Bone Age Estimation with X-ray Images Based on EfficientNet Pre-training Model

Guoyao Hao1, a, †, Yifei Li 2, b, †
1Rutgers University, Piscataway, NJ 08901, USA;
2Chongqing University of Technology, Chongqing, 401135, China.
agh343@scarletmail.rutgers.edu, b11719370423@2017.cqut.edu.cn.
†These authors contributed equally.

Abstract. Human bones have different characteristics in different development stages, so the estimation of bone age can reflect the growth and development level of individuals relatively accurately. Bone age estimation aims to predict the biological age of children, which plays an important role on the diagnosis of some pediatric endocrine diseases. Tradition methods are carried out by doctors, and it is not effective in accuracy and speed. To this end, we proposed a deep-learning based method for bone age estimation. Based on the training set of more than 10000 X-ray images of hand bones from Radiological Society of North America (RSNA), this paper studies the processing, segmentation, feature extraction of X-ray hand bone images by using computer image processing and artificial intelligence learning methods, and uses convolution neural network to process samples and analyze them automatically. The main research work and achievements are as follows: (1) Pre-processing of X-ray hand bone image, unifying the size and cutting, reducing the image area without hand bone; (2) The gray-scale image is transformed into a three-channel image, and pre-processing by EfficientNet of ImageNet. Then convolution neural network is used to learn the features of X-ray hand bone image and evaluate it automatically. Finally, the network is evaluated by the minimum mean square error, so that the minimum mean square error is as close as possible to the minimum value. Through the neural network, the bone age from X-ray hand bone image can be quickly judged, and then it can be applied to clinical research.

1. Introduction
In recent years, with the development of computer science and technology and medical technology, these disciplines gradually produced a lot of cross parts, and had a significant impact on human beings. At present, medicine can use computer image processing technology to carry out various medical activities, such as X-ray photo scanning and analysis. In medicine, human growth is based on age, while biological age can be inferred from bone age. X-ray is used to take regional photos of human left wrist. Bone age plays an important role in many fields, such as adolescent development assessment, genetic disease screening, or talent selection. However, at present, X-ray evaluation is mainly carried out manually. However, this method has great limitations, is easy to make mistakes, and takes a lot of time.

At present, there are many related X-ray bone age image learning projects at home and abroad. With the development of computer technology and graphics technology, the innovation of artificial intelligence technology is constantly emerging, which provides more methods for people to solve interdisciplinary problems. Bone age assessment is also adapted to it, from human judgment to batch evaluation of artificial intelligence deep learning. In this paper, the EfficientNet pre-training model [1]
is based on the three dimensions - network depth, network width, image resolution - to make estimations, and explore the best combination of the three. Therefore, EfficientNet's model size to accuracy ratio is far ahead of other traditional models. In addition, there are similar studies in Applied Science. Researchers have made a system to process X-ray image. The image is pre-processed by a separable convolutional neural network, which becomes a standardized and normalized form. Image registration is performed by segmenting the hand region. The position of the region of interest is aligned in the middle of the image. The hand region is rotated by the angle of four key points. Using deeplab V3 plus Architecture [2] and mobilenet V1 Architecture [3], separable convolution is used as the core arithmetic unit. Finally, in order to avoid the defect of underfitting, different rotation angles, scaling factors and clipping images are used to generate composite data to enhance the training data set. However, previous research projects did not attempt to estimate the bone age of X-ray hand bone images under the EfficientNet pre-training model, so we hope to find a fast and convenient method to process X-ray hand bone images using the existing pre-training model. One can refer to [4-9] and references therein for more details.

Therefore, based on computer deep learning and image processing, this project uses the training set of RSNA 2017 [10] more than 10000 X-ray images of children's hand bones, mainly including image information and age and gender information, to study the pre-processing and segmentation of X-ray hand bone images, feature extraction and automatic assessment of bone age. (1) The pre-processing method of X-ray photograph was studied. Because the size and position of each image in the training set are different, a unified algorithm is needed for cutting. After noise reduction, histogram equalization is used to make the gray distribution uniform. We propose a unified method to cut and store the extracted hand bone images for further deep learning. (2) The EfficientNet pre-training model on ImageNet data set is used to process the hand bone cutting image before. The effectiveness and efficiency of EfficientNet in feature acquisition of hand bone image are studied. (3) The deep learning of a convolution neural network is studied, and the features of X-ray hand bone image are extracted automatically. The processed vector is processed through the full connection layer to get an accurate bone age number, and then compared with the standard answer of the validation set. For proof of the effectiveness of the proposed method, the experiments on the RSNA X-ray dataset validate the effectiveness of our method. The MSE on the test set is average 219, so we conclude that our method is capable of making a precise prediction on bone age.

2. Methods

2.1 Introduction to dataset

The X-ray image data of this project is from RSNA 2017, a competition for the correct identification of children's age through X-ray photos of hands, provided by the children's bone age machine learning challenge Organization Committee of the Radiology Information Committee of the Radiological Society of North America. This data set contains more than 10,000 X-ray images of children's hand bones, mainly including images, age and gender information, and the resolution and size of each image are not uniform. The size of the hand bone image in the data set is different, but the statistics show that the length and width of the image range from 800 to 2200. We divide the X-ray images of RSNA into 70% training set, 15% validation set and 15% test set.
2.2 Introduction of methods

2.2.1 convolution neural network

Compared with the traditional neural network, the convolution neural network is still a hierarchical network, but as an improved version of the traditional neural network, the function and form of its layer have changed greatly.

The architecture of a convolutional neural network is similar to that of a conventional artificial neural network, especially in the last layer of a fully connected layer. Moreover, the convolutional neural network accepts the input of multiple feature graphs rather than vectors. (1) Data input layer: this layer is mainly used to pre-processing the original image data, including de-averaging, normalization, whitening. (2) Convolution computing layer: this layer is the most important layer of convolution neural network, mainly for local correlation and window sliding operation. The local correlation is that each neuron is regarded as a filter, which is called the convolution kernel, and the depth of each neuron is the same as the number of neurons. Window sliding is to use convolution check local data to calculate, and finally, form a new matrix. (3) Excitation layer: this layer maps the output of the convolution layer nonlinearly. The convolution neural network generally uses the ReLU excitation function, which has the characteristics of fast convergence speed and simple tangent gradient. (4) Pooling layer: this layer is between the convolution layer and the convolution layer of continuous convolution. It is used to compress data and parameters and reduce overfitting. In the pooling layer, the scale of features will not change due to compression. Only some irrelevant information is removed, leaving features that can express the image. In addition, when there are many features and some information is not useful for the task, we will use a pooling layer to extract features, remove redundancy and prevent overfitting. (5) Full connection layer: this layer is usually located at the tail of the convolutional neural network. Similar to the traditional neural network connection, the neurons between layers are connected by weights.

Convolutional neural network is superior to traditional neural network because it shares convolution kernel, processes high-dimensional data at high speed, and does not need to manually select features. As long as the weight is trained properly, it has good feature classification. Therefore, we choose convolutional neural network to extract features of X-ray hand bone image.

2.2.2 Transfer learning

With the demand for more and more machine learning, the application space of the pre-training model is larger and larger. A good performance of supervised learning requires a lot of labeled data, which is a
huge project, so the concept of transfer learning is born. Traditional machine learning needs to be based on the same distribution hypothesis, which requires a lot of annotation data. However, in practice, different datasets may have problems such as abnormal data distribution and expired annotation data. Therefore, making full use of the existing labeled data has become the top priority of migration learning.

There are two important concepts in transfer learning: domain and task. Domain is a characteristic domain at a certain moment, and the task is a transaction that needs to be done. Therefore, the transfer learning algorithm based on feature selection will know how to find the common features between the source domain and the target domain and use these features to transfer knowledge. In this way, in a space, the distribution of the source domain data and the target domain data is the same. In the new space, the existing labeled data samples of the source domain can be used for classification training, and finally the target domain data can be classified and tested.

In deep learning, a large number of high-quality annotation data are usually needed. Pre-training and fine-tuning are very popular skills in deep learning. Therefore, ImageNet, which is pre-training, is often selected to initialize the model.

2.3 Model design block diagram and ideas
In this work, our method consists of data pre-processing, feature extraction and aggregation, prediction. Since the image has various sizes in the dataset, we need to apply data pre-processing. Furthermore, we apply data enhancement to improve the diversity of training data. In the feature extraction process, both features, including image and gender feature are extracted, and then the features are aggregated via channel concatenation. Finally, the aggregated feature is sent to fully connected (FC) layers for bone age prediction.

![Figure 2. The image is the design steps of this experiment](image)

2.3.1 Data pre-processing
Data pre-processing contains 4 steps including, image resize and normalization, data enhancement, histogram equalization and channel padding [11]. First, because the size and position of X-ray hand bone images provided by RSNA are different, we need a unified hand bone image to learn the pre-training model. We fill the hand bone image into a square image of the same size, and try to make the hand bone in the middle position. Second, we fine tune the image by rotating and adding Gaussian noise to make the training set more diversified. Then, histogram equalization is carried out to make the gray changes tend to be balanced, so as to increase the contrast of hand bones. Finally, the grayscale image is transformed into a three-channel image and zoomed to meet the input of the convolution neural network in the next step.

2.3.2 Feature extraction and aggregation
As for image data, we use EfficientNet for feature extraction. Transfer learning can greatly shorten the model training time. Therefore, we decided to experiment with the existing pre-training model and fine tune the dataset. The EfficientNet parameter is initialized as the pre-training parameter of ImageNet dataset, and then the model is fine-tuned on our dataset to speed up the network training. Through the EfficientNet-B3 model, 1536 dimensional vectors are extracted from the image as image features. As for the gender feature, the gender information in the original dataset is Boolean. Firstly, the boolean type is transformed into float, and then it is extended to 64-dimensional gender features. In the experiment,
we will discuss how the size of the gender feature influences the prediction performance. Finally, both image and gender features are concatenated into a 1600-dimensional feature for prediction.

2.3.3 Prediction
The features are sent to a linear model, which consists of two FC layers, the first FC layer is related to gender, which is \((1536 + 1) \times 600\), and the second FC layer is \(600 \times 1\). After two FC layers, the prediction results are obtained.

2.4 Implementation details
The details of the parameters used in the experiment are as follows: hidden layer dropout = 0.6, AdamW learning rate = 0.0002, betas = (0.9, 0.999), weight decay = 0.01, CLAHE (Contrast Limited Adaptive Histogram Equalization) clip limit = 2.0, tile grid size = (8,8).

3. Results and Discussion
In this section, we conduct an analysis of the feature size of gender and the influence of pre-trained model. First, we describe the metric used in the experiment. Second, since the gender information is a boolean type with one dimension, while the image feature has 1536 dimensions, we hope to find the most suitable size for feature aggregation. Thus, we analyze the influence of gender feature size in section 3.2. Finally, inspired by Zulkifley et al., who said that the pre-trained model has a slight influence on the final result [12]. To discuss its influence in bone prediction, we apply the analysis with and without a pre-trained model, which is shown in section 3.3.

3.1 Experimental evaluation
For regression problems, we use MSE loss for model evaluation. MSE, namely mean squared error, is a convenient algorithm to balance the "average error." By evaluating the degree of data change, the smaller the MSE value is, which indicates that the experimental data described by the prediction model has better accuracy. MSE statistical parameter is the mean value of the sum of squares of the errors (SSE) corresponding to the predicted data and the original data.

The calculation formula is as follow (n is the number of samples):

\[
MSE = \frac{SSE}{n} = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2
\]

3.2 Effect experiment of the number of gender vectors operation results and analysis
After using the convolutional neural network of EfficientNet pre-training model, we get the following results. To use the loss function to judge whether the whole training result can meet the expected requirements, we need to minimize the loss function. We show the MSE curve during the training process. Therefore, we use the recursive function to evaluate the minimum loss value.

![Different Dimensional Gender Vector Validation](image)

Figure 3. The validation of gender vectors with different dimensions affects the convergence results.
Table 1. The MSE of test set with gender vector is smaller than that without gender vector, which is 108.99

| Gender Vector       | Test Set Loss |
|---------------------|---------------|
| 0 dimensional gender| 161.62        |
| 1 dimensional gender| 108.99        |

It can be seen that the convergence rate of the combination of 16-dimensional gender information and 1-dimensional gender information is the same, and the larger the batch, the smaller difference of the loss. However, if there are no gender characteristics, the convergence speed will be affected to some extent. Overall, there is a big difference between the gender vectors and the non-gender vectors, but the size of loss is not affected by how many dimensions of the gender vectors are set. The final loss of the 1-dimensional gender vector on the test set is 109.

3.3 Effect experiment with or without pre-training model operation results and analysis

And then, through a comparative experiment, we study whether the convergence rate of the pre-training model is better than that of the non-training model. Therefore, the following results are obtained:

![Figure 4. Influence of pre-training model on convergence results](image)

Table 2. Compared with no pre-training model, the MSE of test set with pre-training model is smaller, which is 108.99

| Pre-training Model | Test Set Loss |
|--------------------|---------------|
| with               | 108.99        |
| without            | 142.85        |

It can be seen that using the pre-training model can greatly speed up the convergence speed, and without using the pre-training model, the loss will be more difficult to converge, and the loss will be larger than using the pre-training model.

4. Conclusion and future work

Bone age estimation plays an important role in the evaluation of human development, detection, and prevention of diseases, and other fields. In this work, we propose a new method for bone age estimation, consisting of data pre-processing, feature extraction and aggregation, prediction. The data are pre-processed via resizing, normalization and data enhancement to remove the data bias and enlarge the training set. Then the image feature is extracted via EfficientNet and then aggregated with the gender feature. The fused feature is sent to the FC layer for prediction. Our method obtains a minimum MSE
of 109. We also analyze the influence of gender features and pre-trained model. Compared to the model without a gender vector, gender features can improve the model to higher accuracy. We find that the feature size of gender does not have an obvious influence on the final result. Furthermore, the pre-trained model can help fast convergence and obtain a better result. For future work, there is still a long way to go before the market and mass production can be realized. In the future, we mainly focus on (1) how to utilize the region of interest to obtain more targeted information; (2) data pre-processing.

Reference
[1] Mingxing Tan, Quoc V. Le. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. arXiv:1905.11946v5 [cs.LG]. 11 Sep 2020.
[2] Chen, L. C., Zhu, Y., Papandreou, G., Schroff, F., & Adam, H. Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation. European Conference on Computer Vision. Springer, Cham, 2018.
[3] Andrew G. Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, WeiJun Wang, et al. MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications. arXiv:1704.04861v1 [cs.CV]. 17 Apr 2017.
[4] Wenxiang Zhou. Automatic Evaluation of Bone Age Based on X-ray Images. (Doctoral dissertation).
[5] Xiaoying Pan, Yizhe Zhao, Hao Chen, De Wei, Chen Zhao, & Zhi Wei. Fully Automated Bone Age Assessment on Large-Scale Hand X-Ray Dataset. Hindawi, International Journal of Biomedical Imaging, Volume 2020, Article ID 8460493, 12 pages.
[6] Arsalan Manzoor Mughal, Nuzhat Hassan, & Anwar Ahmed. Bone Age Assessment Methods: A Critical Review. Pak J Med Sci. 2014 Jan-Feb; 30(1): 211–215. doi: 10.12669/pjms.301.4295.
[7] C.Spampinato, S.Palazzo, D.Giordano, M.Aldinucci, R.Leonardi. Deep learning for automated skeletal bone age assessment in X-ray images. Medical Image Analysis. Volume 36, February 2017, Pages 41-51.
[8] Sven Koitka, Aydin Demircioglu, Moon S. Kim, Christoph M. Friedrich, Felix Nensa. Ossification area localization in pediatric hand radiographs using deep neural networks for object detection. PLOS ONE, November 16, 2018.
[9] Vincenzo De Sanctis, Salvatore Di Maio, Ashraf T. Soliman, Giuseppe Raiola, Rania Elalaily, & Giuseppe Millimaggi. Hand X-ray in pediatric endocrinology: Skeletal age assessment and beyond. Indian J Endocrinol Metab. 2014 Nov; 18(Suppl 1): S63–S71. doi: 10.4103/2230-8210.145076.
[10] Halabi, S. S., Prevedello, L. M., Kalpathy-Cramer, J., Mamonov, A. B., Bilbily, A., & Cicero, M., et al. The RSNA Pediatric Bone Age Machine Learning Challenge. Radiology, 2018.
[11] Yiting Xie, David Richmond. Pre-training on Grayscale ImageNet Improves Medical Image Classification. European Conference on Computer Vision. Springer, Cham, 2018.
[12] Mohd Asyraf Zulkifley, Siti Rainahah Abdani, & Nuraisyah Hani Zulkifley. Automated Bone Age Assessment with Image Registration Using Hand X-ray Images. Applied Sciences, 2020.