Age prediction from Hand X-Rays: A Deep CNN Approach

Francesco Pittorino, Giulio Nebbiai

February 2024

Abstract - The analysis of skeletal age through X-ray hand images in paediatric patients is essential for assessing the skeletal development and evaluating growth and development. This study proposes to apply deep learning models to accomplish a regression task in order to predict the age of the patient based on X-ray hand image analysis. The images were studied with three different models: Inception v4, ResNet50 and SqueezeNet v1.1. The architectures' results were compared in terms of size, that is a proxy of complexity, and performances. The outcome was that ResNet50 achieved the best performance, with a MAE on test equal to 11.94 months.

1 Introduction

Inferring the skeletal maturity of paediatric patients is a fundamental task in clinical application, mainly to monitor their growth development. Traditional methods are based on radiologists' interpretation. This approach shows several limitations: it requires an expert and usually some time in order to obtain the predictions, and the different experts' judgments are influenced by subjectivity. In the last years, efforts were made in order to automatize the process and deep convolutional networks represent the state of the arts for this purpose. CNNs automatically extract pertinent features from the raw pixel values of pictures. As a result, manually designed feature extraction procedures are no longer necessary, enabling the network to learn the unique patterns in the input. In the study, three different convolutional architectures were trained and evaluated to accomplish the regression task and were compared in terms of error and complexity: Inception v4, ResNet50 and SqueezeNet v1.1. The dataset processed in the study is obtained from two hospitals of the United States and comprehends 14236 X-ray hand images. The use of hand images is related to its development which presents less variability between different subjects compared to other body parts. However, to produce reliable results, it was needed to apply some preprocessing to the images: resize all images to a standard size, since each original image had a different size, apply CLAHE, to enhance the contrast, normalization of the pixels values into range [0,1],

random cropping and reduction of input channels. The performances of the three models were compared in terms of Mean Absolute Error (MAE).

2 Related Work

Deep learning approaches have led to considerable gains in bone age prediction using X-ray images. As highlighted in the study by Chen et al.[1], X-ray images fail to adequately capture coaxial features (which are the patterns traditional methods are based on) but they become useful when fed to a deep learning model, which automatically extract distinguishing characteristics that aid in the precise assessment of bone age using this type of data. Different deep residual network-based approaches for skeletal bone age prediction are presented in the article by Halabi et al. [2]. The models that achieved the best accuracy are an Inception v3 incorporating even information about the gender of the patience, and a ResNet50 trained on subpatches obtained by dividing the original training images. The article moreover stress the importance of an adequate preprocessing pipeline, taking into account the unique features of medical imaging and in particular of course X-ray images. Togaçar et al [3] in their work apply a SqueezeNet architecture to detect Covid-19 with X-ray chest images. Since their result are good in accuracy and robust, and considering that the nature of input is the same, investigating the performance of such CNN architecture on a slightly different task could be of interest.

3 Dataset

The owner of the data used for the study is "Stanford Medicine's Box Instance" and the images set is freely available. It is organized into three subdirectories, namely the training, the validation and the test set. Moreover, each directory contains a dataframe CSV object that associates to each id of the image the gender and the target value of the regression that is the age in months of the patient. Training and validation data were collected from Children's Hospital Colorado (Aurora, Colo), and Lucile Packard Children's Hospital at Stanford, while the ones related to the test set were obtained to Lucile Packard Children's Hospital. The dataset is split into 12611 images for the training set, 1425 images for the validation set and 200 images for the test set. Here some statistics about the target are displayed.

Statistic	**Training**
Min (in months)	1
Max (in months)	228
Mean (in months)	127.3

For the regression purpose, the gender information wasn't taken into account.

4 Preprocessing

The preprocessing pipeline of the study is implemented in five main steps:

1. Resizing the images: images are resized to a standard shape (256x256), in order to make the model able to process them. The 256 pixels size was chosen trying to optimize the trade-off between two purposes: preserving

the needed data patterns for the regression and limiting needed computation.

2. Applying CLAHE: in this stage, the Contrast Limited Adaptive Histogram Equalization (CLAHE) technique is used to increase the images contrast. This method works by first dividing the image into small, discrete areas known as tiles and then separately performing histogram equalization to each of these areas. Through the "cliplimit" parameter is set the threshold for the contrast equalization. CLAHE is a powerful tool that, however, needs to be tuned to avoid raising noise levels in the background areas of the image. After some trials this goal was reached setting the clipLimit parameter to the value of 4.

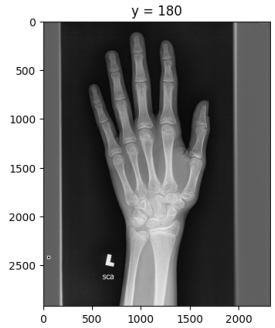

Figure 1: Random image of data before CLAHE

Figure 2: Random image of data after CLAHE

3. Normalization: the pixels values are standardized in the range [0,1] to speed up convergence.

4. Random Cropping: since the X-ray of the hand for each images is centered in the middle, random cropping is applied through a random zooming.

5. Reducing the number of channels from 3 to 1: this was a necessary step in order to limit the number of parameters of each DL model, reducing the ETA for epoch.

5 Learning Framework

Python, version 3.10.12, was used as the programming language for model creation and training. To speed up the training process, the GPUs provided by the environment were utilized, including GPU P100 and GPU T4 x2. The training was executed with a mini batch of 40 examples and each procedure was repeated for 70 epochs, in order to reach the convergence and to refine results. The learning rate was set to 0.001, because turned out to be a

good trade-off between avoiding overshooting and speeding up the training process. Moreover,"Adam" optimizer was used. It dynamically adjusts the learning rate for each parameter, resulting in a faster and more precise learning process. It was set also a callback to stop the algorithm during the training phase: early stopping was triggered when the calculated loss on the validation set does not improve for 16 subsequent epochs. In case of early stopping, the function returned the best weights calculated during the process. Furthermore, each model developed in this study utilized the Rectified Linear Unit (ReLU) activation function, which is defined as:

$$f(x) = \max(0, x)$$

To monitor the models training, were chosen two evaluation metrics: the Mean Square Error, as loss function, and the Mean Absolute Error, measure that was used to check the models' goodness in predicting the real target value.

6 Models

In this section, the implemented neural networks are presented.

- **SqueezeNet v1.1** is a convolutional neural network designed to obtain good performance on benchmark dataset for computer vision with significantly fewer parameters compared to traditional architectures. Limiting the number of parameters is achieved with different strategies. The first is using mainly 1x1 kernels since it has 9X less parameter than a 3x3 filter, 25X less parameter than a 5x5 filter and so on. Another technique used for this purpose is to substitute the lasts dense layers, which are common in traditional CNN architectures, with a max pooling layer. Finally, SqueezeNets rely on the use of the fire module which is made up by a squeeze layer (1x1 convolution) and an expand layer (1x1 and 3x3 convolutions). When using Fire modules is needed to set the number of filters of squeeze layer to be less than the sum of number of filters of 1x1 size and 3x3 in extend layer. This way the squeeze layer helps to limit the number of channels for the feature map used as input to the 3x3 filters in extend layer. This last layer of the module instead increase the depth of feature map to recognize fine patterns. Combining the previous three techniques, the total number of parameter drastically decrease. To the base architecture of SqueezeNet v1.1 is added a fully connected dense layer with a single neuron and linear activation to produce the output and the last Dropout Layer is set to a rate of 0.2, against the 0.5 of standard SqueezeNet v1.1.

- **ResNet50** is a convolutional neural network (CNN) architecture designed for advanced tasks in computer vision. What sets the ResNet50 apart is the innovative use of residual blocks, namely the convolutional block and the identity block, devised to overcome the vanishing gradient problem during the training

of deep neural networks. These blocks include shortcut connections that allow the flow of information to skip one or more layers, facilitating the flow of information through the network. On the other hand, identity blocks are used to maintain the dimensions of the input unchanged. These blocks don't include convolutions with a stride other than 1. An average pooling layer and a flatten layer were added prior to the final layer in order to further reduce the feature map's size and convert the 2D feature maps into a 1D vector. The output is then added with a fully linked dense layer with a single neuron and linear activation.

- **Inception v4** is a convolutional neural network (CNN) specifically designed for advanced image processing tasks. Inception-v4's architecture rely on the use of Inception modules, which are specialized neural network blocks aimed at capturing image features across various scales and resolutions to enhance the model's performance and efficiency. The structure of Inception blocks is characterized by multiple parallel branches, each performing a combination of convolution operations with filters of different sizes, pooling, and other transformations. The features extracted from each branch are then concatenated to form the final output of the Inception block. Additionally, Inception modules incorporate pooling layers and reduction modules, which help reduce dimensionality, thus managing computational complexity and boosting model efficiency. To accomplish the regression task, before the last layer were applied an average pooling layer, to further reduce the size of the feature map, a flatten layer and a dropout layer with a dropout rate of 20% to reduce the number of model parameters. Finally, a fully connected dense layer with a single neuron and linear activation is added to produce the output.

7 Results

In the following table the main characteristics of each model are displayed with the evaluation on test set samples.

Model	MSE	MAE	s/step	Model Weight
SqueezeNet v1.1	270.83	13.11	1s/step	5.14 MB
ResNet50	219.53	11.94	1s/step	90.08 MB
Inception v4	228.37	12.31	1s/step	157.31 MB

In the following graphs the progress of MSE and MAE during the learning phase of each model is plotted.

MSE and MAE on Test set

SqueezeNet v1.1 ResNet50 Inception v4

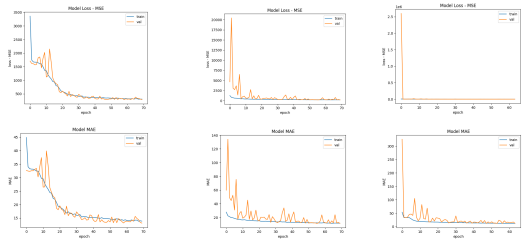

Figure 3: MSE and MAE of the models

8 Conclusions

In the study the application of deep learning models for bone age predic-

tion using X-ray images of the hand is explored taking into account three different architectures: SqueezeNet v1.1, ResNet50, and Inception v4. The results show how all represent an efficient tool for the task, with a MAE value on test of about 1 year. The best performing model was the ResNet50, with a 11.94 MAE. Given ResNet size (90.08 MB), Inception v4 wouldn't be a choice for application since the accuracy is not better than ResNet50, but the size is 70% bigger (157.31 MB). It's important instead to highlight how the SqueezeNet has a MAE higher of 1 month than the ResNet, but having 5.5% of the size (5.14 MB). Further investigation should be done in assessing the performance of SqueezeNet architectures more complex than the 1.1 version for boneage recognition with X-ray hand images. Furthermore retrieving bigger and more varied datasets could improve the generalization and resilience of these deep networks. In conclusion, the study demonstrates how the tested deep learning models can support medical professionals in making precise and effective assessments, which represent a reliable help in clinical decision-making and patient care.

References

[1] X. Chen, J. Li, Y. Zhang, Y. Lu, and S. Liu, "Automatic feature extraction in x-ray image based on deep learning approach for determination of bone age," Future Generation Computer Systems, vol. 110, pp. 795–801, 2020.

[2] Larson DB, Chen MC, Lungren MP, Halabi SS, Stence NV, Langlotz CP. Performance of a Deep-Learning Neural Network Model in Assessing Skeletal Maturity on Pediatric Hand Radiographs. Radiology. 2018

[3] Toğaçar M, Ergen B, Cömert Z. COVID-19 detection using deep learning models to exploit Social Mimic Optimization and structured chest X-ray images using fuzzy color and stacking approaches. Comput Biol Med. 2020