Bioinformatics, CAP 6545
doi: 10.1093/bioinformatics/xxxxx
Due Date: 29 Apr 2021
CNN Category

CAP6545 Final Project Report

The Prediction of Bone Age by Using Deep Convolutional Neural Network

Zhengtai Zhong

Abstract

Motivation: Bone age assessment is a common clinical method to diagnose endocrine and metabolic diseases during the growth of children. In this report, beside the baseline model which is proposed in paper [1], there are also two deep learning method for evaluating bone age. The dataset comes from a bone age assessment competition of RSNA in 2017. To find the better prediction result, there are two types of task for the comparison, one type of task is image segmentation, but since it is almost impossible to label the segmentation of each image individually, the result of paper [1] is used as baseline. Another type of task is classification in two customized models, and the backbone of each model contains Xception [2] or Inception V3 [3].

Results: The result shows that the model with Xception backbone has the lowest mean absolute error (MAE) and the convergence is faster.

Contact: Zhengtai.zhong@knights.ucf.edu

Supplementary information: Supplementary information listed in ReadMe.txt.

1 Introduction

In normal circumstances, bone age should be less than 10% of actual age, but there are exceptions. Although some exceptions are harmless, it is better to be aware, for example, if the bone age is older than actual age, then children will easily stop growing up. If the bone age is younger than actual age, then the child's growth period will be postponed. Other exceptions are much more serious, the difference in bone age and actual age would be regarded as problems, which contains growth disorders, metabolic disorders, and endocrine problems, such as the skeletal dysplasia, or any other unknown factor which coursed by nutritional or metabolic deficiencies. In the case of growth retardation, the bones and height would be affected, but the possibility of reaching normal adult height still exists through treatment.

For decades, bone maturity was usually determined by visual assessment of the shape of hand and wrist bone. The traditional method to estimate the bone age is by (Greulich and Pyle) GP or (Tanner Whitehouse) TW2, both methods are based on bone maturity, but these methods are redundant and objective, and even senior radiologists are not satisfied with the result. Therefore, it is attractive to improve the accuracy of bone age assessment by computer-aided diagnosis system, which is deep convolutional neural network..

2 Related Works

2.1 Preprocessing

The first step is to extract the region of interest, which is a hand mask, from the image and remove all irrelevant pixel. Since diversity of sample images, for instance, the images are taking from different machine and different institution, there is not any background deleting method that can approach the requirement once and for all. Therefore, it is necessary to have a reliable segmentation method manually. This paper [1] proposed a forward mining method, which is combined with manual labeling with automatic labeling. Moreover, the binary graph is segmented, and the connected regions are analyzed according to the post-processed segmentation results. For image segmentation, the author uses U-Net deep network structure. U-net can learn from a small training dataset, which is complement with the forward mining method. In order to locate the region, the upsampling features in the expansive path, and the high-resolution features in the contracting path are connected by skip-connections.

Generalized loss function (L):

$$L = H - \log J$$

Binary cross entropy (H):

$$H = -\frac{1}{n}\sum_{i=1}^{n}(y_i \log \hat{y}_i + (1 - y_i)\log(1 - \hat{y}_i))$$

Differentiable generalization of the Jaccard Index (J)

$$J = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i\hat{y}_i}{y_i + \hat{y}_i - y_i\hat{y}_i}\right)$$

The training dataset in U-net model is 100 masks which is labeled by online service Supervisly. After training, the model is used to segment hands on the remaining training sets. Since the predicted masks are inconsistent with the ground truth by visual inspection, then the author only keeps those acceptable high-quality masks and deletes the rest. Using these high-quality masks to expand the initial training set to increase the number of labeled images which used for segmentation and improving segmentation result. The complete iterative process is shown as below:

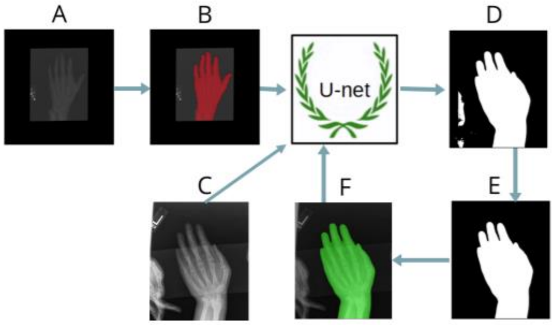


Figure1. Iterative Process

(A) originally input data;
(B) masks label manually with Supervisely;
(C) new input data;
(D) original prediction;
(E) post-processed prediction;
(F) masked raw data for visual inspection.

2.2 Key point detection

Since the purpose of automatic bone age assessment is to verify the importance of specific region of hands. In order to crop this region, all image should be aligned in the same coordinate space. Therefore, detecting the coordinates of each specific key point of the hand, then mapping the parament by scaling, rotation, shift, or mirror image.

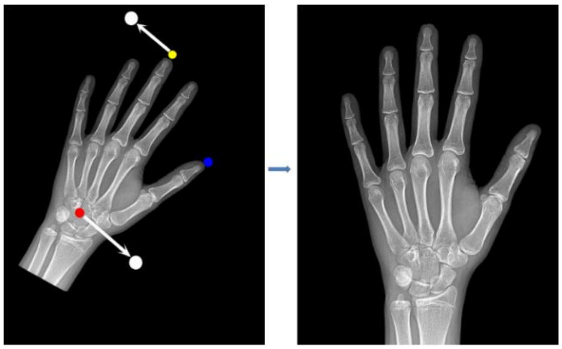


Figure2. Key Point Mapping

(Left) Key points: middle fingertip (yellow dot), the center of capitate bone (red dot), thumb tip (blue dot). Registration position: middle fingertip, center of capitate bone (white point).
(Right) Find the key points, and the affine transformation and scaling is applied.

There are three feature points which are selected in the image: middle fingertip, the center of capitate bone and the thumb tip. All images are scaled with the same resolution: 2080x1600 pixels, and zero padding is required if necessary. In order to build a training set of key point models, the author manually marked 800 images. Key point model is a regression model, and the pixel coordinates of key points are taken as the training target. This regression model is inspired by VGG block, and two drop out layers can produce a better generalized result.

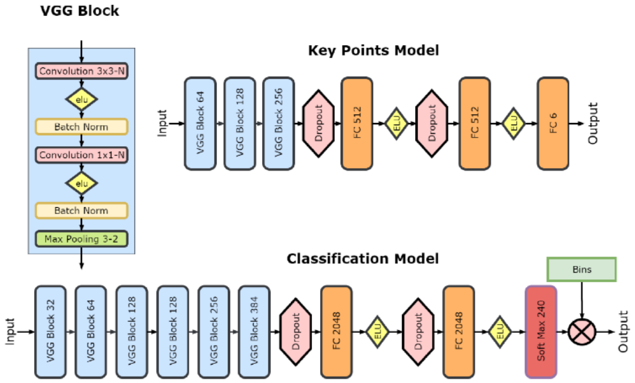


Figure3. VGG style model

The regression model is optimized by Adam and loss function is MSE.

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

In order to reduce computational overhead, all the input images are resized to 130x100, and target coordinate shrink to [-1,1] x [-1,1]. After the key points are detected, its coordinate will be mapping back to the original image size.

2.3 Bone age assessment

The prediction of bone age is a regression task, since the output is a consecutive month. Thus, this proposed VGG style model is trained by minimizing mean absolute error (MAE) through Adam optimizer.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|\hat{y}_i - y_i|$$

3 Method – Inception family

I have tried to reimplement the paper [1] by using U-net and key point model to obtain the bone age assessment result, but the problem is that the workload of labeling the segmentation in each image manually is extremely heavy, since I could not tell the acceptable quality of mask (which should contain three part: finger bone, metacarpal bone and carpal bone) by vision inspection in ten-thousand training dataset. Thus, I was looking for the most recently published model which can be used as the backbone of my customized model. Xception [3] is a suitable model, which is the final version of Inception family with the combination of residual connects. As the comparison, Inception V3 [2] backbone is the replacement by Xception, since the Xception is upgraded from Inception V3.

3.1 Inception family: from Inception module to Separable Convolution

Figure 4 is a canonical Inception module [3]. The basic concept is that feature can learn the relationship between feature channels, and this feature also can learn the relationship between the internal space in a single channel, when this feature is processed by convolution. Therefore, many 1x1 convolution kernels are used in inception module to focus on learning the relation between channels, then use 3x3/5x5 (or two 3x3) kernel to learn the spatial association in a single channel in different dimensions.

If only use 3x3 convolution kernel to represent the spatial association in a single channel based on the assumption of association separation used in figure 4, then we can get the simplified inception module, as shown in figure 5.

The simplified module can be expressed in the form shown in figure 6, because it is equivalent to using a 1x1 convolution kernel to learn the relationship between the features on the input feature maps, and then dividing the feature maps output by 1x1 convolution kernel, and submitting them to the following 3x3 convolution kernel to deal with the correlation of spatial elements.

Use a corresponding 3x3 convolution kernel to process the spatial association on each channel separately. In this way, the separable convolution shown in figure 7 is obtained.

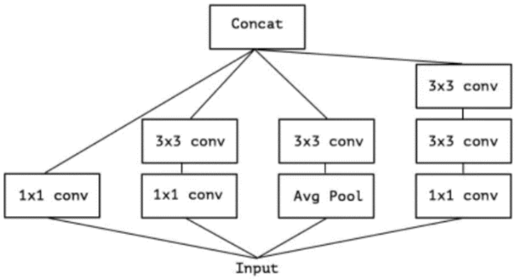


Figure 4. A canonical Inception module

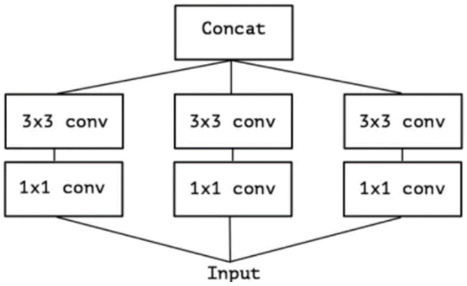


Figure 5. Simplified ver.

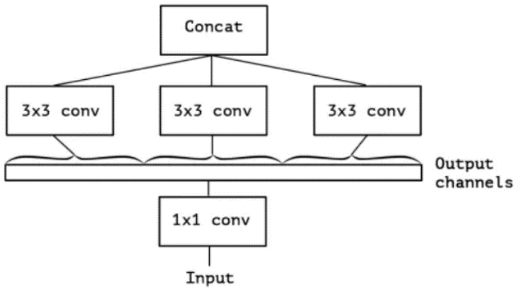


Figure 6. Equivalent ver.

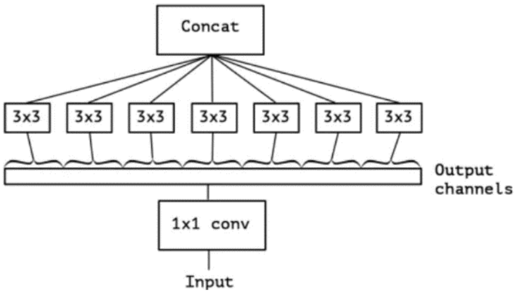


Figure 7. Extreme ver.

3.2 Inception family: from V1 to V4 (X)

The first version of Inception is GoogLeNet, which is a 22-layer network that won the ILSVRC 2014 competition. A year later, the researchers developed Inception v2 and v3 in the second paper, and made many improvements on the original version, the most notable one is to reconstruct the larger convolution kernel into a continuous smaller convolution kernel, which makes learning easier. For example, in Inception v3, the 5×5 convolution is replaced by two consecutive 3×3 convolutions.

The following figure is Xception architecture. It upgraded directly from Inception v3. The structure of residual learning is embedded to this structure. Same as the complex Inception family, it also has three flows: Entry/Middle/Exit, each of flow uses different repeating modules.

Entry flow is mainly used to continuously down sample and reduce the spatial dimension; The Middle flow is essential, which continuously analyzes and filters features, and it also constantly learn the relationship and optimize the characteristics; In the end, it is to summarize and sort out the features for presentation by fully connected layer.

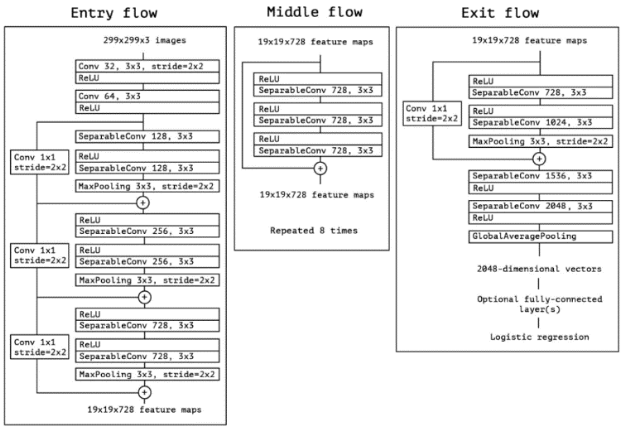


Figure 8. Xception Architecture

4 Experiments

4.1 Inception family: from Inception module to Separable Convolution

The dataset is from a bone age assessment competition of RSNA in 2017, and there are 6,833 boy’s radiographs and 5,578 girl’s radiographs, and the sample distribution is shown as the figure. The horizontal axis is bone age in month, and the vertical axis is the number of samples. After standardization, the bone age data of samples is converted into a z score from -3 to 3.

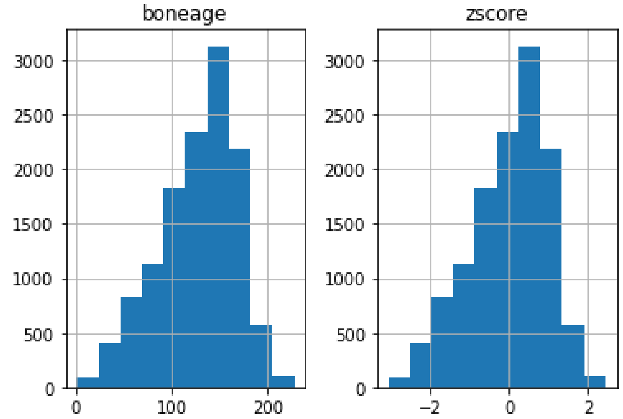


Figure 9. Dataset Distribution

After combined the two datasets, and drop the nonvalue, the data distribution as shown on the left, which need to be balanced. Since the image file name is corresponding to CVS file, the regular data augmentation methods, such as rotation, flipping and random cropping is inefficient. In this case, the final dataset is randomly select 500 images from 20 categories, so that ten-thousand samples were used, and 80% of them is training image.

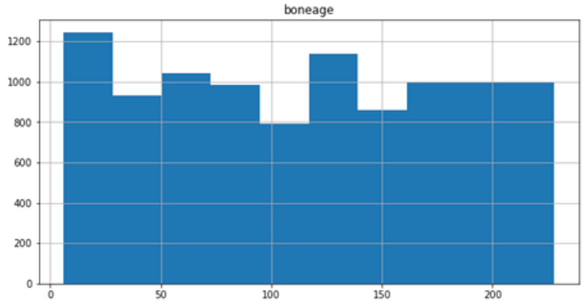


Figure 10. Balanced Dataset

4.2 Two Customized model with different backbone: Inception V3 & Xception

One thing should be noted that the default input image size for Inception V3 is 224x224, while the input image size for Xception is 299x299.

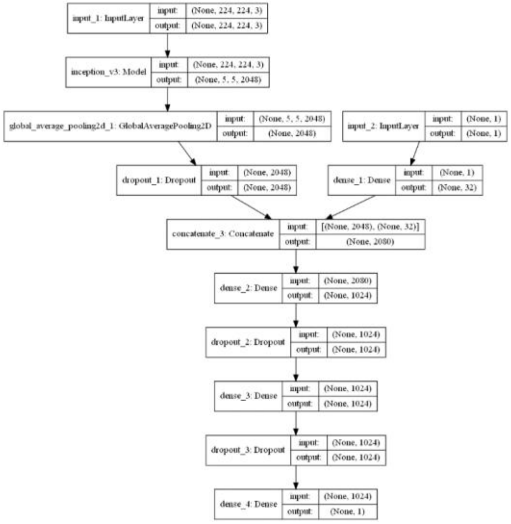


Figure 11. Customized model with Inception V3 backbone

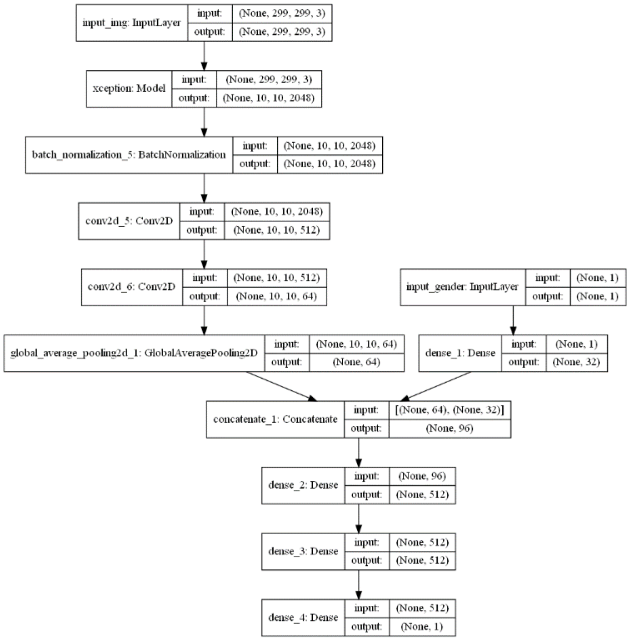

input_img: InputLayer
input: (None, 299, 299, 3)
output: (None, 299, 299, 3)
xception: Model
input: (None, 299, 299, 3)
output: (None, 10, 10, 2048)
batch_normalization_5: BatchNormalization
input: (None, 10, 10, 2048)
output: (None, 10, 10, 2048)
conv2d_5: Conv2D
input: (None, 10, 10, 2048)
output: (None, 10, 10, 512)
conv2d_6: Conv2D
input: (None, 10, 10, 512)
output: (None, 10, 10, 64)
input_gender: InputLayer
input: (None, 1)
output: (None, 1)
global_average_pooling2d_1: GlobalAveragePooling2D
input: (None, 10, 10, 64)
output: (None, 64)
dense_1: Dense
input: (None, 1)
output: (None, 32)
concatenate_1: Concatenate
input: [(None, 64), (None, 32)]
output: (None, 96)
dense_2: Dense
input: (None, 96)
output: (None, 512)
dense_3: Dense
input: (None, 512)
output: (None, 512)
dense_4: Dense
input: (None, 512)
output: (None, 1)

Figure 12. Customized model with Xception backbone

4.3 Result and Comparison

Both customized models tend be convergence around 30 epochs.

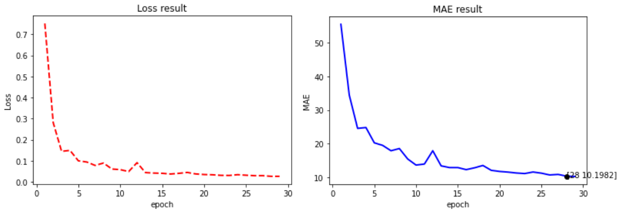


Figure 13. Loss & MAE (Inception V3 backbone)

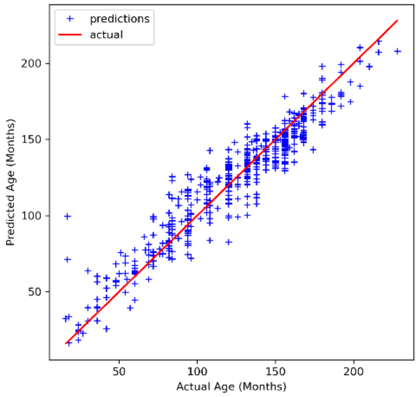


Figure 14. Actual Age vs. Prediction Age (Inception V3 backbone)

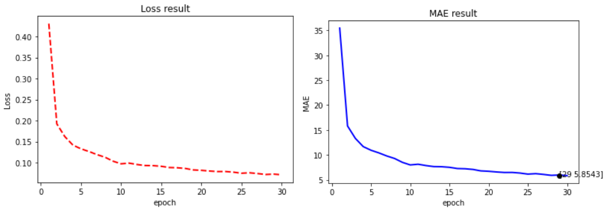


Figure 15. Loss & MAE (Xception backbone)

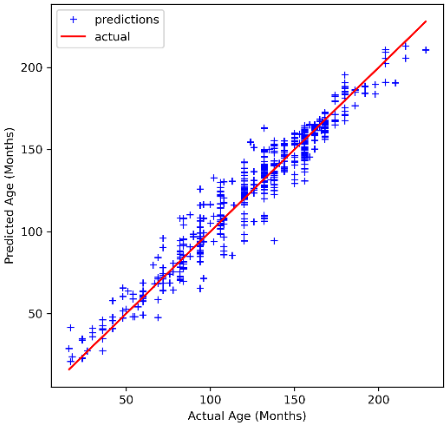
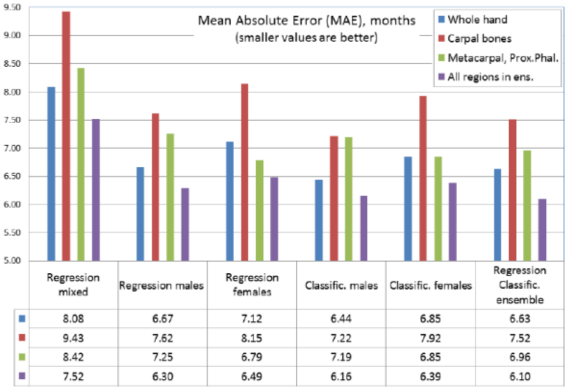


Figure 16. Actual Age vs. Prediction Age (Xception backbone)



	Regression mixed	Regression males	Regression females	Classific. males	Classific. females	Regression Classific. ensemble
🟦	8.08	6.67	7.12	6.44	6.85	6.63
🟥	9.43	7.62	8.15	7.22	7.92	7.52
🟩	8.42	7.25	6.79	7.19	6.85	6.96
🟪	7.52	6.30	6.49	6.16	6.39	6.10

Figure 15. Baseline MAE [1]

	Baseline	V3 backbone	X backbone
MAE (months)	6.1	10.1982	5.8543

Table 1. MAE results

Since the key point model [1] has many different combinations result, in this case, the lowest MAE, which produced by regression classification ensemble with whole hand region image, is used for the comparison. Xception backbone model has the lowest MAE value, but V3 backbone model has the highest MAE value. Probably because the separable convolution kernel in Xception has a better learning ability than the regular Inception module, and the residual connects in Xception architecture also accelerate the convergence with a higher accuracy.

5 Conclusion and Discussion

In this project, I have successfully decreased the mean absolute error for the bone age assessment, and the customized model is more accurate than the proposed method [1]. Since the difference is 0.35 month, which approximate 10 days. Although this project is using CNN models, which is a classification task, in fact, it can be regarded as a regression task. Moreover, in term of MAE, Xception network is almost 2 times powerful than Inception V3, not only because of separable convolutions, but also the deployment of residual model.

References

[1] Vladimir Iglovikov, Alexander Rakhlin, Alexandr Kalinin, Alexey Shvets. “Pediatric Bone Age Assessment Using Deep Convolutional Neural Networks”. arXiv:1712.05053 [cs.CV]

[2] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, Zbigniew Wojna. “Rethinking the Inception Architecture for Computer Vision”. arXiv:1512.00567 [cs.CV]

[3] François Chollet. “Xception: Deep Learning with Depthwise Separable Convolutions”. arXiv:1610.02357v3[cs.CV]