Prediction of Adolescent Bone Age by Hand X-rays

Wang Chenyi(515030910362), Yu Linlin(515030910078),

Guo Xueqi(515030910103), Wang Zhenglong(515030910102)

Abstract

Bone age is an important indicator of bone and physical growth in adolescents and is widely used in many fields such as medicine, sports and forensic science. In view of the shortcomings of the existing automated bone age assessment methods, this report presents an attention-based deep convolutional neural network for automated bone age assessment. Combining with power feature learning capabilities of deep learning, the method can eliminate the need for tedious hand-crafted feature extraction in images in traditional work. After training, the model achieves a Mean Absolute Deviation (MAD) of 9.7845 months on a data set of 12611 hand bone X-ray images.

Ⅰ. Introduction

While automated approaches to bone age analysis have previously been developed and are commercially available today, none are widely available and radiologists are stuck with the task of flipping through the Greulich and Pyle atlas to find the most similar example every time they are presented with a bone age study. With such a cumbersome and dated method, bone age analysis is one of the “low hanging fruits” of medical imaging in this renaissance of artificial intelligence.

In 2017, the RSNA held a global machine learning competition to develop the best algorithm to predict skeletal age from pediatric hand x-rays. As a baseline, this experiment reproduces the procedure that one group have completed. With a model based on attention mechanism and several improvement over it, we eventually achieved a Mean Absolute Deviation (MAD) of 9.7845 months. Furthermore, several trials based on the original method will be put forward.

Ⅱ. Idea & Rationale

The algorithm is mainly based on the following rationale:

A. A few specific pieces of bones are primary factors in determination of bone age prediction.

The sample images used in this experiment are hand x-rays, which is donated by two U.S. hospitals. According to [1], hand skeleton can be divided into radius, ulna, metacarpal bone, proximal phalanx, middle phalanx, distal phalanx, etc. The following table shows the different weight of each bone.

Bones	Weights in CHN		Weights in TW2
	male	female	
radius	9.52	8.39	10
ulna	0.00	0.00	10
metacarpal bone I	1.97	4.51	3.34
metacarpal bone III	7.75	7.76	2.5
metacarpal bone V	4.53	6.61	2.5
proximal phalanx I	2.09	3.87	3.33
proximal phalanx III	11.19	9.21	2.5
proximal phalanx V	6.96	7.97	2.5
middle phalanx III	7.69	7.89	2.5
middle phalanx V	1.14	3.83	2.5
distal phalanx I	10.20	8.61	3.33
distal phalanx III	6.69	7.06	2.5
distal phalanx V	1.61	3.33	2.5
capitatum bone	14.29	10.76	7.1
hamate bone	14.13	10.75	7.1

Table 2-1 Weights of Different Bones On Skeleton Age Prediction

From Table 2-1, notice that radius, proximal phalanx III, distal phalanx I, capitatum bone and hamate bone are relatively decisive in skeleton age prediction and such bones as lunare bone are of less importance. So, although a sample image includes the whole hand bone, actually there are only a part of pixels contributing to the final result. Extracting this part of image, we can obtain sufficient information and get access to a really approximate prediction.

Thus, we conjecture that attention mechanism will have great effect in this project. Essentially, attention mechanism is to mark each feature of a set of interesting features with Softmax Function. For example, a group of interesting features may come from a picture, and then every pixel in the picture is scored. Another instance is that if the group of interesting features is a sentence, then every word in the sentence is scored. Normally, the input of the attention mechanism is the current state $\mathbf{h}_t$ and a set of features $\mathbf{f} = (\mathbf{f}^1, \mathbf{f}^2, \cdots, \mathbf{f}^n)$ and the output is the SoftMax score $\mathbf{s} = (\mathbf{s}^1, \mathbf{s}^2, \cdots, \mathbf{s}^n)$ for these n features, which can be used to filter the features or re-integrate the features into the system in subsequent processing.

With attention mechanism, in the skeletal age prediction task, higher computational resources are invested in the skeletal region with high recognition, while the use of computational resources is reduced in other background regions with sparse information, so that the model can obtain more detailed information. In this experiment, we build an attention mechanism to turn pixels in the GAP on an off before the pooling and then rescale the results based on the number of pixels. The model could be seen as a sort of 'global weighted average' pooling. It is very similar to the kind of attention models used in NLP. It is largely based on the insight that the winning solution annotated and trained a UNET model to segmenting the hand and transforming it. Afterwards, the transformed image, which is called attention map, will be used for subsequent processing.

B. Gender might be a potential influencing factor on bone age prediction.

According to Table 2-1, we contend that there exists diversity in the growth and development of skeleton between two genders since one kind of bone has different weights in male and female. Therefore, we modulate the original Neural network structure to import gender as a potential influencing factor. While the origin method didn’t take gender into account, with a mean absolute difference of 13.70 months, our model, which independently calculates the skeleton age prediction of both male and female, eventually has a mean absolute difference of 9.7845 months, that is approximately 4 months more accurate than before.

C. Other exploit we have made in this experiment.

We wonder whether and how the size of convolution kernel in attention layer will influence the accuracy of the final prediction. By replacing the original kernel of 1 * 1 with one of 3 * 3 and 5 * 5, we discover that the prediction results become more precise as the size of the convolution kernel increases. But we still need a further discussion whether it’s the optimal choice.

Furthermore, we revised the contribution of gender information to the whole network. According to [4], the best alternative is that gender contributes 32 input. In the experiment, we change the number and observe whether it's adaptable in our experiment.

Another try is changing the pre-training model with VGG19, ResNet and inception_v3, however, the results are far from feasibility. The result of these trials will be given in the part of Experiment & Discussion

III. Experiments & Discussion

A. Experiment Environment

OS: windows 10, macOS Mojave

Programming Environment: anaconda 5.3, pycharm 2018.2.4, python 3.6.6, cuda 9.2, cuDnn 7.1.4, tensorflow-gpu 1.11.0, keras 2.2.4, pandas 0.23.4, scikit-learn 0.19.2

(using computational resources at the Maryland Advanced Research Computing Center (MARCC))

B Neural Network Structure

The original network structure is shown below. (The numbers in brackets represent the dimensions of the input.)

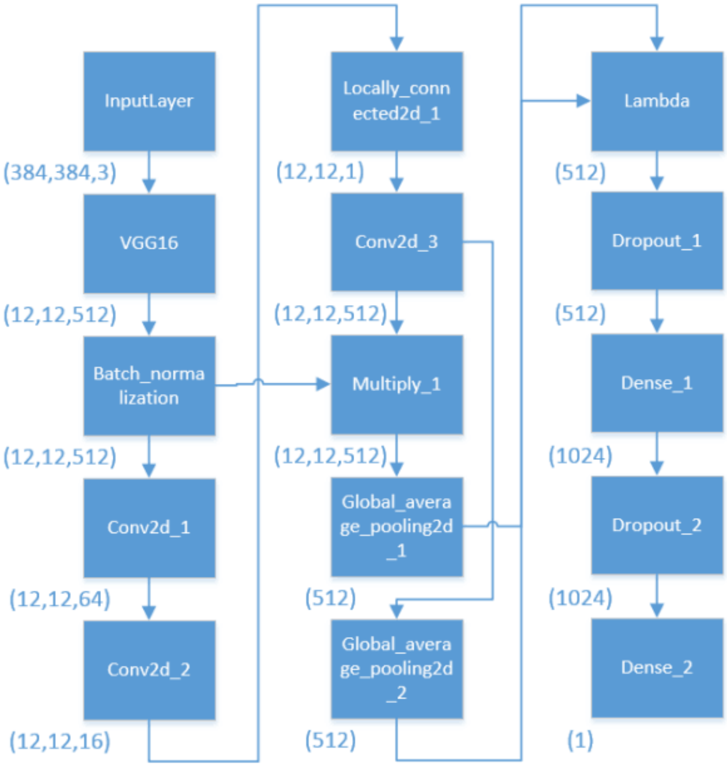


Figure 3-1 The Original Neural Network Structure

In this network structure, we use VGG16 as a pre-training model and the attention layer includes Conv2d_1, Conv2d_2, Locally_connected2d_1 and Conv2d_3. In this layer, a normal image with 384 * 384 pixels will be transformed into a data structure called attention map. Some examples of attention map are given in figure 4-2. And the next few layers are used to predict the bone age based on the image transformed by the attention layer. The samples of the final result are given in figure 4-3.

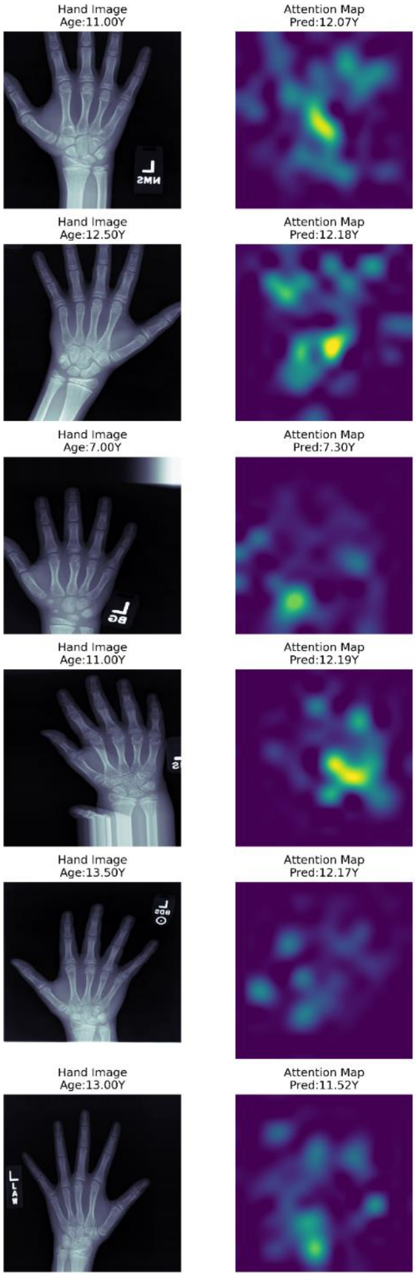


Figure3-2 Attention Map

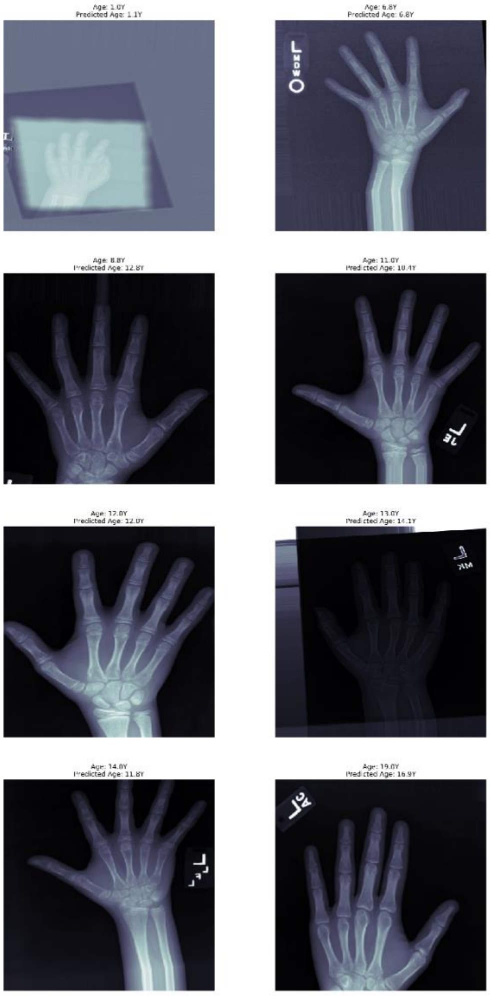


Figure3-3 Final Result

To import gender as an influencing factor, we modulate the original network structure and add several layers before the attention layer. The modulated network structure will be given in figure 4-4.

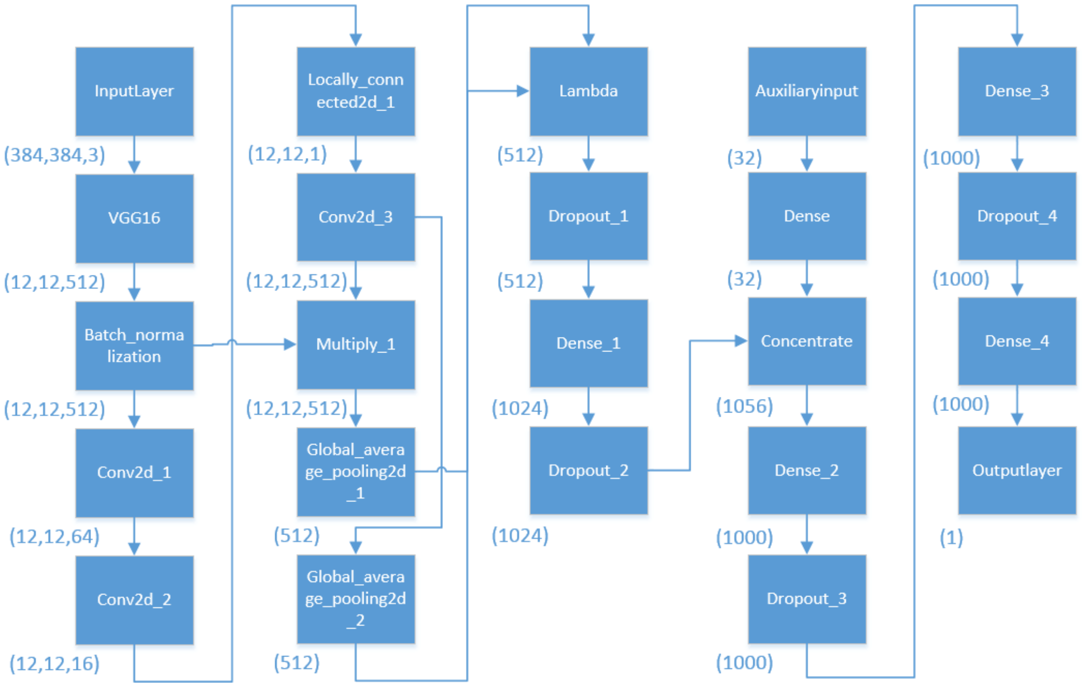


Figure 3-4 modulated network structure

Stress that Auxiliaryinput includes gender information. And by extracting a layer from the last cascade layer, flattened it and connected it to a gender network that was used to input binary gender information (0 for women and 1 for men), and fed it through a 64-neuron dense connection layer. Pixels contribute 1024 inputs, while gender contributes 64 inputs[4]. The reason for choosing this ratio is that instead of wanting the network to be too biased towards gender input, it wants to give it the ability to influence overall forecasts. For the need of research, actually we experiment on change this ratio and observe whether it's the optimal choice. Additional full connectivity layers give the network more learnable parameters to adjust during training so that it can infer the relationship between pixels and gender information.

C. Result & Discussion

(The MAD of each trial can be observed in Table 3-1)

Model	**MAD**
origin	13.6961
pre_train=vgg19	15.2857
pre_train=resnet	26.9282
pre_train=inception_v3	35.3809
kernel_size=(3, 3)	13.0673
kernel_size=(5, 5)	12.9808
female only	10.6902
male only	10.8182
no_balance	14.1151
shape_of_dense_2 = 16	11.2869
shape_of_dense_2 = 32	10.969
shape_of_dense_2 = 64	9.7845
shape_of_dense_2 = 128	10.2238
shape_of_dense_2 = 256	10.8532

Table 3-1 MAD of various models

After the network structure has been fixed, we also implement several trials, including replacing the VGG16 model in pre-training model with VGG19 (Figure 3-6), Resnet (Figure 3-7) and Inception_v3 (Figure 3-8). According to Table 3-1, it seems that images pre-trained by VGG16 attain the best performance in bone age prediction, which is different from our previous guess that taking Inception_v3 as a pre-training model will achieve a better performance. We infer this outcome may result from an improper adjusting of super parameter such as learning rate or the bias of the training set that causes a better training while a worse test.

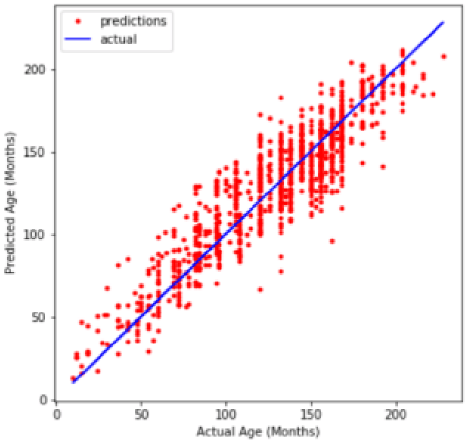


Figure 3-5 VGG16

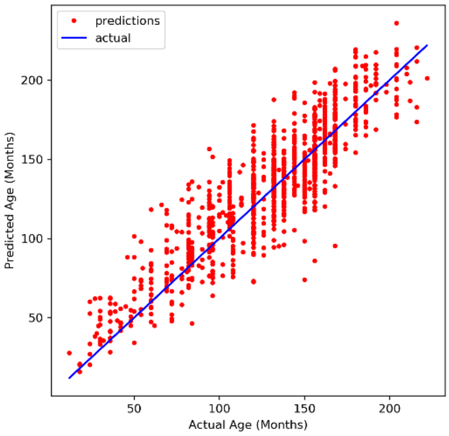


Figure 3-6 VGG19

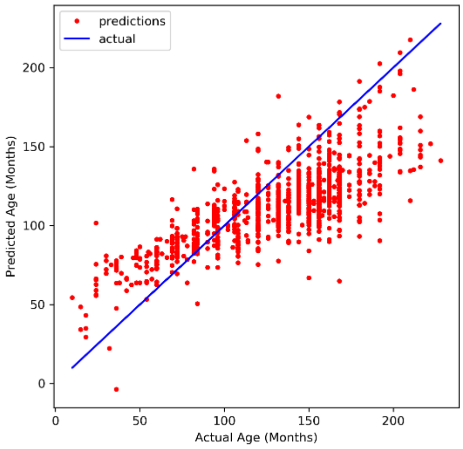


Figure 3-7 Resnet

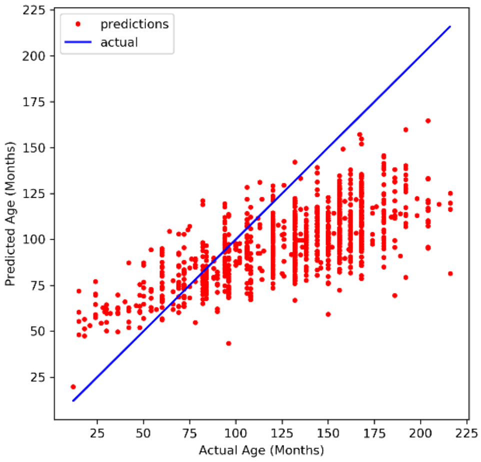
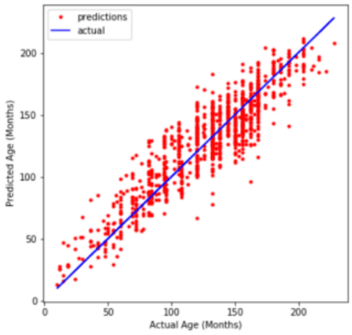
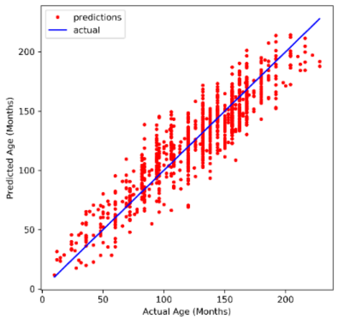


Figure 3-8 Inception_v3

On the choice of the convolution kernel size, a 5*5 kernel (Figure 3-11) perform the best compared with a 1*1 kernel (Figure 3-9) and 3*3 kernel (Figure 3-10). However, the attention map with a 5*5 kernel is inexplicable. We speculate that the attention layer with a 5*5 kernel brings about overfitting thanks to an increase of the amount of parameters while using 5*5 kernels. Due to the similar accuracy of models with a 5*5 kernel and a 3*3 kernel, we prefer to choose a 3*3 kernel in the attention layer.





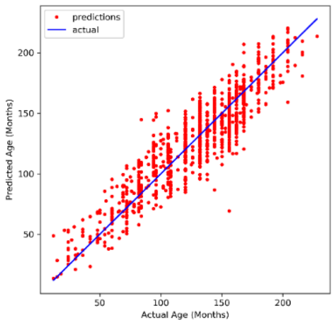
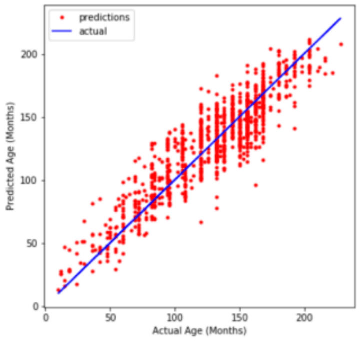
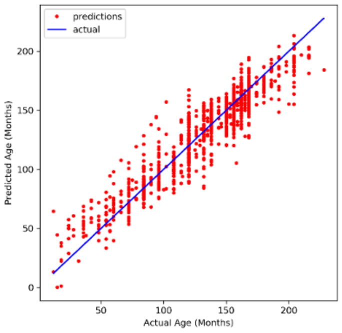


Figure 3-9 1*1 kernel Figure 3-10 3*3 kernel Figure 3-11 5*5 kernel

Another trial is to make prediction independently for both male and female, which has a great improvement of approximately 3 months in MAD. This experiment indicates that there indeed exists difference between male and female in skeletal development.





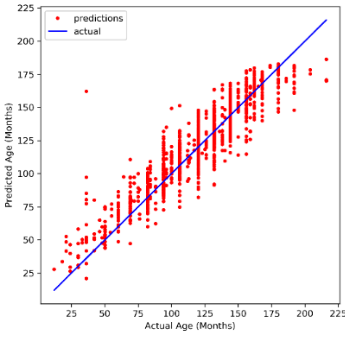


Figure 3-12 original Figure 3-13 male only Figure 3-14 female only

Before the pre-training model, we balanced our data first in bone age and gender information. If we don't do that, which is shown in Figure 3-16, the accuracy of the prediction will decrease. Interestingly, the rate of convergence slows down simultaneously.

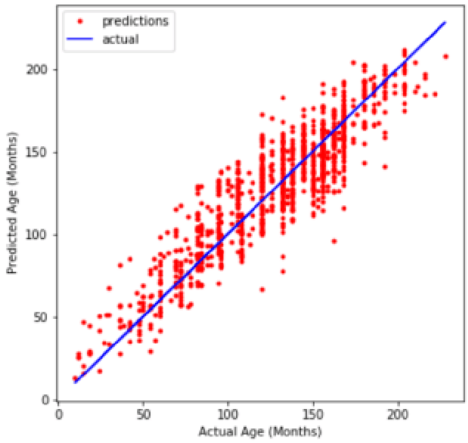


Figure 3-15 original

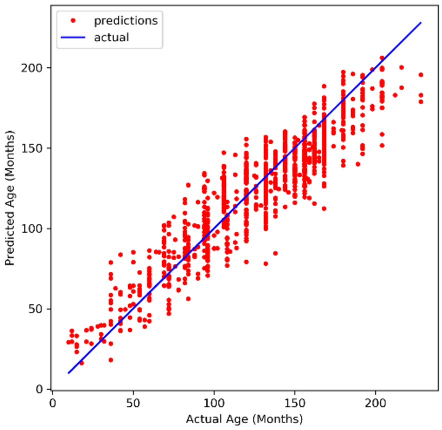


Figure 3-16 no data-balance

The final attempt is to change the contribution of gender information to the whole network. In the original model, gender contributes 32 inputs towards the concatenate layer. We try 16 (Figure 3-17), 32 (Figure 3-18), 64 (Figure 3-19), 128 (Figure 3-20), 256 (Figure 3-21) inputs respectively and empirically discover that 64 inputs from gender information obtain the best performance. As mentioned earlier, the ratio of pixel contribution to gender contribution depends on several aspects such as the medical information gain from image and gender, the network structure etc. Equally, we assert that it's empirical but theoretical.

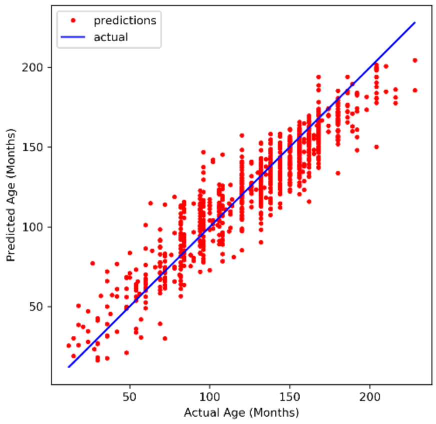


Figure 3-17 16 inputs

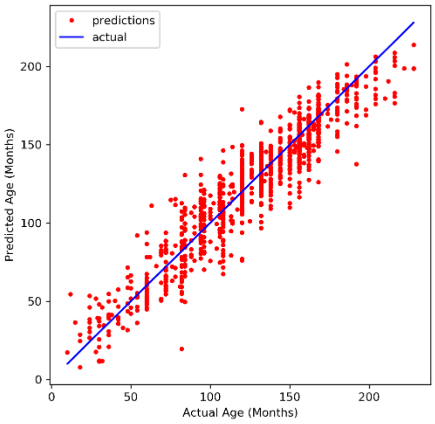


Figure 3-18 32 inputs

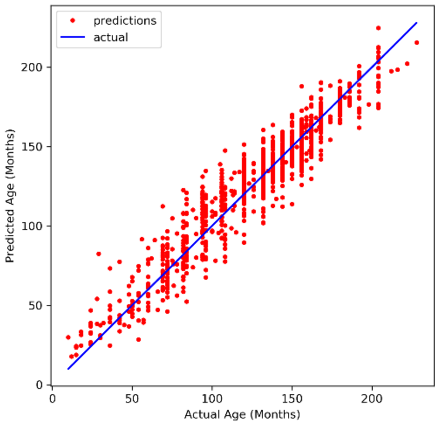


Figure 3-19 64 inputs

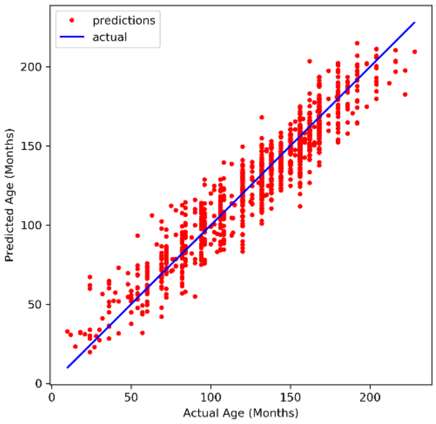


Figure 3-20 128 inputs

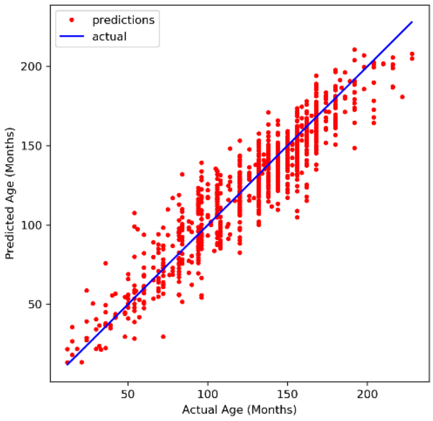


Figure 3-21 256 inputs

IV. Conclusion

The attention mechanism is effective in the task of skeleton age prediction. Implementing the attention mechanism, we can find the most decisive pixels in an image, thus increasing the accuracy of the following prediction.

VGG16 performed best in several pre-training models that had been tried. When choosing a pre-training model, we need to be very careful. If the task is very different from the pre-training model training scenario, the prediction results obtained by the model will be very inaccurate. We guess the reason why other pre-training models are not performing well as following: It may be related to the number of parameters and the depth of the model. Perhaps the number of parameters is large, and the shallower pre-training model is more suitable for this task.

Training the male and female hand X-rays separately can make the results obviously better, which shows that there is a difference in the bone structure of the hands of men and women. Thus, after adding gender information into the network structure, we achieve the best performance of 9.7845 months in MAD.

When the convolution kernel is changed from 1*1 to 3*3 and 5*5, the information in the picture is more fully utilized because the parameters in the operation process are increased. Thus, the prediction result is improved to some extent. In this task, since the results of 3*3 and 5*5 kernel calculations are not significantly different, but the 5*5 kernel calculations are large, it is better to consider using 3*3 kernels.

V. Future Work

A. Grayscale graph should be processed by one channel.

Due to the fact that most medical images are grayscale graph while the standard network architecture usually accepts an image of RGB format. In that case, we have to transform the dataset, which, to some extent, reduces the accuracy of prediction. If we could modulate the network to accept a grayscale graph, this model will have broader application scenarios.

B. If we have pre-training weights for medical images, the pre-training model may perform better.

The pre-training weights in this experiment is based on ImageNet, which is of more universality but lack of speciality in iatrology.

C. The uneven distribution of bone age may have potential effect on prediction.

Though we balance the sample amount of gender, the distribution of bone age is uneven. We postulate that this will have a negative influence on prediction.

D. The scanned image is noisy and can be converted to original image.

The images in the dataset are scanned, thus contains more noisy than the original ones. It, obviously, will affect the accuracy of prediction. If the original images are available,, the performance may be better.

Reference

[1] Zhang Shaoyan, Liu Lijuan, et al.Standard of wrist bone development in Chinese[J] Chin J Sports Med , Sept 2006 , Vol.25, No.

[2] 孙亚圣,姜奇.基于注意力机制的行人轨迹预测生成模型[J] Journal of Computer Applications, 2018-09-18

[3] 王景樟.基于深度学习的骨龄自动评测系统的研究与实现[J]

[4] http://www.itsiwei.com/21761.html

[5] https://www.kaggle.com/kmader/attention-on-pretrained-vgg16-for-bone-age/notebook?scriptVersionId=3164183