Fully End-To-End Super-Resolved Bone Age Estimation

Mohammed Gasmallah, Farhana Zulkernine, Francois Rivest, Parvin Mousavi, and Alireza Sedghi

School of Computing, Queen's University, Kingston ON K7L3N6, Canada

Abstract. With the release of large-scale bone age assessment datasets and competitions looking at solving the problem of bone age estimation, there has been a large boom of machine learning in medical imaging which has attempted to solve this problem. Although many of these approaches use convolutional neural networks, they often include some specialized form of preprocessing which is often lengthy. We propose using a subpixel convolution layer in addition to an attention mechanism similar to those developed by Luong et al. in order to overcome some of the implicit problems with assuming particular placement and orientation of radiographs due to forced preprocessing.

Keywords: Image Processing · Bone Age Estimation · Convolutional Neural Networks.

1 Introduction

In radiology, bone age estimation is useful for a variety of reasons. Bone age is an indicator of the skeletal and biological maturity of a person and can often be different from the chronological age of an individual [6]. Bone age estimation is often requested for diagnosing pediatric diseases which indicates the maturity of a child's skeletal structure [6, 1]. Other measures have far too much variation in development to be used as established techniques for skeletal maturity [1].

Bone age estimation using radiographs is invaluable for pediatricians and orthopedic surgeons [1]. The most employed methods for bone age estimation are the Greulich and Pyle and Tanner-Whitehouse (TW2) atlases [1, 6, 7, 3]. A radiologists spends approximately thirty minutes per patient, comparing the radiograph to reference bone ages and estimating the age of the patient based off these references [3, 4]. This can be quite a time-consuming task and the accuracy and efficiency of the process is mainly determined by the experience of the reviewer in question [4].

Computerized methods and computer-assisted automated systems can help radiologists save precious time and increase accuracy at estimating bone age [4, 2]. The Radiological Society of North America (RSNA) has released a dataset and held a competition to assess and evaluate machine learning and automated bone age estimation methods [2]. In many of these methods, artificial neural networks are used to preprocess the dataset and transform the radiograph into

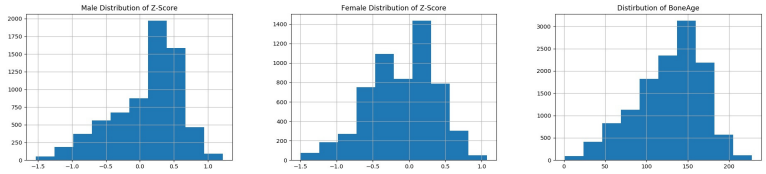
some standardized form to create a better and augmented training dataset [2]. For the same reason, we explored the technique of super-resolution by adding a subpixel convolution layer before the feature extractor in a deep convolutional network based on the work of Shi et al. [8]. Our model thus super resolves the input by extracting higher resolution information before passing it to feature extraction phase.

The rest of the paper is organized as follows. Section 2 discusses the methods and our proposed neural network model including the dataset and training and evaluation protocol. Section 3 presents the results and some additional images. Finally section 4 concludes the paper and discusses some future work.

2 Methods and Proposed Neural Network Model

We describe the dataset used to train the model and our proposed augmented neural network model, in this section. We discuss in detail our implementation and training/evaluating protocol as well as reasoning for our design choices.
Dataset: The dataset that was posted on Kaggle is from the Radiological Society of North America's pediatric bone age assessment challenge in 2018 [**?**]. The dataset contains 12611 images each labelled with the gender and bone age of the patient estimated by six reviewers. The ground truth was also provided in the dataset [2].

In order to make sure the network does not predict based on the non-uniform distribution of the dataset (see Fig.1) and train a good model, we resample the dataset based on bone age and sex. Additionally, we separate an initial 15% of the dataset to use for testing and another 25% for validation. The remainder is the training set and is augmented using rotation, horizontal/vertical flips and small affine transformations. This allows the network to learn the features that are necessary to evaluate the bone age regardless of rotation and other affine transformations of the image. It also allows the network to generalize better to radiographs which it has not seen yet. Finally, the images are normalized between 0-1 by dividing all pixel values by 255. [2].



(a) Male Distribution (b) Female Distribution (c) Total Distribution

Fig. 1: Distribution of bone age by gender and stand-alone of gender.

In order to make the regression smaller and more contained, we decide not to regress over the actual bone age of the patients. Instead, we regress over the z-score distributions of the patients' bone age using equation 1. This keeps

the values that the network must regress much smaller and allows for generally smaller gradients to flow through the network.

$$z_n = \frac{x - \mu}{2\sigma} \tag{1}$$

Where x is the bone age, μ is the mean of the bone ages, σ is the standard deviation of the bone ages and z is the z-score.

Model Architecture: As our base model, we use the network architecture that won the bone age estimation contest. We add on three blocks of convolutional neural network layers that lead to a subpixel phase shift layer. The subpixel phase shift layer is presented in Shi et al.'s paper and transforms depth wise information into space wise information [8]. The input image is then resized to match the size of the output of the subpixel phase shift layer and concatenated in order to feed into a feature extractor (such as VGG16 or Resnetv2 50). The output of the feature extractor is passed on to an attention module composed of a series of layers similar to those found in [5]. This accentuates particular areas of the feature extractor's output. The sex of the patient is also processed and then concatenated with the output of the feature extractor and attention module. Finally, two dense layers attempt to regress the z-score using this final output. The network architecture can be seen in Fig.2.

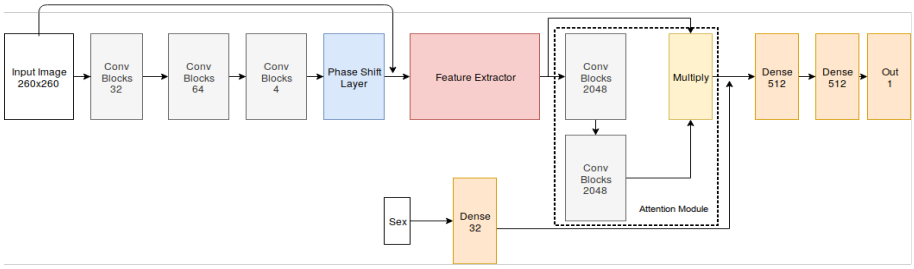


Fig. 2: Basic architecture used for bone age estimation. Feature extractor is replaced with either VGG or Resnetv2 50.

Variants for the purpose of testing include a network without the phase shift layer and any preceding convolutions, and two networks which either use the VGG16 feature extractor or the Resnetv2 50 feature extractor. Imagenet trained weights were loaded and the feature extractor was not trained. The final layers of the feature extractors were fed into our regressor and the attention module as illustrated in Fig.2.

Training and Evaluation: After we separate the dataset into the training, validation and testing sets, we began training the network using the ADAM optimizer. The ADAM optimizer is an optimizer which computes adaptive learning rates for differing parameters. The network is trained using a learning rate re-

duction scheduler, an early stopping mechanism to avoid overfitting and we only save the network which performs best on the validation set. The input dimension for the images is set to 260x260 and the batch size is set to 2 as any value higher causes certain variant networks to fail as they cannot fit in the memory of one graphics card. During training, the network is trained using a root mean squared error (RMSE) loss (see equation 2), however, a custom metric of mean average error is calculated so that we can directly compare performance when we relate it to the differences in months between the predicted value and the label.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(P_i - O_i)^2}{n}} \tag{2}$$

3 Results

Our results focus on the performance of the network using subpixel layers versus those that do not. In order to evaluate the networks, we decided to use the testing set to report the mean average deviation (MAD) and root mean squared error (RMSE) in both years and months. The networks reported in Table 1 have only been trained for a max of 25 epochs. The state of the art performance from the RSNA 2018 challenge was an ensemble network which was trained for 300 epochs on this problem [2]. It should be noted that the state of the art was trained on the same dataset as our network, but uses an ensemble method in order to outperform other methods. Additionally the images in Fig.3 provide insight on each networks' validation as it trains.

Table 1: Metrics of results comparing MAD score against RMSE in years and months.

Networks	Mean Average Deviance (months)	RMSE (years)	RMSE (months)
RSNA state of the art[2]	6.12	-	-
Resnetv2 50 **with subpixel layers**	36.31 ± 22.30	3.58	43.02
Resnetv2 50 with no subpixel layers	36.63 ± 23.00	3.60	43.26
VGG16 **with subpixel layers**	16.48 ± 13.46	1.77	21.28
VGG16 with no subpixel layers	17.93 ± 15.94	2.00	23.99

4 Future Works and Conclusion

Overall the use of the subpixel layers before the feature extraction network lead to inconclusive results regarding an increase in performance of the network. Although the networks with the subpixel layers do take longer to train and converge, they also tend to be much smoother when they reach convergence (see Fig.3). We have provided an example of the output of the super-resolution layer as well as the input in Fig.4. What is interesting to note is that the super-resolved hand includes higher activations in specific areas of the bone. Particularly, certain bone edges seem to be highly valued. This is the type of information we

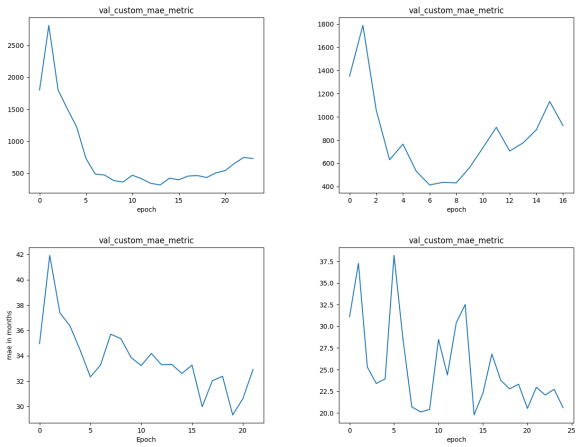


Fig. 3: Validation mean average error. Top left is VGG with super resolution layers, top right is VGG without the super resolution layers. Bottom left is Resnet with super resolution layers and bottom right is Resnet without the super resolution layers.

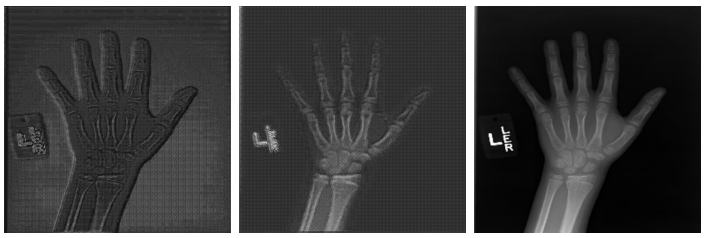


Fig. 4: The image on the far right is the input to the network, left is the super-resolution output for the network with Resnet and the middle is the super-resolution output for the network with VGG16 (both have brightness adjustments in order to be able to view the hand better).

expect that the network is learning and propagating through the super-resolution layers which is similar to those described in [6].

Although the networks we have described do not perform as well as many of the other networks as in [7, 2], the purpose of this endeavour was to view whether the subpixel convolution layers as introduced by [8] would be useful for the purposes of extracting and super-resolving information that the network deems important for the purposes of bone age estimation. In this endeavour, we believe our results are inconclusive. Although all networks are trained the same way and the subpixel convolution layers have lower MAD scores, and are more consistent, they do not perform significantly better (see Table 1). Regardless, the subpixel convolution layers are very useful and show that networks for bone

age estimation do not require more complex preprocessing steps but instead require important and useful layers that can aid or skip the preprocessing steps altogether and a more varied data augmentation phase.

There is still much to be done in this field. Bone age estimation is still an open problem in machine learning, although networks often perform better than radiologists and often more consistent [2, 7] they are less trusted and not as well understood. Additionally, the problem is quite simplistic and allows for a variety of different approaches to be taken. This makes it a useful problem to evaluate new and novel layers and how they perform in practical scenarios.

Future Works: There is still much to be done with the implementation of the subpixel convolutional layers. For one, we attempted early on to develop a multi-stage training step which required freezing particular layers of the network in order to focus on training either the subpixel layers or the regression layers. This proved to be unsuccessful, but we believe that modifying the training it may be possible to get the multi-stage training step to work. This changes the framing of the problem and intuitively, we believe that this will increase the network performance. Finally, looking into whether training from scratch or retraining the feature extractor may be beneficial.

References

1. Gilsanz, V., Ratib, O.: Hand Bone Age A Digital Atlas of Skeletal Maturity. Springer-Verlag, Los Angeles, 1st edn. (2005). https://doi.org/10.1007/978-3-642-23762-1
2. Halabi, S.S., Prevedello, L.M., Kalpathy-cramer, J., Mamonov, A.B.: The RSNA Pediatric Bone Age Machine Learning Challenge pp. 1–6 (2018). https://doi.org/10.1148/radiol.2018180736
3. Iglovikov, V., Rakhlin, A., Kalinin, A., Shvets, A.: Pediatric Bone Age Assessment Using Deep Convolutional Neural Networks (December) (2017). https://doi.org/10.1007/978-3-030-00889-5_34,
4. Kim, J.R., Shim, W.H., Yoon, H.M., Hong, S.H., Lee, J.S., Cho, Y.A., Kim, S.: Computerized bone age estimation using deep learning-based program: Evaluation of the accuracy and efficiency. American Journal of Roentgenology **209**(6), 1374–1380 (2017). https://doi.org/10.2214/AJR.17.18224
5. Luong, M.T., Pham, H., Manning, C.D.: Effective Approaches to Attention-based Neural Machine Translation. Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing pp. 1412–1421 (2015). https://doi.org/10.18653/v1/D15-1166, https://re-work.co/blog/deep-learning-ilya-sutskever-google-openai
6. Manzoor Mughal, A., Hassan, N., Ahmed, A.: Bone Age Assessment Methods: A Critical Review **30**(1), 211–215 (2014). https://doi.org/10.12669/pjms.301.4295
7. Mutasa, S., Chang, P.D., Ruzal-Shapiro, C., Ayyala, R.: MABAL: a Novel Deep-Learning Architecture for Machine-Assisted Bone Age Labeling. Journal of Digital Imaging pp. 1–7 (2018). https://doi.org/10.1007/s10278-018-0053-3
8. Shi, W., Caballero, J., Huszár, F., Totz, J., Aitken, A.P., Bishop, R., Rueckert, D., Wang, Z.: Real-Time Single Image and Video Super-Resolution Using an Efficient Sub-Pixel Convolutional Neural Network (2016). https://doi.org/10.1109/CVPR.2016.207