velocity, weight velocity, and stages of puberty. *Arch Dis Child.* 1976;51(3):170–9.

3. Visual dictionary [cited at 2018 Jan 5]. <http://visual.merriam-webster.com/human-being/anatomy/skeleton/hand.php>.
4. RSNA. Pediatric Bone Age Challenge. IL, USA: RSNA [cited at 2017 Dec 13]. <http://rsnachallenges.cloudapp.net/competitions/4>.
5. Vicente G, Ratib O. Hand bone age: a digital atlas of skeletal maturity. Berlin: Springer; 2005.
6. Gaskin CM, Kahn SL, Bertozzi JC, Bunch PM. Skeletal development of the hand and wrist: a radiographic atlas and digital bone age companion. Oxford: Oxford University Press; 2011.
7. Thodberg HH, Kreiborg S, Juul A, Pedersen KD. The BoneXpert method for automated determination of skeletal maturity. *IEEE Trans Med Imaging*. 2009;28(1):52–66 (**abstract**).
8. Kim JR. Computerized bone age estimation using deep learning based program: evaluation of the accuracy and efficiency. *Am J Roentgenol*. 2017;209(6):1374–80. <https://doi.org/10.2214/AJR.17.18224>
9. Spampinato C, Palazzo S, Giordano D, Aldinucci M, Leonardi R. Deep learning for automated skeletal bone age assessment in X-ray images. *Med Image Anal*. 2016;36:41–51.
10. Wikipedia. Backpropagation. Wikipedia; c2017 [cited at 2017 Dec 11]. <https://en.wikipedia.org/wiki/Backpropagation>.
11. Srivastava Nitish, Hinton Geoffrey, Krizhevsky Alex, Sutskever Ilya, Salakhutdinov Ruslan. Dropout: a simple way to prevent neural networks from overfitting. *J Mach Learn Res*. 2014;15:1929–58.
12. BAIR. Caffe Installation. CA, USA: BAIR [cited at 2017 Dec 13]. <http://caffe.berkeleyvision.org/installation.html>.
13. BVLC/caffe. c2017 [cited 2017 Dec 13]. github.com/BVLC/caffe/tree/master/models/bvlc_reference_caffenet.
14. Szegedy C, Liu W, Jia Y, Sermanet P, Reed S, Anguelov D, Erhan D, Vanhoucke V, Rabinovich A. <https://arxiv.org/abs/1409.4842>.
15. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. <https://arxiv.org/abs/1512.03385>.
16. Lawrence I-Kuei Lin. A concordance correlation coefficient to evaluate reproducibility. *Biometrics* 1989;45(1):255–268. <https://doi.org/10.2307/2532051>.
17. Deng J, Dong W, Socher R, Li LJ, Li K et al. ImageNet: a large-scale hierarchical image database. In: Proceedings of IEEE conference on computer vision and pattern recognition. 2009.
18. Ontell FK, Ivanovic M, Ablin DS, Barlow TW. Bone age in children of diverse ethnicity. *AJR*. 1996;167:1395–8.
19. Zhang A, Sayre JW, Vachon L, Liu BJ, Huang HK. Racial differences in growth patterns of children assessed on the basis of bone age. *Radiology*. 2009;250:228–35.
20. Romany FM. Deep-learning-based automatic computer-aided diagnosis system for diabetic retinopathy. *Biomed Eng Lett*. 2017;8(1):41–57.
21. Salamon J, Bello JP. Deep convolutional neural networks and data augmentation for environmental sound classification, 2016, *IEEE Signal Proc Lett*.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.