Research on the Intelligent Assessment Algorithm of Bone Age Based on Attention Mechanism

X X Zhao¹, D Y Li¹, J Li¹, J Kang¹ and L Yang*¹

¹ China Telecommunication Technology Labs, China Academy of Information and Communications Technology, Beijing, 100191, China.
yanglei@caict.ac.cn

Abstract. Bone age is a reliable index to reflect the maturity of physical development, which is of great significance to evaluate the growth and development of children and adolescents, diagnosis and treatment of diseases. Traditional bone age assessment based on artificial has many problems, such as its time-consuming and subjective result, which may lead to great fluctuation of assessment results. Based on the X-ray image of the hand bones, this study proposes an intelligent prediction model of bone age in Deep Learning based on attention mechanism, combined with the traditional methods of bone age interpretation in Deep Learning. In the pre-processing stage, U-Net is used to remove the background of X-ray image of hand bones, and the dense connection network of attention mechanism is used to extract image features, and the mean absolute error function is introduced to improve the accuracy of this model. In the data set of RSNA competition, the mean absolute error of the method proposed in this study is 0.38 ± 0.10 years old, and obtained the best results reported at present.

1. Introduction

Bone Age Assessment (BAA) is one of the most commonly used clinical diagnostic techniques in pediatric radiology and legal medicine to evaluate the maturity of children’s bone development [1][2]. Traditionally, radiologists or endocrinologists make artificial interpretation according to the Greulich Pyle (GP) map method [1] or Tanner Whitehouse (TW) scoring method [2]. GP Atlas method compares hand X-ray image with Reference Atlas [1], while TW scoring method is a scoring system for examining a group of specific bone areas (such as epiphysis/metaphysis and carpus) and synthesizing all scores [2]. Generally, the GP diagram method is simpler, and the TW score method is more accurate and reproducible. However, both evaluation methods require not only doctors’ diagnostic experience, also cost a lot of time. For this reason, Thodberg et al. developed a computer program to assist medical workers to achieve robust and accurate evaluation results [3]. In this method, region of interest (ROI) is extracted from the image (such as the size or shape of a specific bone [4][5], and then regression /classification methods (such as morphological feature aggregation [5], support vector machine [6], or random forest [7]) are applied to predict bone age.

In recent years, deep learning has been widely used in medical image analysis. Compared with traditional machine learning, convolutional neural network in deep learning can automatically extract task-related features from the original image and complete the end-to-end learning task. In terms of bone age prediction, Spampinato et al. constructed BoNet architecture for bone age prediction, and compared its performance with three common CNN models (OverFeat, GoogLeNet and OxfordNet) [8]. Lee et al. used depth CNN to remove background and detect hand from X-ray image, and used
three CNN models (AlexNet, GoogLeNet and VGG-16) as a regression network [9]. A recent study used a depth residual network to evaluate bone maturity in hand X-ray images [10]. In the above research, CNN models of different depths are applied to the whole image [8][10] or the hand image [9] extracted from the original image, but little attention is paid to local features. Inspired by the TW scoring method, the local characteristics of independent anatomical structure may be significant for bone age assessment [11]. Local anatomical structures (such as hand bones and finger joints) can be segmented not only by traditional methods (such as active contour model [12]), but also identified by CNN structure [13]. Cao et al. used the bone age prediction method with multi region characteristics and obtained the best evaluation result [14] with error value less than 6 months. However, this scheme needs to locate the target region first, which improves the complexity of data preprocessing. Attention mechanism presents a pioneering approach for capturing long-range dependencies, via aggregating query-specific global context to each query position. Through attention mechanism, important features of sparse data can be quickly extracted, so as to improve the accuracy of network reasoning. It has been widely used in natural language processing and image analysis [15]-[18]. Therefore, by introducing attention mechanism, this study improves the weight of key parts in bone age prediction, weakens irrelevant areas, and replaces the process of regional training in reference [14].

The main work of this study includes:

1) The U-Net segmentation network [19] is used to remove the background of the picture. At the same time, in order to improve the accuracy of prediction, SSD (single shot multibox detector) target detection algorithm [20] is used to align the hand feature points, and image registration is realized by using the feature point information to carry out the image transformation[21][22].

2) In this paper, the DenseNet with attention mechanism [23][24] is proposed for X-ray bone age prediction. In feature extraction, the DenseNet with attention mechanism improves the importance of key areas in bone age prediction.

3) In this paper, the data of 2017 RSNA competition is used to train and verify the proposed method. The experimental results show that: Comparing with the existing bone age prediction methods, the prediction results of state-of-the-art are obtained in this paper, and the mean absolute error (MAE) of bone age is reduced to 0.38 ± 0.10 years old.

2. Materials and methods

In this study, the framework of bone age assessment is shown in Figure 1. First of all, this study used data preprocessing to improve the ROI region of the model input. Then, the Dense Block and the Attention Mechanism Block modules are used to develop a CNN’s model for bone age regression.

![Figure 1. Prediction framework of bone age.](image)

2.1. Data set

This paper is based on the 2017 RSNA competition data set, which is a high-quality X-ray bone age data set provided by Stanford University, University of Colorado and UCLA for the North American Radiology Society (RSNA) radioinformatics Committee (RIC) children's bone age machine learning challenge. The data set contains 12611 X-ray hand bone images, in which the age range of subjects is
0-19 years (228 months), and each image is labeled with the real bone age. The bone age distribution histogram of RSNA data set is shown in Figure 2.

![Figure 2. Bone age distribution histogram of RSNA data set.](image)

2.2. Data preprocessing
Data preprocessing is shown in Figure 3. In this study, U-Net [19] was first used as the segmentation network of the hand bones’ main body to achieve background removal. Then, image enhancement is achieved through grayscale conversion. In order to improve the prediction accuracy, this paper uses image cutting to reduce image noise outside ROI region and uses the SSD target detection algorithm [20] to detect the feature points in the hand bone area. Finally based on feature point coordinates, this paper uses Affine transformation to accomplish the image alignment [21][22].

![Figure 3. Image preprocessing process.](image)

2.3. Bone age prediction network based on attention mechanism

2.3.1. The Dense Block. In this study, the bone age prediction network is mainly composed of DenseNet's Dense Block (Figure 4). Figure 4 illustrates the layout of the resulting DenseNet schematically. Consequently, the layer receives the feature-maps of all preceding layers, $X_0, \ldots, X_{j-1}$ as input:

$$X_j = H\left([X_0, X_1, \ldots, X_{j-1}]\right)$$  \hspace{1cm} (1)

where $[X_0, \ldots, X_{j-1}]$ refers to the concatenation of the feature-maps produced in layers 0, 1, \ldots, $l-1$. The Dense Block is an iterative cascade of previous feature mapping. Each layer can directly access all previous layer features and fully reuse the previous feature mapping information. In addition, from the point of view of parameter use, dense connection modules are much more efficient than other modules [24].
2.3.2. The Attention Mechanism Block. The attention mechanism aims at strengthening the features of the query position via aggregating information from other positions. We denote \( X = \{ X_i \}_{i=1}^{N_p} \) as the feature map of one input instance, where \( N_p \) is the number of position in the feature map \( \left( N_p = W \cdot X \right) \). \( X \) and \( Z \) respectively denote the input and output of the Attention Mechanism Block, which have the same dimensions. The Weight coefficient of Attention Mechanism Block can then be expressed as:

\[
Z_i = X_i + W_z \sum_{j=1}^{N_p} \frac{f(X_i, X_j)}{C(X)} W_v X_j
\]

(2)

Where \( i \) is the index of query positions, and \( j \) enumerates all possible positions. \( f(X_i, X_j) \) denotes the relationship between position \( i \) and \( j \), and \( C(X) \) is a normalization factor. \( W_z \) and \( W_v \) denote linear transform matrices. The paper denotes \( W_{ij} = \frac{\exp(\langle X_i, X_j \rangle)}{\sum_m \exp(\langle X_i, X_m \rangle)} \) as normalized pairwise relationship between position \( i \) and \( j \). The Attention Mechanism Block is illustrated in Figure 5.

**Figure 4.** Dense connection module.

**Figure 5.** Architecture of the Attention Mechanism Block.
The feature maps are shown by their dimensions, e.g., CxHxW. Symbol ⊙ is matrix multiplication, and symbol ⊕ is broadcast element-wise addition. For two matrices with different dimensions, broadcast operations first broadcast features in each dimension to match the dimensions of the two matrices. The network results used in this study are shown in Table 1. The input of DenseNet is 224 × 224 normalized gray-scale X-ray image; the initial convolution layer contains 2K convolutions, in which K is the growth rate of the characteristic image, the convolution size is 7 × 7, and the step size is 2; then four dense connection modules are connected, and the number of characteristic images in all other layers is also determined according to the parameter k. The attention mechanism is used to replace the global average pooling (GAP) [25] layer in the original DenseNet structure.

Table 1. The Arrangement of Channels

| Layers                  | Output Size | DenseNet+ (k=32) |
|-------------------------|-------------|-------------------|
| Convolution             | 112×112     | 7×7 conv, stride=2|
| Pooling                 | 56×56       | 3×3 max pool, stride=2|
| Bottleneck Block(1)     | 56×56       | [1×1 conv] ×6     |
| Transition Layer(1)     | 28×28       | [1×1 conv] ×12    |
| Bottleneck Block(2)     | 28×28       | 1×1 conv          |
| Transition Layer(2)     | 14×14       | 2×2 avg pool, stride=2 |
| Bottleneck Block(3)     | 14×14       | [1×1 conv] ×24    |
| Transition Layer(3)     | 7×7         | 1×1 conv          |
| Bottleneck Block(4)     | 7×7         | 2×2 avg pool, stride=2 |
| Attention Mechanism     | 7×7         | 1×1 conv          |
| Block                   | 1×1         | 1000D fully-connected, softmax |

3. Results and Discussion

3.1. Loss function

In reference [26], the author proved that regression model is better than classification model in solving bone age prediction task. Because MSE function has the mathematical properties of convexity and strict differentiability, all the regression models of bone age prediction, without exception, adopt MSE loss function with excellent performance and easy to optimize. MSE is defined as:

\[ MSE = \frac{1}{N} \sum_{i=1}^{N} \left( \bar{y}_i - y_i \right)^2 \]  

(3)

Where, \( y_i \) and \( \bar{y}_i \) represent the real value and predicted value respectively, and \( N \) is the number of samples.

In contrast, MAE loss function is not completely smooth, which makes optimization difficult. Although there is no research on the application of MAE cost function in bone age prediction, MAE cost function has achieved great success in facial age estimation [26]. Therefore, this paper discusses whether MAE cost function has good prediction performance in bone age automatic prediction. MAE losses are defined as:
\[ MAE = \frac{1}{N} \sum_{i=1}^{N} |y_i - \hat{y}_i| \]  

(4)

3.2. Experiment configuration
In the experiment, the hand bone X-ray image data set (12611 in total) was divided into three sub sets: training set, test set and verification set. The number of samples in each subset is 9458, 2020 and 1133 respectively. All the X-ray images of hand bones were first adjusted to 1300 × 1300 by filling operation, and then scaled to 224 × 224 and input into the network. All models are implemented on NVIDIA K40 GPU.

In the experiment, the same dataset is used to explore influence of different loss functions and network structures on prediction results. Note that DenseNet MAE didn't use the Attention Mechanism Block which DenseNet+ MAE use.

3.3. Image preprocessing and influence on prediction results
In order to generate the training image of U-Net network, this paper uses the semi-automatic segmentation method to remove the background of 3000 X-ray images from the hand bone X-ray image data set. Then the images are randomly divided into 4:1 training/testing data sets. The training parameters of U-Net are: batch size = 24, epoch = 5000, optimizer = Adam, learning rate = 0.001, loss function = cross entropy. Using the U-Net model after training, the background of the remaining 9611 images in the X-ray image data set of the hand bone was removed. The results are compared with the real situation, as shown in Table 2.

| Table 2. The Arrangement of Channel |
|-----------------------------------|
| ACCURACY | DICE | IOU | RECALL |
|----------|------|-----|--------|
| 0.98     | 0.85 | 0.84| 0.92   |

\[ \text{Accuracy} = \frac{TP}{TP + FP}; \text{Dice} = \frac{2 \times |\text{Seg} \cap \text{Re}|}{|\text{Seg}| + |\text{Re}|}; \text{IOU} = \frac{TP}{TP + FP + FN}; \text{recall} = \frac{TP}{TP + FN}. \]

\[ a \text{ Accuracy}= \frac{TP}{TP + FP}; \text{Dice} = \frac{2 \times |\text{Seg} \cap \text{Re}|}{|\text{Seg}| + |\text{Re}|}; \text{IOU} = \frac{TP}{TP + FP + FN}; \text{recall} = \frac{TP}{TP + FN}. \]

\[ b \text{ where, TP: true positives; FP: false positive; FN: false negatives; TN: true negatives; Ref is the ground truth and Seg is the segmented results by U-Net.} \]

3.4. Influence of Different Loss Functions on Prediction Result
In the research, MAE and MSE are used as loss function to train DenseNet+ model. Network training parameters: batch size = 32, epoch = 1000, optimizer = Adam, learning rate = 0.001. Except for the difference of loss function, the network training parameters of the two models are the same, input data sets are the same and the network structure is the same. The experimental results of the two models are compared in Table 3.

| Table 3. Results of Different Loss Function Densenet + Models |
|-------------------------------------------------------------|
| Model            | Mean Absolute Error |
|------------------|---------------------|
| DenseNet+ MAE    | 0.482±0.103         |
| DenseNet+ MSE    | 0.564±0.183         |

The experimental results show that the performance of DenseNet+ MAE is better than that of DenseNet+ MSE. First of all, since MAE is a loss function, optimizing MAE will naturally improve the performance of MAE. In addition, MAE is less sensitive to outliers than MSE. Because MSE loss will square the error before the average error, the model trained by MSE usually pays more attention to reduce the larger error. However, in practice, the tag noise is inevitable, so the model with MSE loss
will sacrifice other accurate tags and concentrate on reducing a small number of inaccurate tag errors. Therefore, the performance of DenseNet+_MAE is better than that of DenseNet+_MSE in all evaluation items.

3.5. Influence of Different Network Structures on Prediction Results

Using the same X-ray image data set of hand bone, the DenseNet+_MAE method proposed in this paper is compared with other BAA methods. The results are shown in Table 4. It can be seen from Table 4 that the dense network with MAE as loss function has better prediction accuracy. In addition, in DenseNet, the accuracy of bone age prediction is further improved after the attention mechanism is integrated, which is mainly because the model pays more attention to epiphysis/metaphysis and carpal region after the attention mechanism is added (Figure 6), which is consistent with the method of TW scoring bone age judgment mainly based on such regions, further proving that the network model of attention mechanism is helpful for the analysis to improve the prediction accuracy of BAA.

Table 4. Comparison of Accuracy of Several Bone Age Automatic Prediction Algorithms

| Method           | Time | MAE   |
|------------------|------|-------|
| Giordano et al [27]. | 2016 | 1.82  |
| Fine-tuned GoogLeNet [8]. | 2016 | 0.82  |
| DenseNet[24].     | 2018 | 0.73  |
| Cao et al [14].   | 2019 | 0.52  |
| DenseNet_MAE      | 2020 | 0.62±0.09 |
| DenseNet+_MAE     | 2020 | 0.38±0.10 |

(a) The actual bone age is 11.5 years, and the predicted bone age is 11.2 years
(b) The actual bone age is 8 years old, and the predicted bone age is 7.6 years old

Figure 6. Schematic diagram of attention mechanism.

4. Conclusions
This study builds a bone age prediction process based on deep learning algorithm, elaborates the data preprocessing process in detail, builds and trains a dense depth convolution neural network model integrating attention mechanism, and analyzes the influence of different loss functions on the model accuracy. Through the data set of RSNA competition, the best prediction accuracy is obtained.

5. References
[1] Bayer L M 1959 Radiographic atlas of skeletal development of the hand and wrist: second edition. *California medicine*, 91(1).
[2] Ehrenberg A, Tanner J M, Whitehouse R H, Marshall W A and Goldstein H 1977 Assessment of skeletal maturity and prediction of adult height (twii-method) Applied Statistics, 26(1), p.80.
[3] Martin D D, Schittenhelm J and Thodberg H H, 2016 Validation of adult height prediction based on automated bone age determination in the paris longitudinal study of healthy children Pediatric Radiology, 46(2), p.263-269.
[4] Henrik T H 2009 An automated method for determination of bone age The Journal of Clinical Endocrinology & Metabolism, vol.94, p.2239-2244.
[5] Thodberg H H, Kreiborg S, Juul A and Pedersen K D 2009 The bonexpert method for automated determination of skeletal maturity IEEE Transactions on Medical Imaging, 28(1), p.52-66.
[6] Davis L M, Theobald B J, Lines J, Toms A and Bagnall A 2012 On the segmentation and classification of hand radiographs International journal of neural systems, vol.22, p.1250020, 2012.
[7] Stern D, Ebner T, Bischof H, Grassegger S and Urschler M 2014 Fully automatic bone age estimation from left hand Mr images In International Conference on Medical Image Computing and Computer-Assisted Intervention, p.220-227.
[8] Spampinato C, Palazzo S, Giordano D, Aldinucci M and Leonardi R 2017 Deep learning for automated skeletal bone age assessment in x-ray images Medical image analysis, 36, p.41-51.
[9] Lee H, Tajmir S, Lee J, Zissen M, Yeshiwas B A, Alkasab T K and et al 2017 Fully automated deep learning system for bone age assessment Journal of digital imaging, vol.30, p.427-441.
[10] Larson D B, Chen M C, Lungren M P, Halabi S S, Stence N V and Langlotz C P 2018 Performance of a deep-learning neural network model in assessing skeletal maturity on pediatric hand radiographs Radiology, 287(1), p.313.
[11] Stern D, Payer C, Lepetit V and Urschler M 2016 Automated age estimation from hand MRI volumes using deep learning In International Conference on Medical Image Computing and Computer-Assisted Intervention, p.194-202.
[12] Cootes T F, Edwards G J, Taylor C J 2001 Active appearance models IEEE Transactions on pattern analysis and machine intelligence, vol.23, p.681-685.
[13] Lee S, Choi M, Choi H S and et al FingerNet: Deep learning-based robust finger joint detection from radiographs in 2015 IEEE Biomedical Circuits and Systems Conference (BioCAS) p.1-4.
[14] Cao S, Chen Z, Li C and et al Landmark - based multi - region ensemble convolutional neural networks for bone age assessment International Journal of Imaging Systems and Technology, vol.29, p.457-464.
[15] Luong M T, Pham H, Manning C D 2015 Effective Approaches to Attention-based Neural Machine Translation Computer Science.
[16] Xu K, Ba J, Kiros R and et al 2015 Show, Attend and Tell: Neural Image Caption Generation with Visual Attention Computer Science, p.2048-2057.
[17] Yang Z, He X, Gao J and et al 2016 Stacked Attention Networks for Image Question Answering 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR).
[18] Hu J, Shen L, Sun G 2018 Squeeze-and-Excitation Networks 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).
[19] Ronneberger O, Fischer P, Brox T 2015 U-Net: Convolutional Networks for Biomedical Image Segmentation in International Conference on Medical image computing and computer-assisted intervention, p.234-241.
[20] Liu W, Anguelov D, Erhan D and et al 2016 SSD: Single Shot MultiBox Detector In European conference on computer vision, p.21-37.
[21] Wang J, Shi J, Wu X X 2008 Survey of image mosaics techniques Application research of computers, 56(4), p.517-320.
[22] Pluim J P, Maintz J A and M A 2003 Mutual-Information-Based Registration of Medical Images: A Survey IEEE transactions on medical imaging, vol.22, p.986-1004.

[23] Wang F, Jiang M, Qian C and et al 2017 Residual Attention Network for Image Classification In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition p.3156-3164.

[24] Huang G, Liu Z, Laurens V D M and et al 2016 Densely Connected Convolutional Networks in Proceedings of the IEEE conference on computer vision and pattern recognition, p. 4700-4708.

[25] Lin M, Chen Q, Yan S 2013 Network in Network Computer Science, arXiv preprint arXiv:1312.4400.

[26] Xing J, Li K, Hu W, Yuan C and Ling H 2017 Diagnosing deep learning models for high accuracy age estimation from a single image Pattern Recognition the Journal of the Pattern Recognition Society, vol.66, p.106-116.

[27] Giordano D, Kavasidis I, Spampinato C 2016 Modeling skeletal bone development with hidden Markov models Computer Methods and Programs in Biomedicine, vol.124, p.138-147.