Predicting Bone Age from Hand Radiographs Using Deep Convolutional Neural Networks

Caroline Kimmel
Mathematical and Computational Science
Stanford University
ckimmel@stanford.edu

Amin Ojjeh
Computer Science and Philosophy
Stanford University
amino@stanford.edu

Samir Safwan
Mathematical and Computational Science
Stanford University
ssafwan@stanford.edu

Abstract

Skeletal age (SA), measured against a child's age (CA), is integral for understanding and managing various endocrinopathies and growth disorders in pediatric patients [1]. Skeletal age assessment via hand radiograph is the most reliable and widespread clinical procedure for predicting SA. Existing popular assessment methods, however, date back to the 1960s. In 2017 the Radiological Society of North America (RSNA) challenged the data science community to leverage computer vision to predict SA more efficiently and accurately than existing methods. The primary goal of this paper is to use convolutional neural networks to build a model to predict bone age with high accuracy given RSNA's provided data set. The secondary goal is to leverage deep learning visualization techniques for better interpretation of our results. Overall, our selected final model achieved a competitive mean absolute deviance of 7.279 months on the test set provided by RSNA.

1 Introduction

In fall 2017, the Radiological Society of North America (RSNA) challenged the data science community to leverage computer vision and CNNs to predict SA. They shared their challenge and accompanying hand radiograph data set via Kaggle, an online platform for predictive modeling and analytics competitions [10]. RSNA aims to yield an accurate model from the competition that leverages technological innovation to promote excellence in patient care [2].

In this paper, we fit a deep convolutional neural network to RSNA's provided data to predict bone age with high accuracy. Our final model architecture is influenced by the published models of successful Kaggle competitors in conjunction with additional tuning and alteration by us. We employed visualization techniques to gain a deeper understanding into how convolutional neural networks tackle the bone age prediction challenge. Our goal is to accurately and efficiently predict SA given the hand radiograph data set.

Our model takes in as input an arbitrary-sized black-and-white hand radiograph, paired with an indicator of male-female gender. We then use a convolutional neural network, comprised of an Inception v3 network reading the hand radiograph and a fully-connected network reading the gender, to output a predicted numerical bone age of the radiograph in months.

CS230: Deep Learning, Spring 2018, Stanford University, CA. (LateX template borrowed from NIPS 2017.)

Formally, this is a regression task with inputs $X \in \{(\{0,\ldots,255\}^{H\times W}, \{0,1\})\}$ and outputs $\hat{y} \in \{\mathbb{R} > 0\}$. We selected mean absolute deviance (MAD), calculated as the mean of the absolute values of the difference between the model estimates and those of the reference standard bone age, as our loss criterion, as instructed by RSNA's challenge. We define $MAD = \frac{1}{n}\sum_{i=1}^{n} |y_i - \hat{y}_i|$.

Our baseline goal is to achieve MAD of 0.7 years (8.4 months), outperforming Bonexpert, the only automated SA assessment tool on the market at present [11].

2 Related work

The most common clinical methods for measuring SA are the Greulich and Pyle (GP) and Tanner and Whitehouse (TW) methods [1]. Both rely on human observation, making them subjective and variable. Using the GP method, used by 76% of pediatricians, a doctor compares a given hand radiograph to a series of hand radiograph templates to determine the closest match. The templates consist of 31 hand radiographs of male children between 0 and 19 years of age and 27 hand radiographs of female children between 0 and 18 years of age. A well-trained radiologist can use this method in 1.4 minutes on average, making it relatively quick and easy to use; however, it is less reliable than the TW method [1]. The TW method takes a radiologist 7.9 minutes on average. The radiologist computes a score for the radiograph by assigning points to the radius, ulna, short bones, and carpal bones via detailed structural analysis. Separate scores are used for male and female children [1].

There exists a fully-automated SA assessment tool: BoneXpert, marketed to pediatric clinics, was presented in 2008 [11]. Its accuracy is 0.71 to 0.72 years, and its precision is 0.17 to 0.18 years. These results make it the most state-of-the-art SA assessment tool used in practice [6]. However, the software does not account for carpal bone development and is limited to a SA of 2.5 to 17.0 years for males and 2.0 to 15.0 years for females, restricting its applicability [1]. RSNA's Kaggle competition yielded many deep neural network models for predicting SA from their data set.

The competition winners Mark Cicero and Alexander Billbily incorporated an Inception V3 network to design the best model. They published their methodology online with the title "Machine Learning and the Future of Radiology: How we won the 2017 RSNA ML Challenge" [8]. Their model incorporated gender as a predictor, and they achieved a mean absolute deviation (MAD) of 0.36 years, or 4.32 months using an ensemble of deep convolutional neural networks (note that the individual best model in the ensemble had MAD of 5.99 months) [8]. Another competitor Kevin Mader designed his model using a pretrained VGG16 network with additional convolutional layers using the attention mechanism [3].

3 Dataset and Features

RSNA provides training and unlabeled test data for the SA assessment challenge on Kaggle [3]. The training set consists of 12,612 unique arbitrary-sized hand radiograph images and a CSV file of identifier, bone age, and gender for each. The test set has 200 unique hand radiograph images and a CSV file of identifier and gender for each. The identifier for each image is its file name, which is a unique number appended to ".png". For example, the second image below is identified as "1459.png".

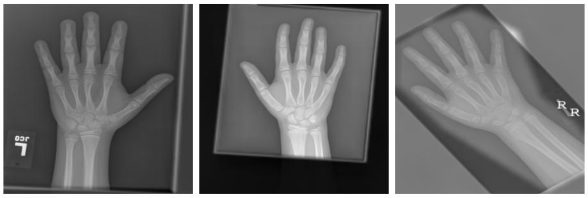


Figure 1: Example Hand Radiographs

Both the training and test sets consist of a variety of left and right hand radiographs (some images have both hands) at various angles. All radiographs in the sets are arbitrarily-sized. For our model, we split the provided labeled data into training and validation sets according to an 80:20 split, respectively, and incorporated random horizontal and vertical flips. Our CS 230 mentor David Eng provided us with a labeled test data set from RSNA on which we evaluated our architecture.

4 Methodology

Firstly, we resized all of the images to size 299 by 299 as specified for a pretrained *Inception v3* model, and converted the images to RGB to have a 3-channel input to the CNN. As a base model, we used a pretrained *Inception v3* model with a modified linear output layer for regression. Setting all gradients to false besides the new final layer, we were able to achieve mean absolute deviance (MAD) loss of about 33 after twenty epochs using the Adam optimizer and batch-size of 8.

Following our baseline model, we attempted to follow a similar approach as the winners of the RSNA challenge to incorporate gender into our model, unlike our baseline, which did not. To do so, we passed in gender as a binary variable (1 for male and 0 for female) into a hidden layer with 16 hidden units and ReLU activation. Meanwhile, images are fed into the *Inception v3* model. By modifying the available source code for PyTorch's Inception v3 model, found on GitHub [7], we were able to delete the fully connected layers after the final *Inception* module, flatten the output and concatenate this with the outputs of the layer that gender was initially fed into. Following this, we implemented two fully connected layers with ReLU activations, and a final linear output layer. Here is a diagram of this model:

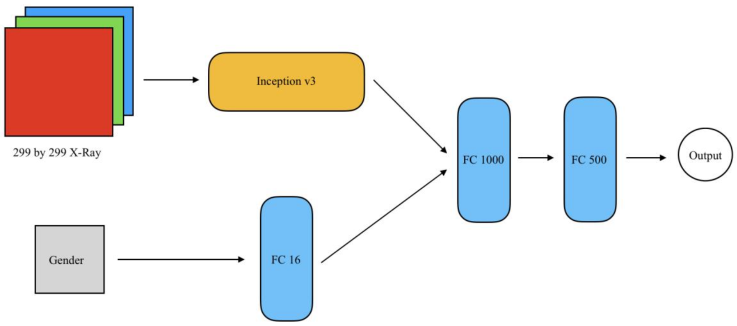


Figure 2: Model Architecture 1

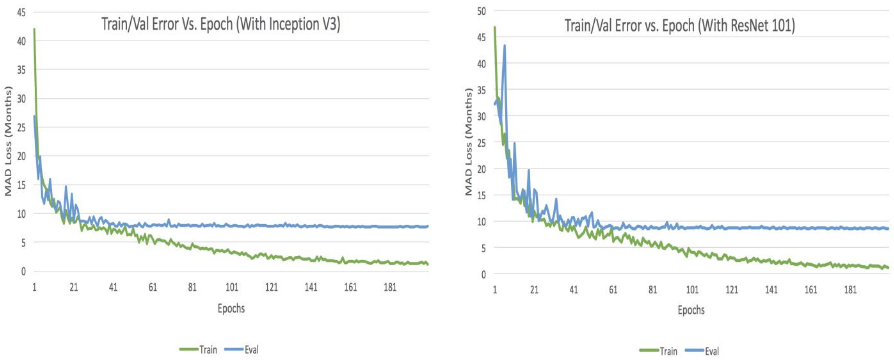
Further, we additionally tried using *ResNet 101*, a 101-layer convolutional neural network architecture that was used successfully in the ImageNet Challenge, to replace the *Inception v3* in our model, also removing the final fully connected layers; however, after hyper-parameter tuning, this model (Model Architecture 2) was not as successful as the one with the *Inception v3* module. The performance of these models will be assessed and compared in the proceeding section. We did not use pre-trained weights for either model, as Amazon Web Services's and Stanford's Institute of Computational and Mathematical Engineering's GPUs provided enough computing power.

5 Experiments and Results

For training, we used the mean absolute deviation (MAD) loss criterion to evaluate our model, per the expectations of the RSNA challenge. Furthermore, we used the Adam optimizer and an initial learning rate of 0.001 and implemented learning rate reduction when validation accuracy plateaued, as suggested by the RSNA challenge winners [8]. For both architectures, we explored using batch sizes of 8 and 16, and we found that batch size of 16 allowed us to achieve better results. Additionally, we found that using random horizontal and vertical flips improved our overall performance.

We trained our models for 200 epochs (about 30 hours for Model Architecture 1 and 50 hours for Model Architecture 2) and used the weights at the epoch when the minimum error on the validation set was observed. Here are plots of train/val error vs. epoch for the model with the *Inception v3* module and the *ResNet 101* module:

Figure 3: Train/Val Loss vs. Epochs



After hyper-parameter tuning, we found that the Model Architecture 1 (with *Inception v3*) performed better than Model Architecture 2 (with *ResNet 101*) with a MAD accuracy of 7.279 months compared to 8.766 months. Note that Model Architecture 1 outperformed our baseline goal of 8.4 months MAD accuracy. Thus, for our final model we selected the architecture that included the *Inception v3* module with the following hyper-parameters:

Figure 4: Hyper-Parameters for Best Model

Architecture	Model Architecture 1 (See Figure 2.)
Optimizer	Adam$(\beta_1 = 0.9, \beta_2 = 0.999, \varepsilon = 10^{-8})$
Learning Rate	0.001
LR Decay	ReduceLROnPlateau$($*factor* $= 0.8$, *min_lr* $= 10^{-4}$, *patience* $= 10$, *cooldown* $= 5)$
Batch Size	16
Epochs	200
Augmentations	Random Horizontal and Vertical Flips

6 Discussion and Next Steps

To understand what our model was learning, we used activation maps to visualize important characteristics of our inputs at various stages of our convolutional neural network. The figure below shows activation maps for two different radiographs with different hand orientations as the inputs progress through the model.

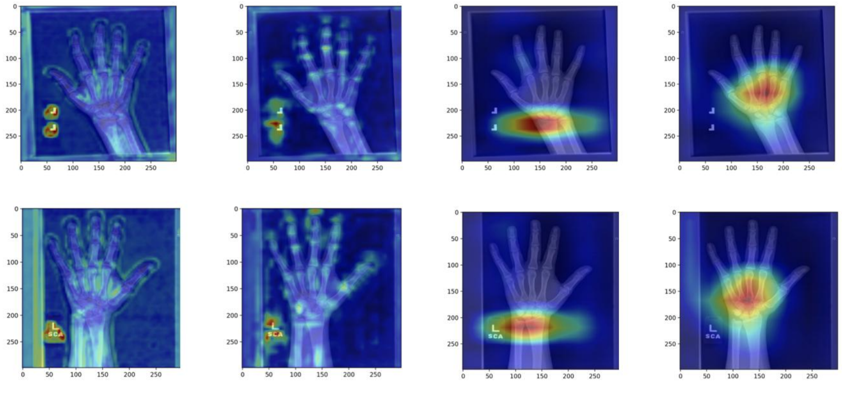


Figure 5: Activation Maps of Two Hand Radiographs at Various Convolutional Layers

The left-most images (earlier in the CNN) show activations at specific locations on the hands, such as individual epiphyses in the metacarpals and distal phalanges, which are important features in the prediction of skeletal bone age according to Dr. Dipak Sheth, a New England doctor familiar with skeletal bone age prediction methods. On the other hand, the right-most images (later in the CNN) show activations of general regions, emphasizing features like the wrist or multiple fingers.

Our selected model had strong overall performance on the test set and, in particular, had the highest predictive accuracy on radiographs of hands with greater skeletal bone age. However, after randomly sampling various radiograph that had absolute deviance higher than the observed MAD for the entire test set, it appears that the lower quality radiographs, as the one below on the right, on average had worse predictions than the higher quality radiographs, as shown in the example on the left.

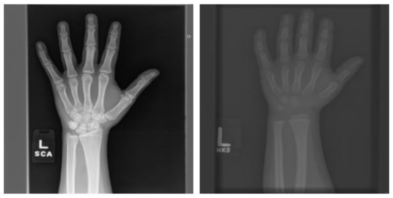


Figure 6: Test Examples - Female 192.59 Months (left) and male 73.26 months (right)

There are various steps we could take in the future to improve our performance on low quality images. Because each radiograph was arbitrarily sized to start with, shrinking and resizing inputs to 299 by 299 decreased the quality of various images. To prevent such an issue, some next steps that we plan to take would be to use larger input images and take advantage of more data augmentation strategies. Additionally, we found that, on average, our model had weaker performance on hand radiographs with skeletal bone age less than 36 months (3 years). However, much of the error due to younger radiographs can be explained by the lack of younger-aged images in the dataset provided by RSNA. To improve our results on these inputs, we could look into finding more data from different parties and re-balancing our validation set in a way to target these errors. By getting more data or using a validation set that contains a larger percentage of younger-aged radiographs, our model would be able to learn more features in these images, which would allow for improvement.

7 Final Remarks

Overall, our team was able to attain a model that successfully learns detailed hand features to predict skeletal bone age. Considering the smaller size of our input radiographs, the performance of our model (MAD test error of 7.279 months) is very competitive with the winners of the RSNA challenge, which achieved an overall MAD test error of 5.99 months on their best model using 512 by 512 radiographs as inputs.

Over the last decade, developments in computer vision have increased our ability to extract information from images tremendously. This technological development has not only been a step forward in the computer science industry but also has changed the healthcare sector. Due to algorithmic and computing improvements we are able to solve interdisciplinary problems throughout society. It is thus essential to understand what data and architectures are best for the task at hand, and to apply them to problems that can have the most significant impact.

8 Contributions

Here is a breakdown of our team member contributions to the the project:

- Topic research and data set selection - Caroline, Amin, Samir
- Research on RSNA data and Kaggle competition - Amin, Samir
- Additional research on bone age prediction - Caroline, Amin
- Project proposal write-up - Caroline, Amin, Samir
- Code:
  - `build_dataset.py` - Amin, Samir
  - `train.py` - Caroline, Samir
  - `model`: `inception.py, net.py, net2.py, data_loader.py` - Caroline, Samir
  - `evaluate.py` - Caroline, Samir
  - `CAM.py` - Samir
- Project milestone write-up - Caroline, Amin, Samir
- Project poster - Caroline, Amin, Samir
- Project final write-up - Caroline, Amin, Samir

Additionally, we would like to give a special thanks to our project mentor, David Eng, who worked closely with us in developing our model and to Dr. Dipak Sheth who briefly spoke to us about important features in predicting bone age.

References

[1] Sanctis, Vincenzo De, et al. "Hand X-Ray in Pediatric Endocrinology: Skeletal Age Assessment and Beyond." Advances in Pediatrics., U.S. National Library of Medicine, Nov. 2014, www.ncbi.nlm.nih.gov/pmc/articles/PMC4266871/

[2] "Competition." RSNA Pediatric Boneage Challenge, rsnachallenges.cloudapp.net/competitions/4#learn_the_details

[3] Mader, Kevin. "RSNA Bone Age." Kaggle, www.kaggle.com/kmader/rsna-bone-age/data

[4] Mader, Kevin. "Attention on Pretrained-VGG16 for Bone Age." Kaggle, www.kaggle.com/kmader/attention-on-pretrained-vgg16-for-bone-age

[5] De, V, et al. "Hand X-Ray in Pediatric Endocrinology: Skeletal Age Assessment and beyond." Advances in Pediatrics., U.S. National Library of Medicine, Nov. 2014, www.ncbi.nlm.nih.gov/pubmed/25538880

[6] Spampinato, C, et al. "Deep Learning for Automated Skeletal Bone Age Assessment in X-Ray Images." Egyptian Journal of Medical Human Genetics, Elsevier, 29 Oct. 2016, www.sciencedirect.com/science/article/pii/S1361841516301840

[7] Pytorch. "Pytorch/Vision." GitHub, github.com/pytorch/vision/blob/master/torchvision/models/inception.py

[8] "Machine Learning and the Future of Radiology: How We Won the 2017 RSNA ML Challenge." 16 Bit Blog, www.16bit.ai/blog/ml-and-future-of-radiology

[9] "Pytorch Documentation." Pytorch, pytorch.org/docs/stable/index.html

[10] "Kaggle: Your Home for Data Science." Kaggle, www.kaggle.com/

[11] BoneXpert, www.bonexpert.com/

[12] Pytorch, pytorch.org/

[13] Scipy, www.scipy.org/

[14] NumPy, www.numpy.org/

[15] Tqdm. GitHub, github.com/tqdm/tqdm.