Bone Age Prediction: Detect Bone Growth using VGG19 with Attention Mechanism

Nguyen Duc Bao Minh, Nguyen Dinh Tung, Nguyen Kien Trung,
Le Quoc Trung, Nguyen Huy Tung, Do Bao Phuc

March 30, 2025

Abstract

Bone age assessment using X-ray images is a standard clinical procedure to detect any anomaly in bone growth in human. Usually, the manual screenings is assessed through X-ray images of the non-dominant hand using the Greulich-Pyle (GP) or Tanner-Whitehouse (TW) approach. This is our group's report about the application for bone age prediction automatically using VGG19 transfer learning enhanced with attention memchanism. The proposed system would takes input as 384x384 (RGB 3-channels) and then the output would be prediction of the bone age through the VGG19 network that incorporates multiple layers of spatial attention mechanism to emphasize the important features for more accurate boneage prediction. From the experimental results, the proposed VGG19 achives the lowest mean absolute error and mean squared error on the test set of 16.2751 months and 20.6213 months^2.

I. Introduction

Bone age refers to the actual level of bone growth development, which can be assessed continuously as the skeleton bone grows, as such, it will change in shape and size over time.

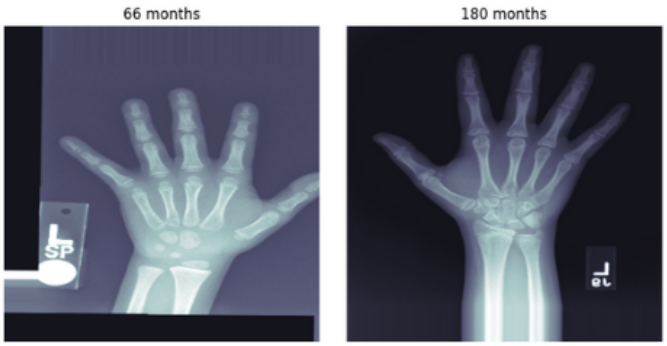


Figure 1: Comparison of hand X-rays between 5-year-old and 15-year-old subjects showing bone development differences

As shown in Figure 1, there are significant differences in bone structure between different age groups. The image clearly demonstrates there are significant changes as a toddler grows older until they reach the maturity age of 18 years. The assessment is done using a non-dominant hand due to the nature of bone ossification. It is vital to predict a child's age through bone X-ray images to understand if there are any health issues (such as: genetic disorders, hormonal problems, and endocrine disorders).

II. Materials and Methods

2.1 Workflow overview

The proposed automated bone age assessment system will involve two stages, which are Data Preprocessing Stage and Prediction Stage. Figure 2 shows the full workflow of the system, starting from the X-ray input until the predicted output of the bone age assessment. The first 4 modules perform a data preprocessing task that transforms the input image into a standardized form so that the VGG19 model can better predict the bone age wth high accuracy.

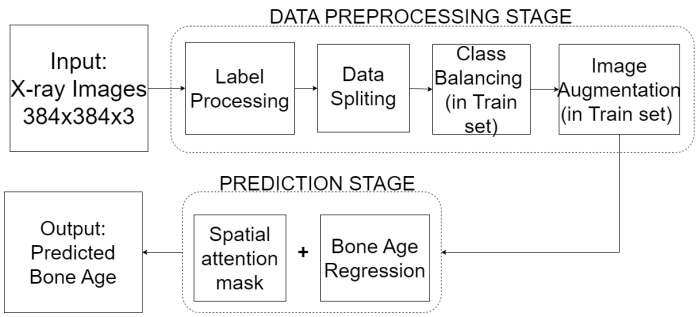


Figure 2: The workflow of the proposed automated assessment of the bone age

2.2 Input

Dataset Firstly, the dataset that we used in the project is the RNSA Bone Age from the 2017 RSNA Pediatric Bone Age Challenge. It is developed by Stanford University and the University of Colorado and was annotated by multiple expert observers. The RSNA Bone Age dataset containing 12.611 images. For each image, the ground truth bone age and the sex are provided.

2.3 Data Preprocessing Stage

2.3.1 Label Processing

To understand the dataset further, we label each image with its corresponding gender, then we use z-score to normalize the bone age. Figure 3 below shows three histograms after applying z-score. The boneage histogram's distribution is right-skewed, with most subjects clustered around 80 months to 180 months; there are fewer samples for very young and very old ages. The boneage zscore has symmetric around mean (z=0), shows proper normalization. And then male subjects appear to be more then female subjects.

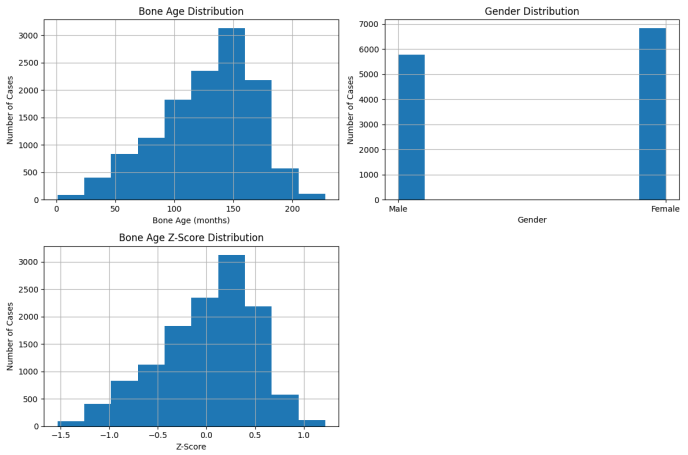


Figure 3: The histogram distribution of Bone age, Bone age z-score, and Male/Female

2.3.2 Data Spliting

The Bone Age dataset is then split into 2 subset, train set (9458 images) and validation set (3153 images). Then from the validation set, there will be 1024 image chosen to be test set.

2.3.3 Class Balancing

From Figure 3, the dataset has shown a slight imbalanced. A stratified sampling strategy was taken in the Training set only. We group the data by bone age category (10 bins) and genders (male/female). There are 500 samples rows per group, so the new training dataset results in 10 bone age categories $\times$ 2 genders $\times$ 500 samples $\equiv$ 10.000 total rows. Some groups had lesser than 500 samples, we allow duplicates existing rows to reach 500, while the others had more than 500 sample, we keep only 500 samples. The new bone age training dataset is shown in Figure 4:

2.3.4 Image Augmentation

Image augmentation for model training is crucial for improving model performance and generalization. For our project, only Training set is applied with augmentation function, while the validation and test set is intact. Below is all the augmentation details that we applied:

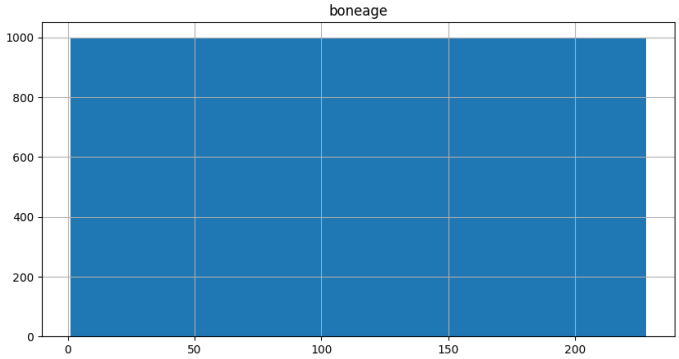


Figure 4: New Training dataset after Class Balancing

Table 1: Data augmentation pipeline configuration

Augmentation Processing Step	**Train**	**Valid**	**Test**
Input image size	384×384		
Color channels	3 (RGB)		
Horizontal flip	√	–	–
Vertical flip	–	–	–
Height shift (15%)	√	–	–
Width shift (15%)	√	–	–
Rotation (±5°)	√	–	–
Shear (1%)	√	–	–
Zoom (±25%)	√	–	–
Fill mode	Nearest	–	–
Batch size	32	256	1024

Here are some example from the train set after augmentation:

2.4 Prediction stage

In this work, a spatial-based attention mechanism is embedded in the bone age regressor network. The purpose of this attention network is to emphasize more weightage on the region of interests, such that age can be better differentiated. In our project, we transfer learning the VGG19 model from pre-trained ImageNet weights, we also apply batch normalization output features. For the attention mechanism, we have a locally connected 2D, which activate sigmoid function and dynamically hightlight bone-relevant regions. On the other hand, we also apply global average pooling (GAP). In the end, a regression head is used instead of default classification head of the VGG19 architecture. Only the attention and regression layers are trained from scratch.

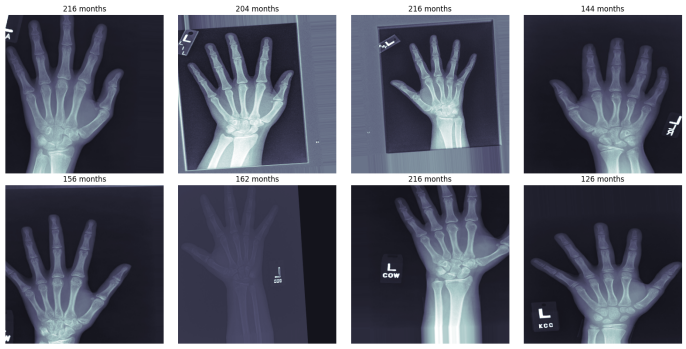


Figure 5: Augmentation example

Limitations When training, because of limited resources (time and computational power), the model is only trained in 5 epochs, which leads to it is not trained unitl convergence, the Mean Average Loss is still decreasing (and can still decreasing we assume). Also the model does not apply residual and feed-forward connections.

III. Results and Discussion

This section describes the evaluation metrics used for performance analysis of the test set:

3.1 Performance Metrics

Three evaluation metrics are used to validate the proposed method for bone age regression: mean absolute error (MAE), mean absolute deviation (MAD), and root mean squared error (RMSE). These metrics were computed using scikit-learn's implementations (`sklearn.metrics`), which employ the following standard formulas:

- **Mean Absolute Error (MAE)**:

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

  where y_i is the true bone age, $\hat{y}_i$ is the predicted bone age, and n is the number of samples.

- **Mean Absolute Deviation (MAD)**:

$$\text{MAD} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \text{median}(y)|$$

  This measures the average absolute difference from the median bone age, providing a robust measure of dispersion.

- **Root Mean Squared Error (RMSE)**:

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

The MAE measures the average of absolute differences between predicted and annotated bone age values, while MAD characterizes the variability in the ground truth data. The RMSE gives more weight to larger errors through squaring. All metrics were computed using:

- `mean_absolute_error()` for MAE
- Custom calculation for MAD (using numpy's `median()` and `mean()` functions)
- `mean_squared_error(squared=False)` for RMSE

3.2 Attention Mechanism Results

The attention value for random 6 images are very low, varies from the lowest: 2.21432e-5 to highest: 0.04. That is why we have to apply normalization for the attention map, adjust vmin and vmax to fit the range as below:

Overall, the attention mechanism successfully identifies relevant bone structures in older children.

3.3 Regression Results

The Figure 7 plot clearly demonstrates a strong positive correlation between the actual age (x-axis) and the predicted age (y-axis). As the actual age increases, the predicted age tends to increase as well. There are fewer data points on the lower age and older age. But the model perform equally well across the entire age range.

3.4 Evaluation

Key observations from these results:

- The **MAE of 16.2751 months** indicates the model prediction to the ground truth is usually 16 months
- **MAD of 13.5901 months** shows the inherent variability in the ground truth data
- **RMSE of 20.6213 months** suggests occasional larger errors due to its sensitivity to outliers
- The MAE/MAD ratio of 1.20 suggests the model's error is slightly higher than the dataset's natural variability

IV. Conclusion

In this project, we have learn to do overall workflow of a system processing medical image!

Table 2: Model architecture summary

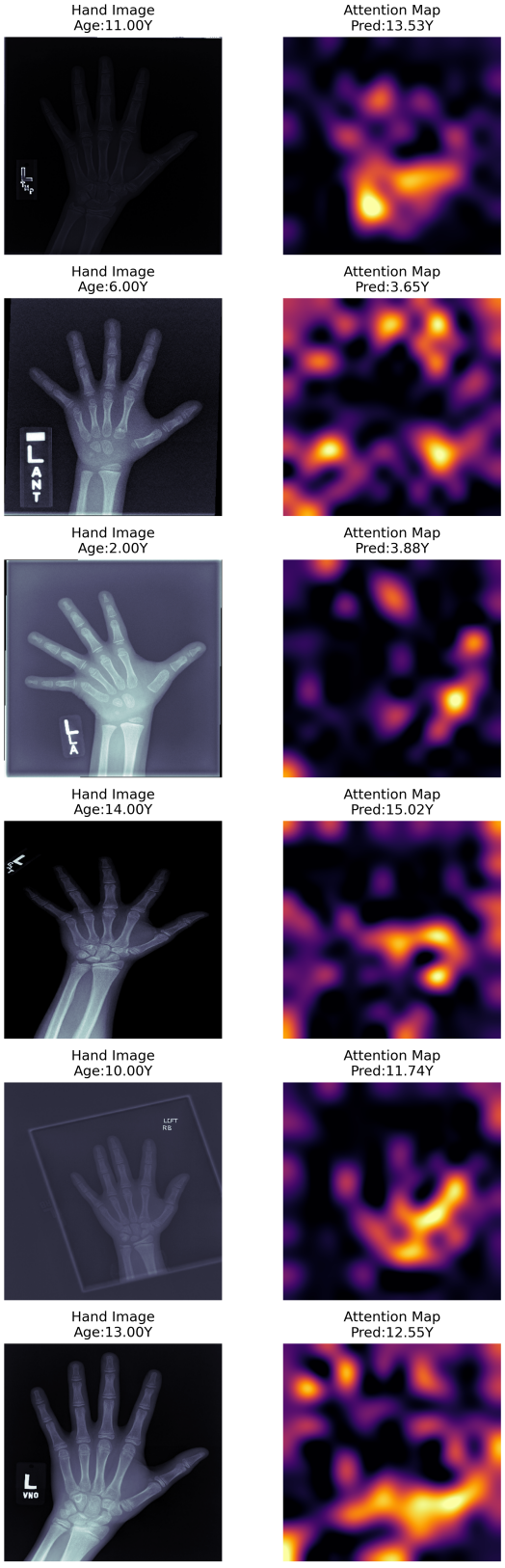
Layer (Type)	Output Shape	Param #	Connected to
input_2 (InputLayer)	(None, 384, 384, 3)	0	[]
vgg19 (Functional)	(None, 12, 12, 512)	20,024,384	input_2[0][0]
batch_normalization	(None, 12, 12, 512)	2,048	vgg19[0][0]
conv2d (Conv2D)	(None, 12, 12, 64)	32,832	batch_normalization[0][0]
conv2d_1 (Conv2D)	(None, 12, 12, 16)	1,040	conv2d[0][0]
locally_connected2d	(None, 12, 12, 1)	2,448	conv2d_1[0][0]
conv2d_2 (Conv2D)	(None, 12, 12, 512)	512	locally_connected2d[0][0]
multiply (Multiply)	(None, 12, 12, 512)	0	conv2d_2[0][0], batch_normalization[0][0]
global_average_pooling2d	(None, 512)	0	multiply[0][0]
global_average_pooling2d_1	(None, 512)	0	conv2d_2[0][0]
RescaleGAP (Lambda)	(None, 512)	0	global_average_pooling2d[0][0], global_average_pooling2d_1[0][0]
dropout (Dropout)	(None, 512)	0	RescaleGAP[0][0]
dense (Dense)	(None, 1024)	525,312	dropout[0][0]
dropout_1 (Dropout)	(None, 1024)	0	dense[0][0]
dense_1 (Dense)	(None, 1)	1,025	dropout_1[0][0]
Total params:			20,589,601
Trainable params:			563,681
Non-trainable params:			20,025,920

Table 3: Model performance on test set (n= 1024 in 32 batch chunk)

Metric	Value	Units
Mean Absolute Error (MAE)	16.2751	months
Mean Absolute Deviation (MAD)	13.5901	months
Root Mean Squared Error (RMSE)	20.6213	months

Table 4: Performance comparison with state-of-the-art methods

Method	**MAE (months)**	**MSE (months)**	**MAD (months)**	**RMSE (months)**
Proposed VGG19+Attention	16.28	–	13.59	20.62
AXNet	7.70	108.87	–	–
Deeplasia (SOTA)	–	–	3.87	7.67



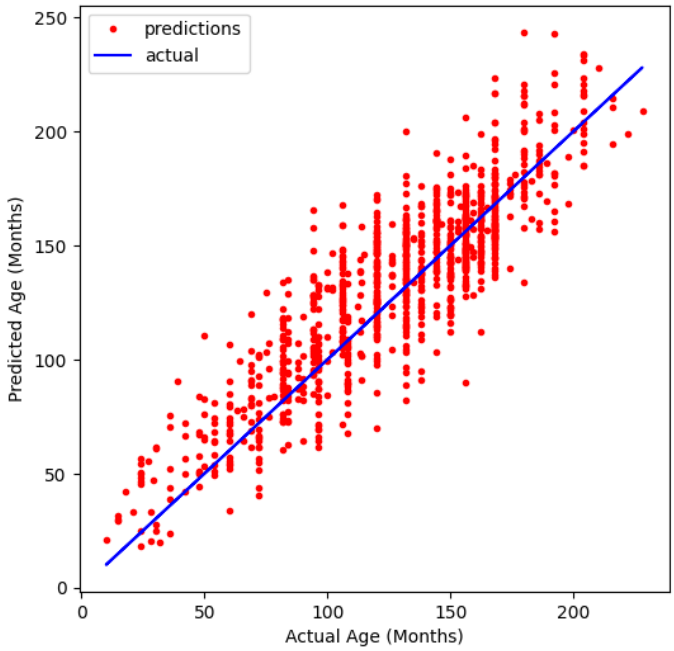


Figure 7: Regression plot

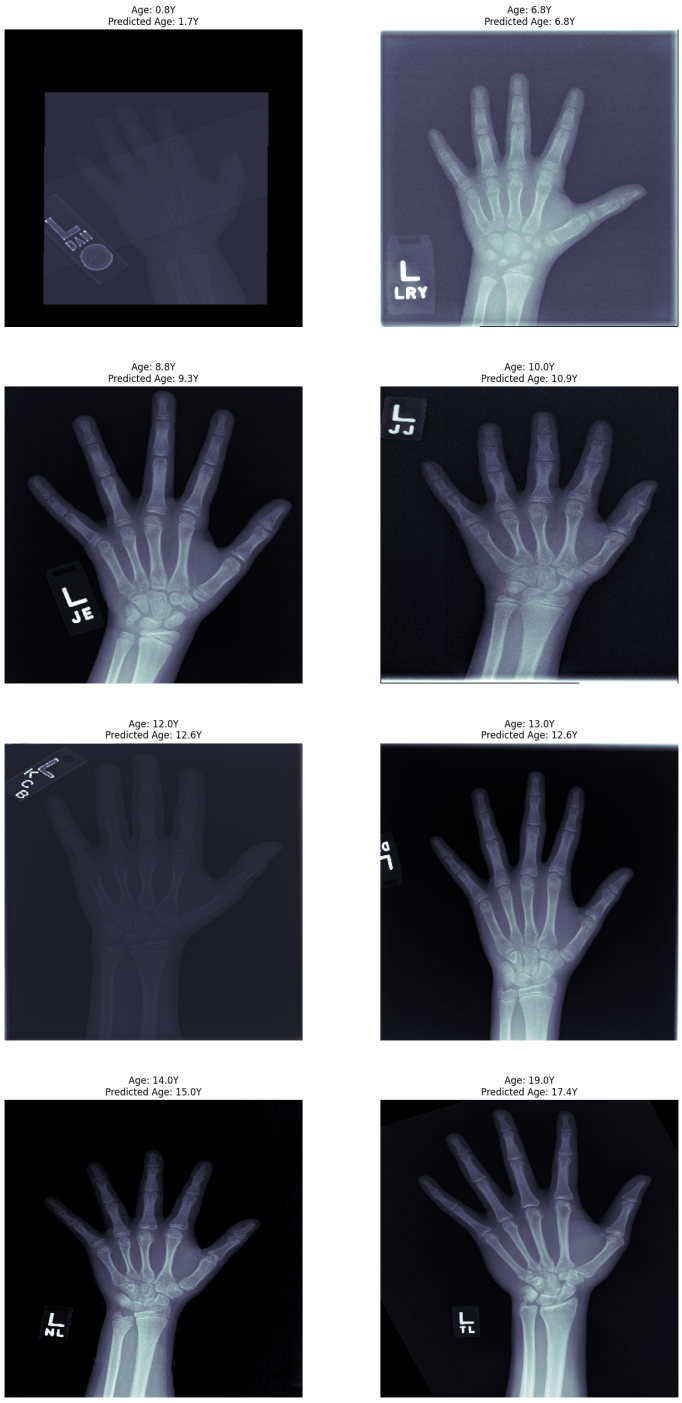


Figure 8: Prediction sample