Fine-tuning of Pre-trained Deep Networks for Bone Age Assessment

Myrthe Wouters (Snr: 1273195)

1. Introduction

Skeletal maturity determined by hand bone age study is used as an adjunct in the care of patients with different diseases (e.g., endocrine, scoliosis)[3]. Traditionally, left hand radiographs are evaluated in clinical practice by one of two methods: the Greulich and Pyle method or the Tanner-Whitehouse method.[3] However, during the past few years, computer vision for bone age determination has shown to provide accurate results[1]. Publications on computer vision for bone age assessment generally take one out of two different approaches. The first approach includes entirely re-training well-known CNN architectures from scratch using both image and gender input, conducted by amongst others the winners of the RSNA Bone Age Challenge[2] and at least a few other publications[1,3]. With this approach, the winners of the RSNA Bone Age challenge reached a Mean Absolute Error (MAE) of 4.2 months[2]. On the other hand, multiple publications[4,5] use (fine-tuned) pre-trained models for bone age determination. This approach usually results in lower MAE scores. However, publications in the latter branch do not take into account gender inputs and/or do not report upon results on a separate test set. In this research, I will focus on deploying fine-tuned, pre-trained deep networks, taking into account both image and gender input. More specifically, the research question is defined as follows:

To what extent can we accurately determine bone age using fine-tuned pre-trained networks, including both image and gender inputs?

2. Experimental Procedure

For this research, I used the RNSA Bone Age dataset[6]. The dataset contains 12,611 hand X-rays with accompanied gender labels and ages ranging from 1 to 228 months for training and 200 X-rays, gender labels and ages for testing. The training set was divided into an 80:20 train/validation split, resulting into training and validation sets consisting of 10,088 and 2,523 instances respectively. During preprocessing stages, images were resized to 256x256 pixels with padding for computational reasons.

The models in this research incorporate a fine-tuned pre-trained architecture (VGG16 or ResNet18) on the ImageNet dataset to handle the image input and a gender network. The model is shown schematically in Figure 1. I used real-time data augmentation on the training set, including random rotations, width and height shifts, zooms and horizontal flips in order to prevent overfitting and increase the generalizability of the model[7]. An example image is presented in Figure 2. The models were trained for 20 (30) epochs for the model incorporating VGG16 (ResNet18) using the Adam optimizer, as used in other publications[2,3,4,5].

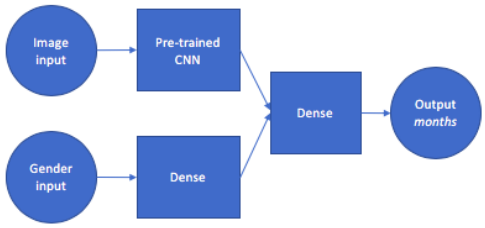

Figure 1: Model architecture

Due to Google Collaboratory resource constraints, I could not do extensive hyperparameter tuning through Bayesian optimization. However, I explored the following parameters experimentally: learning rate, number of fine-tuned layers in the pre-trained CNN and number of neurons in the final dense hidden layer. For details on the implementation and code, I refer to the GitHub repository[8].

Performance of the model is evaluated through the Mean Absolute Error (MAE), used in multiple other publications[2,3,4,5]. As an initial baseline, I consider the mean of the validation set, resulting in a MAE of 33.7 months. A more domain-based baseline is the MAE of pediatric radiologist. Prior research[3] has found a MAE of 6.0 months for radiologist on the RSNA bone age dataset.

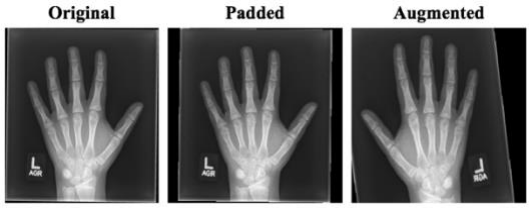

Figure 2: Example image

3. Results

Table 1 shows results of the current paper in comparison with existing literature. For the VGG16 and ResNet18 models including both image and gender input, I achieve MAE of 15.9 and 11.7 months respectively on the test set of 200 instances. Hence, the VGG16 (ResNet18) model is able to predict bone age within, on average, a deviation of 15.9 (11.7) months. MAE on the *validation* set for both the VGG16 and the ResNet18 model including gender was 10.6 months. Striking is the discrepancy between the validation and test MAE for especially the VGG16 model, which could potentially be caused by the considerably smaller size of the test set or the absence of batch normalization layers with slight regularization effects in the VGG16 architecture.

The VGG16 and ResNet18 model taking only image data as input achieve MAE of 16.9 and 14.6 respectively. These results suggest a positive influence of gender input on the model performance. However, these results should be interpreted with caution, as the models including image input only were trained with optimal parameters from the models including both gender and image input, instead of being tuned separately.

Authors	Model	MAE (months)	
		Image + Gender input	Image input only
Current paper	VGG16 (finetuned)	15.9	16.9
Current paper	ResNet18 (finetuned)	11.7	14.6
Sarić et al.[5]	VGG16 (finetuned)		12.2*
Sarić et al.[5]	ResNet50 (finetuned)		10.2*
Cicero & Bilbily[2]	InceptionV3 (retrained)	4.3	
Reddy et al.[3]	Xception (retrained)	4.7	

* Results on validation set

Table 1: Results and comparison to some existing literature

4. Discussion and conclusion

In summary, we can determine bone age using fine-tuned pre-trained networks including both image and gender inputs with a MAE of 11.7 months for the best-performing model incorporating the ResNet-18 architecture. These results are considerably better than the initial baseline of 33.7 months. However, results are not yet close to the baseline of radiologists' performance or the literature incorporating completely retrained (ensemble) models, reaching a MAE of approximately 4-5 months[2,3]. Limitations of the current approach include the relatively small image size and the lack of extensive hyperparameter tuning and image enhancement techniques. Furthermore, the models do not include explicit regularization, as literature suggests that data augmentation alone can achieve the same performance as regularized models and exhibits higher adaptability to changes in model architecture.[7] Nonetheless, the models in this paper might need additional explicit regularization in order to achieve higher generalization performance. To conclude, fine-tuning of pre-trained models for bone age assessment including both image and gender inputs does not provide as accurate results as completely retrained (ensemble) models. Future research is encouraged to experiment with bigger image sizes, extensive hyperparameter tuning and potential image enhancement techniques as segmentation and normalization[3] for improvement of finetuned pre-trained model performance.

1 https://pubs.rsna.org/doi/full/10.1148/radiol.2017170236
2 https://www.16bit.ai/blog/ml-and-future-of-radiology
3 https://link.springer.com/article/10.1007/s00247-019-04587-y
4 https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8935965
5 https://ieeexplore.ieee.org/document/8820451
6 https://www.kaggle.com/kmader/rsna-bone-age
7 https://arxiv.org/abs/1806.03852
8 https://github.com/myrthewouters/deep-learning