Fully Automated Deep Learning System For Bone Age Assessment

Weiyi Tang
tangwy@seas.upenn.edu

Shiao Li
shiaoli@seas.upenn.edu

Tiening Li
tiening@seas.upenn.edu

Songyue Lu
songyuel@seas.upenn.edu

***Abstract*—In this project, we want to implement a fully automated deep learning pipeline to segment a region of interest, standardize and preprocess input radiographs, and perform Bone Age Assessments. We first normalize our input images, because they have different sizes and background color. We then try different ways including Image Processing Technology, Single Shot MultiBox Detector (SSD) and Mask R-CNN to remove annotation markers and background border from image to make our data more clearly. We finally feed processed pictures to GoogLeNet to predict the age of bone. As for the result, we can get accuracy of roughly 95% if allowing error within 2 years, while roughly 80% if allowing error within 1 years.**

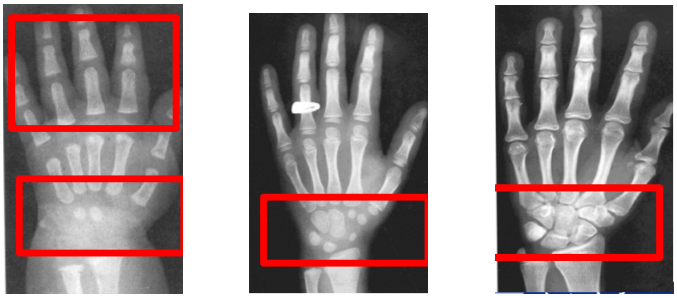
***Index Terms*—Bone Age Assessment, CNNs, Image Processing, SSD, Mask R-CNN**

I. Introduction

Skeletal maturity progresses through a series of discrete phases, particularly in the wrist and hands (As shown in Fig 1). As such, pediatric medicine has used this regular progression of growth to assign a bone age and correlate it with a child's chronological age. If discrepancies are present, these helps direct further diagnostic evaluation of possible endocrine or metabolic disorders. Alternatively, these examinations may be used to optimally time interventions for limb-length discrepancies. Bone age assessment is the ideal target for automated image evaluation as there are few images in a single study (one image of the left hand and wrist) and relatively standardized reported findings (all reports contain chronological and skeletal ages with relatively standardized keywords, like "bone age" or "year old"). This combination is an appealing target for machine learning, as it sidesteps many labor-intensive preprocessing steps such as using Natural Language Processing (NLP) to process radiology reports for relevant findings.

II. Related Works

Our project is to reproduce a paper [1]. This paper first implemented detection CNNs to detect bones and tissues, construct a hand/wrist mask, and apply a vision pipeline to standardize images. Then they fed their processed image to GoogLeNet to predict bone age. We tried different methods to achieve their first phase, then decided to utilize Mask R-CNN to reach our goal. We also implemented GoogLeNet to help us predict bone age.



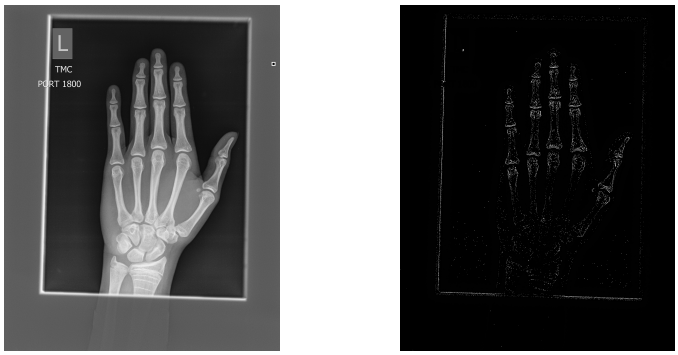
6 months old | 5 years old | 14 years old

Fig. 1: discrete phases of different age

III. Approach

A. Image Processing Technology

The first approach we have tried to remove markers and background is first segment the hand and fingers from the image, then create another image with just the hand silhouette. Once we have the silhouette image we can erode the image to make it a little smaller by using OpenCV package. But this method failed, because import pixels of hand will lose though we can successfully remove markers(As shown in Fig. 2).



Original Image | Result Image

Fig. 2: Failed result

B. Single Shot MultiBox Detector

Then we are thinking about utilize CNNs to help us find the position of hand. There are many cases that can detect human's real hand instead of hand bone, so we want to try to

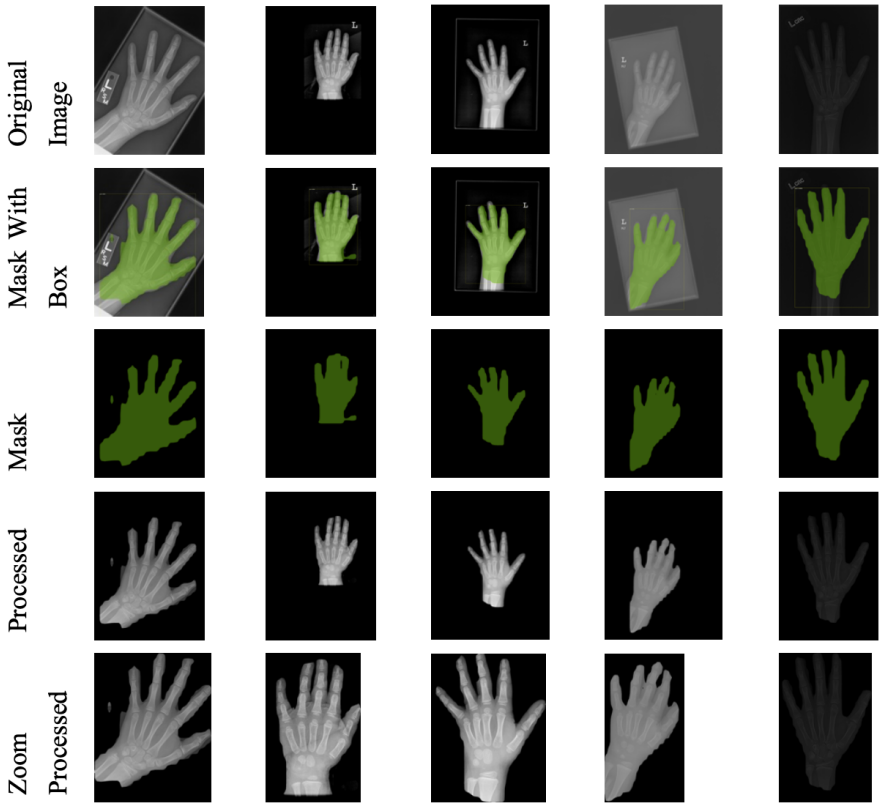
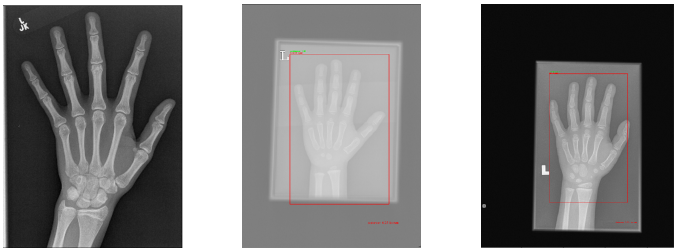


Fig. 3: Result of Mask R-CNN

use real human hand as training date set and see how good trained model can be to detect hand bone. We tried a neural network based on Tensorflow named Single Shot MultiBox Detector (SSD) that was trained on EgoHands dataset by Indiana University. The EgoHands dataset contains 48 Google Glass videos of complex, first-person interactions between two people and provides high quality, pixel-level segmentations of hands, which is frequently used as training and validation sets of hand detection on images or real-time webcam video streams. The SSD model we used implement transfer learning by taking a pre-trained model (especially its weights) named ssd_mobilenet_v1_coco in Tensorflow model zoo to detect hands after we train it again on frozen graphs from EgoHands dataset. According to the model name we know that it is a SSD model trained with COCO dataset, which is another widely used set in object detection. However, there is a fatal defect that both COCO and EgoHand only contain figure of hands in normal shape rather than X-ray photo, which looks quite different. The x-ray photos highlighted hand bones and make other tissue more transparent, so it is difficult for model trained by normal hand images to recognize them as "hands". This unreliable performance is shown in Fig. 4, only some hands of the x-ray photos can be detected (with red bounding box).



failed Successful Successful

Fig. 4: Cases of failure and success by using SSD

C. Mask R-CNN

We finally decided to utilize Mask R-CNN to extract hang bone from image. Mask R-CNN is conceptually simple: Faster R-CNN has two outputs for each candidate object, a class label and a bounding-box offset; to this it adds a third branch that outputs the object mask. Mask R-CNN is thus a natural and intuitive idea. But the additional mask output is distinct from the class and box outputs, requiring extraction of much finer spatial layout of an object. And Mask R-CNN usually gives good result with small train data set, which is good for us, because we need to make all the train data set by ourselves.

We select 50 images from the rsna training data set, and use VIA(VGG Image Annotator) only to label the hand position in the image. This is because we only care about the hand's figure and not interested in other parts. The learning rate is 0.001, number of epochs is 30, training steps for every epoch is 200. The result is shown in Fig. 3.

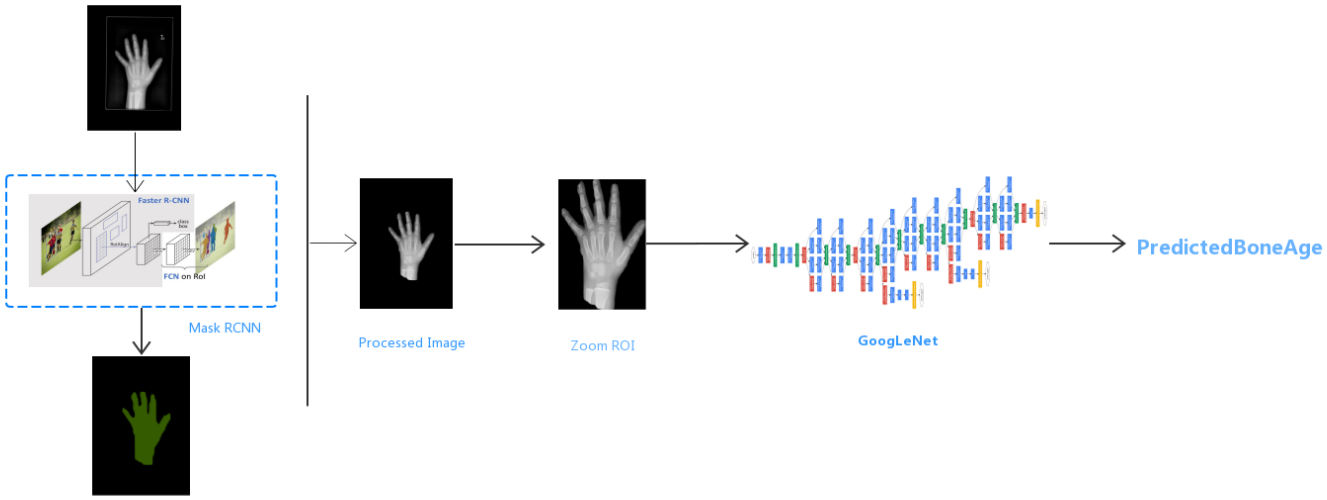


Fig. 5: Bone Age Assessment Flow Chart

TABLE I: Result

Group	Accuracy(allow error within 2 years)	Accuracy(allow error within 1 years)	mAP	RSME (month)
A1	94%	77%	34.6%	10.92
B1	95%	83%	33.5%	8.40
A2	94%	76%	40.3%	11.40
B2	96%	85%	38.8%	7.34

D. Final step of image preprocessing

When we have more clear images, we need to group our images, in others words, we need to label our data. We first divide the dataset into male and female. Then we utilize two ways to label bone age for each gender. We group data at 12-month intervals (e.g. 11 months old belongs to label 0) and 6-month intervals (e.g. 11 months old belongs to label 1). In the end, we have four groups: A1 (female, 6-month intervals), B1 (male, 6-month intervals), A2 (female, 12-month intervals), B2 (male, 12-month intervals). And we will have results for each group.

E. Image classification

Next step is to build a system to predict the boneage of the input radiograph. We choose deep learning since transfer learning can bring us a better result. AlexNet,VGGNet and GoogLeNet are the candidates of our system. We finally use GoogLeNet who has sparse connectivity and won the champion of ILSVRC14. Also, GoogLeNet has higher prediction accuracy than AlexNet and is more efficient than VGGNet. Firstly, we use gzip and pickle module to pack all the training images in size of 500. Meanwhile, all the images are resized into 227×227×3. The learning rate is set to be 0.001, the batch size is 32 (75% GPU usage). We use categorical cross-entropy as Loss Function to define the distance between the actual output (probability) and the expected output (probability) (in 'googlenet.py' line 153-155). We also set 10% of each training data to be the validation to prevent overfitting. Then, we use A1, B1, A2, B2 as training data and get four models, which can be used to predict the boneage of the test data. The whole flow chart of our approach can be seen in Fig. 5.

IV. Result

The result table can be seen in TABLE I. Here, we discuss the accuracy which allows error within 2 years and 1 year, which is the same with medical standard. Our result implies that we can use the system to correctly predict the boneage of the input radiograph. The prediction rate of each picture is 0.5s per image, which is higher than manual efficiency. We also categorize the test dataset by different age periods: 4-7 years, 7-13 years, 13-15 years, 15+ years. When the bone age

TABLE II: Compare With Other Team

Group	Image Size	Accuracy(allow error within 2 years)	Accuracy(allow error within 1 years)	mAP	RSME (month)
This project	227x227	94%	77%	34.6%	10.92
Lee from Harvard University	512x512	90.39%	84.57%	33.9%	18.12
Iglovikov from Russian	2000x1500	Not mentioned	Not mentioned	75.4%	8.08

is too big or too small, the accuracy will become lower since we do not have enough training data. mAP and RSME can be computed by the equation below.

$$mAP = \frac{1}{Q_r} \sum_{q \in Q_r} AP(q)$$

$$RSME = \sqrt{\frac{1}{m} \sum_{i=1}^{m} (h(x^{(i)}) - y^{(i)})^2}$$

From the TABLE II , we can find that the Russian team has a lower RSME and high mAP. This is because that they extract features to train. And their ROI are finger phalanx, ulnar radius and carpal bone, and they conclude that using carpal bone as feature can get a better accurate result. Furthermore, their image size is the largest, which means that their figures have more information to train. And the bone age between 7-15 years have a better prediction result. This is because we have the most data in this age range. And although the features are obvious when bone age is under 4 years. However, it is difficult the data for the new baby.

V. Discussion

In order to get more accurate result, several steps are considered to take beyond our project. First, when pre-process the raw data, images of right hand are better to be flipped horizontally as the model is trained based on left hand images. Besides, images with askew hand need to be rotated upright, and resolution of resized image should be as large as possible to keep more information as we can. Meanwhile, the x-ray photos of whole hand can be finely divided into three segments that contain finger phalanx, ulnar radius and carpal bone individually for training and classifying.

Furthermore, some research of Stanford University indicate that applying VGGNet on BAA can get a considerable accuracy as well as low RMSE, so implement VGGnet and the other networks on BAA is worth trying.

Finally, if we can plot attention map, which means the region that affects the result most, we can have a better performance by focusing on this region.

References

[1] Lee, H., Tajmir, S., Lee, J., Zissen, M., Yeshiwas, B. A., Alkasab, T. K. et al. (2017). *Fully Automated Deep Learning System for Bone Age Assessment*. Journal of Digital Imaging, 30(4), 427–441. doi: 10.1007/s10278-017-9955-8.

[2] Iglovikov, V., Rakhlin, A., Kalinin, Shvets, A. et al. *Pediatric Bone Age Assessment Using Deep Convolutional Neural Networks*. doi: 10.1101/2341200.

[3] Mansourvar M., Raj R. G., Ismail M. A. et al. *Automated web based system for bone age assessment using histogram technique*. Malaysian Journal of Computer Science, 25(3), 107–121.

[4] Byeong-Uk, B., Woong, B., Kyu-Hwan, J. et al. *Improved Deep Learning Model for Bone Age Assessment using Triplet Ranking Loss*.