RSNA Bone-age Detection using Transfer Learning and Attention Mapping

Juan Camilo Castillo, Yitian Tong,Jiyang Zhao,Fengcan Zhu

***Abstract*— A fast, automated and accurate machine learning model for bone age assessment is proposed in this project. Bone age assessment is a common clinical practice in the diagnosis of child development, usually apply to those before 18. Nowadays there are various software model for bone age detection using computer vision. The error of BoneXpert, the most popular and state-of-the art systems in use for now, is about 8.4 months[1]. In our project, we trained various models with regression and Convolutional Neural Network with transfer learning, along with multiple image processing and feature extraction methods. Finally using the VGG16 pretrained model with attention mapping focused architecture we were able to achieve a mean absolute error (MAE) of 9.82/10.75 months for male and female patients which matches our goal of reducing the MAE under a year.**

I. Introduction

With the advancements in Machine Learning, Image Processing, and Statistical Learning, many problems in other fields have seen breakthrough technologies with new and innovative solutions. Medical Imaging in particular has seen a great deal of focus from the machine learning community, and as a result has produced novel ways of solving old problems. One problem our group in particular looks to focus on is in predicting bone age from a series of x-rays images. Given a training set of x-rays of an individuals hands and associated gender, our goal is to predict the bone age within a year tolerance. Associated problems include finding high level descriptors that accurately give insight to bone age. And it's no doubt that individual's gender affect the results. With the help of image processing and Convolutional Neural Networks (CNN) based transfer learning, we have got some satisfying results.

II. Related Works

Medical Imaging is a difficult problem because such images usually contain large homogeneous regions with little color variation. Most common approaches include creating handcrafted feature extractors that take domain specific knowledge into consideration. Recent software solutions such as BoneXpert, have been developed and approved for the clinical use in Europe. BoneXpert uses the Active Appearance Model[2],a computer vision algorithm, which reconstructs the contours of bones of a hand. Then the system determines the overall bone age according to their shape, texture, and intensity based on the Greulich and Pyle (GP)[3] or Tanner-Whitehouse (TW) techniques[4]. However, it is sensitive to the image quality and does not utilize the carpal bones, despite their importance for skeletal maturity assessment.

Feature extraction is a crucial step in most computer vision problems. The conventional feature detectors such as SIFT(Scale-Invariant Feature Transform) and SURF(Speeded-Up Robust Features)[5] usually perform well when applied to typical images taken by a digital camera or camcorder. However, when dealing with medical imaging problem, the performance is often not as good due to the little color variation of medical pictures. For bone age detection, even a small difference in the structure or size of crucial parts of hand bones can result in significant gap between the prediction result and the true value of age.

With the development of deep learning and supporting hardwares including GPU, many models with deep learning algorithm were trained for bone age prediction in the past decade. The error of these models are approximately 1 year [6], which is acceptable enough for clinical practice. Popular Neural Networks such as Inception,VGG are utilized for training the models, and they are all proved to be efficient[7].

III. Data Engineering

A. Dataset Analysis

The dataset was first used on Radiological Society of North America (RSNA) 2017 challenge and then it was released on kaggle for public access [8].

This dataset consists of X-ray scans of hands for people from ages 0 to 20. The training set have 12612 distinguish hand scans labeled by its owners age and gender. Since the challenge organizer did not release labels for the test set, we will split part of the training set for validation. Sample images from the training image set are shown in Figure. 1.

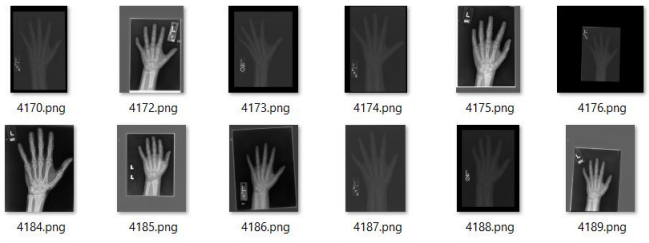

Fig. 1. Samples from RSNA dataset

Distribution of ages in the dataset is shown in Figure. 2. A large portion of the samples have age around 150 months while only a few samples falls near the two edges. The distribution of genders is not balanced as well, there are 6833 images for the male and 5778 images for the female.

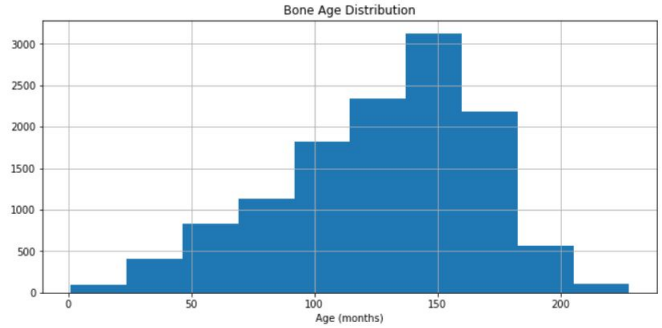

Fig. 2. Distribution of ages

B. Image Preprocessing

The idea of preprocessing of the hand radiography is to crop out the unnecessary part in the image. As shown in Figure. 1, there are labels in the image such as the 'L' sign which are irrelevant to detection of bone age. Also, the surrounding area of the hand in some images in Figure. 1 is large and irrelevant to train our model. The preprocessing is therefore to focus on extracting only the hand part in the image.

Since radiography images are naturally gray scale images with moderate contrast, a gradient based image segmentation is implemented in our algorithm. A illustration of the hand extraction process is shown in Figure. 3. Sobel gradient extractor is used to draw the gradient map of the original hand radiography. Based on this gradient map and predefined threshold for markers, a watershed image segmentation algorithm is implemented to mask all the connected objects in the original image and separate them from the background that have intensity lower than a threshold. The largest connected object is treated as the hand mask and apply back the the original image to extract the hand image.

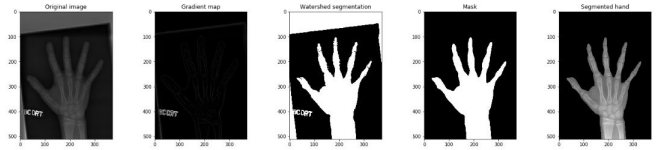

Fig. 3. Hand Extraction process

For the deep learning model, image augmentation is applied specifically. Since there are mostly left hands in the dataset images, a random horizontal flip is applied to the training set images for adding complexity to the dataset. For same purpose, shearing and zooming are applied to the images. Moreover, the hand in the image can be badly angled. A small rotation range for the image is applied to the deep learning model.

IV. Methods

A. Regression

For an image regression problem with high dimensional features and majority of them are indiscriminate features ,it's barely possible for simple regression method to achieve some good results, .However, this method can provide a basic understanding of the dataset. The result we got from regression can offer some guidance of how to improve our preprocessing and feature reduction methods.

Principal component analysis (PCA) is a statistical procedure which is usually used for dimensionality reduction. The basic idea of PCA is to discard the linearly-correlated variables and to keep the dimensions with the highest variance. It is employed in our regression method and proved useful for minimizing prediction errors. For the Linear Regression and Logistic Regression, we first resize the preprocessed images to 128*128, and divide dataset into two thirds are training and the rest images are test.

1) Linear Regression: Linear Regression fits a linear model with coefficients to minimize the residual sum of squares between the observed responses in the dataset, and the responses predicted by the linear approximation. Mathematically it solves a problem of the form:

$$min \parallel X\theta - Y \parallel^2$$

where X is the matrix of the whole training dataset and each row of X is the flattened vector of each bone image and Y is the vector of corresponding Age of each bone image.The size of X is m*n, which means the training set contains m samples and each sample has n features.Thus we created a mapping matrix between an image and it's bone age and used the mapping matrix for bone age prediction.

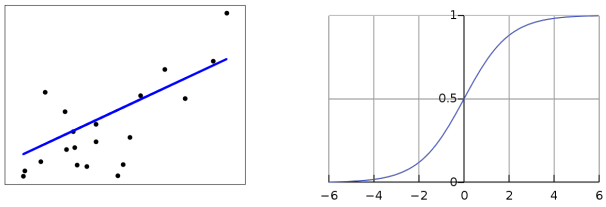

Fig. 4. Linear regression and logistic regression

2) Logistic Regression: Logistic Regression is also a widely-used method for machine learning projects. Although Logistic Regression is usually used as a classification algorithm, we can still use it here. For our project, there are limited number of regression results so the idea is to transfer the regression problem to a multi-classification problem. The prediction result is from 1 to 200 months so there are 200 different categories.We used Sigmoid function as the logistic mapping function and measured the relationship between each image and it's corresponding label. Then when a new picture is fed to the model, it can quickly be classified into a category so as to gain the result of bone age prediction.

However, even after preprocessing and dimensionality reduction, there are still many indiscriminate features we take into our regression training model, which leads to heavy overfitting. This shallow regression based models might not an ideal solution for our problem, so we'll introduce the other method we used for the project, deep learning.

B. Deep Learning

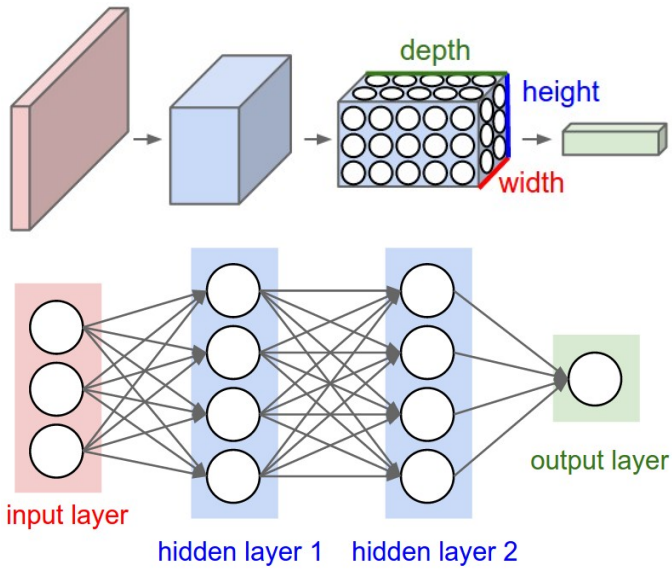

Fig. 5. CNN (top) vs ANN (bottom)

1) Convolutional Neural Networks: Convolutional Neural Networks (CNN) are a special type of Artificial Neural Networks (ANN) that specifically deal with raw images as the inputs. These inputs are then processed through a series of layers that preform a convolution followed by a set of non-linear operations. This specific type of Neural Network has been shown to be a very powerful model in many computer vision tasks , including object detection and image segmentation because of its ability to encode intrinsic characteristics about the image. Like the visual cortex found in humans, CNN's model the spatial locality of the input image, in the sense that every set of neuron in the brain are connected to only a few other neurons; as such the architecture of a CNN have layers of discrete filters, that learn the spatial shift invariant features of the image. A comparative image of the two different models are shown in Figure 5

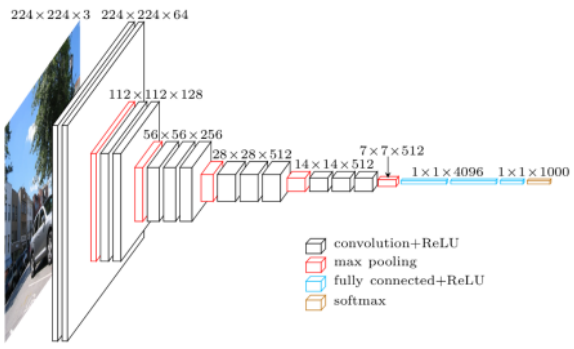

Fig. 6. VGG-16 Network

2) Transfer Learning: In its most simplest form Transfer Learning is the method of using a model or solution that was learned to perform one task and applying it to a another similar task. In deep learning this commonly refers to using a trained Neural Network as the starting off point for another Network that will need to learn a task such as classification or regression from a similar dataset. There have been many proven examples showing the benefits of transfer learning regarding images and with the ever growing amount of data used on images, transfer learning is becoming a powerful method for creating complex models that work from a small datasets.

Transfer learning is typically done in one of many ways:

1) By reusing the model architecture and using the pre-trained weights as an initialization point and then retraining the entire network end to end. This has the benefit of creating a model that works specifically for a problem and leverages the generalization of the first initial layers of a deep network.
2) Using the same model architecture and then using the pretrained weights on the first set of layers and freezing those layers during the training process, consequently only allowing the final layers to be trained. This usually leverages the fact that most of the beginning layers of a deep network can be thought of as a set of feature extracting layers, while the final layers perform that specific task in mind. As a results this process is effectively using a learned feature extract and training an associated network to do a another task.
3) Finally, the last common method of employing transfer learning is by using pretrained weights, freezing the first set of layers and then augmenting the final layers with a completely new architecture, curtailed for a specific problem.

3) Regression Architecture: For this project we employed the use of a CNN paired with the transfer learning methodology to help in the regression task of predicting bone age. The CNN used was the Oxford Visual Geometry Group (VGG) CNN, which took second place in the ImageNet ILSVRC-2014 Challenge. Figure 6 shows the architecture of VGG; taking particular note of the final fully connected (FC) layers that will be replaced with a custom architecture. The first custom architecture built for this problem used simple fully connected layers after a batch normalization layer of the VGG portion. Figure 7 shows the breakdown of this first model. This second model takes advantage of recent advances in understanding the characterization of the feature maps as individual descriptors of points in the image that help in the regression or classification task. With this in mind, we can take the average of each feature map and create a new weight vectors which reinforces the points of interest in the image. By passing these new features back in we can create a class activation map or attention map that visualizes the areas in the image that most heavily weigh in the regression task. The complete description of the model is shown in Figure 8.

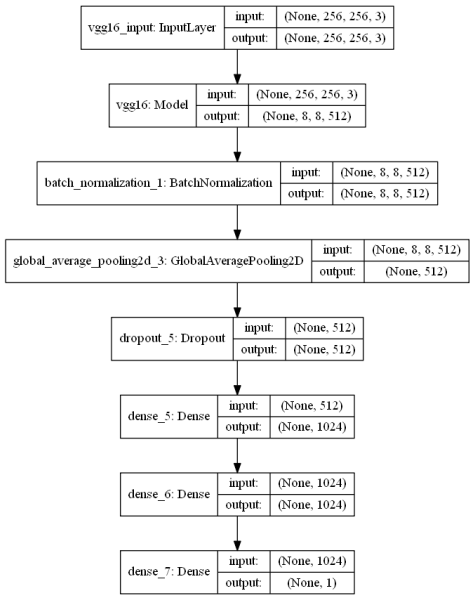

Fig. 7. First Model with FC layers at the output

V. Result

Same image preprocessing and feature selection methods were applied to the dataset before it was fed into the regression models and the transfer learning models. Due to the possible influence of bone development by gender, we experimented training with the whole dataset together and each gender separately. Mean absolute errors (MAE) of months were used to measure the difference between the predicted age and the actual age from the label. 20 percent of the dataset was randomly selected as the validation set, MAE achieved using the four models on the validation sets are shown in Table 1.

TABLE I
MAE FOR VALIDATION RESULTS

Models	Male	Female	Both Gender
Linear Regression	30	29	
Logistic Regression	36	33	
VGG16 Transfer learning + FCs	15.21	16.15	16.88
VGG16 Transfer learning + attention mapping	9.82	10.78	11.45

It's clear that the traditional Linear Regression and Logistic Regression performance much worse than the Transfer Learning models.The first model consisting of only fully connected layers at the output and was trained using Adam Optimization for 30 epochs.The second model was also trained using Adam Optimization but was trained for 50 epochs, and compared to the first model that was fed 256x256 input images, the second model was fed 500x500 images. Data augmentation was used to reduce the possibility over-fitting considering the large number of parameters in Model 1 (1.5 Million) and Model 2 (.5 Million) and comparatively small dataset(12k Images). This data augmentation applied a random rotation of up to 5 degrees, random horizontal flip,

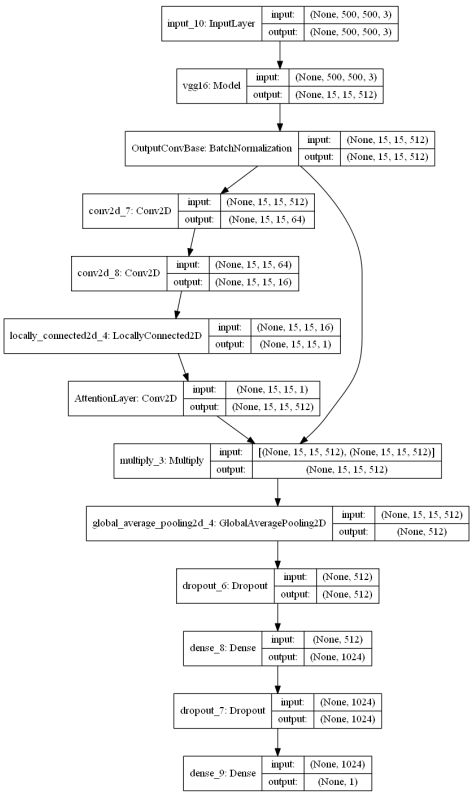

Fig. 8. Second Model with Attention Mapping layers at the output

a height shift of 0.15, a width shift of 0.15 ,a shear range of 0.01 and a zoom range of 0.25.

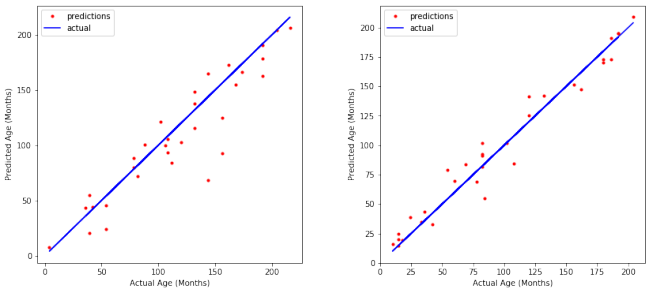

Fig. 9. Age predictions from Model 1 (left) and Model 2 (right) vs true age predictions

Figure 9 show the resulting predictions after training. The blue line is a plot of the ground truths, while the red points are individual predictions out of each model. Figure 10 shows the output of the multiply layer for each input image.

VI. Discussion

For our regression models, we tried various ways to improve the result of prediction. The main problem is overfitting because only some important parts of hand bones have influence on deciding the age of bones. But we used the whole pictures for training, which resulted in a great number of indiscriminate features being taken into account. We tried to reduce the number of feature points by using

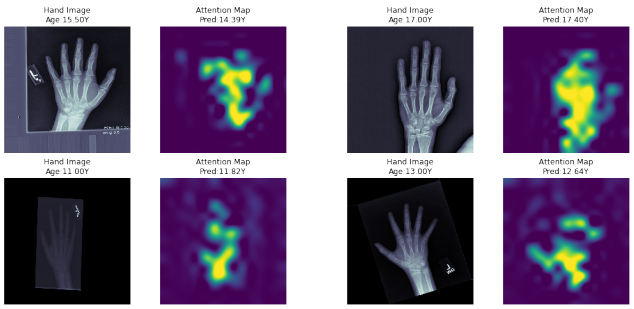

Fig. 10. Attention Mapping for images in Validation Set

feature descriptors such as SIFT of SURF discussed above. However, both SIFT and SURF tend to use the key points with large gradients. They focus too much on the pixels that are different from ambient ones. For medical images that are highly homogeneous like X-ray bone images, the focus will be on the edge of hands' contour which plays negligible roles in age prediction. Vladimir[7] proposed using deep learning methods to detect the crucial parts trained by several hundreds of manually labeled images to build a model that can automatically detect the regions of interest over X-ray scans of hands. But this requires medical background knowledge of bones and manually ROI selection. If we can expect the help from people with medical background and have more time, we believe we can optimize our models better.

Our models have higher prediction accuracy for the age at the very end as shown in Figure 9. It may be caused by the fact that girls grow up much faster than boys around the age 10, which leads to high variances when training two genders together. And the high prediction variance can be also caused by puberty and that's when human bones grow fastest in people's lifetime. According to Figure 10 for the attention mapping, the carpal and metacarpal bones contain more information on the bone age prediction than other areas of the hand.

Comparing the images produced by the two deep learning models; Figure 9 shows how the harder to classify points are those between 100 to 175 months, which corresponds to the period when adolescence begins and ends. This as a results explain the large amount of variance found during those months. We believe the second model performed better than the first because of the added layers found after the VGG section and before the fully connected section. This model essentially performs an ROI highlighting that puts particular parts of the image in focus. This consequently allows the fully connected layers to perform better decision making process.

VII. Conclusion

We achieved MAE of 9.82/10.75 months for male and female using VGG16 pretrained model and attention mapping. The result is similar to the 9.84/11.16 achieved by Fully Automated BAA [6] using the same dataset. The most salient features for predicting the age of an individual clearly seems to be the bones found in the wrist and middle of the hand. Future work can include trying different architectures and analyzing the associated efficacy of the implemented designs.

References

[1] Daniela Giordano, An Automatic System for Skeletal Bone Age Measurement by Robust Processing of Carpal and Epiphysial/Metaphysial Bones, IEEE TRANSACTIONS ON INSTRUMENTATION AND MEASUREMENT, VOL. 59, NO. 10, OCTOBER 2010

[2] T. F. Cootes, G. Edwards, and C. J. Taylor, Active appearance models, IEEE Trans. Pattern Anal. Mach. Intell., vol. 23, no. 6, pp. 681685, Jun. 2001.

[3] Greulich WW, Idell Pyle S. Radiographic atlas of skeletal development of the hand and wrist. Am J Med Sci;238: 393,1959

[4] Tanner JM, Whitehouse RH, Cameron N. Assessment of skeletal maturity and prediction of adult height (Tw2 method). 1989.

[5] Dusty Sargent, Chao-I Chen, Chang-Ming Tsai, Yuan-Fang Wang, Daniel Koppel, "Feature detector and descriptor for medical images," Proc. SPIE 7259, Medical Imaging 2009: Image Processing

[6] Hyunkwang Lee,Fully Automated Deep Learning Systemfor Bone Age Assessment,J Digit Imaging (2017) 30:427441

[7] Vladimir Iglovikov, Alexander Rakhlin, Alexandr Kalinin, and Alexey Shvets, Pediatric Bone Age Assessment Using Deep Convolutional Neural Networks,Dec. 14, 2017

[8] Radiological Society of North America (RSNA). RSNA bone age dataset. Link:https://www.kaggle.com/kmader/rsna-bone-age. 2018

[9] Wikipedia page for Regression and PCA and document for Scikit-Learn package