Final Project: Bone Age Prediction Using Deep Learning

Ekaterina Ustiukhina
Dept. of Electrical Engineering and Automation
Aalto University
Espoo, Finland
ekaterina.ustiukhina@aalto.fi

***Abstract*—This project focuses on predicting bone age from X-ray images using deep learning. During the project two model architectures were evaluated: a baseline model utilizing minimal preprocessing, and an improved model incorporating advanced augmentations and hyperparameter tuning. The improved model demonstrated a significant reduction in Mean Absolute Error (MAE) compared to the baseline, showcasing enhanced prediction accuracy and generalization. Key techniques included advanced augmentations such as rotation and contrast adjustments, and optimized hyperparameters.**

I. Introduction

Bone age prediction is a critical task in pediatric endocrinology, enabling clinicians to assess growth abnormalities. Traditional methods rely on manual readings, which are time-consuming and subjective. In this work, we leverage deep learning to automate bone age prediction, aiming to improve accuracy and efficiency.

Instructions on how to run the Jupiter Notebook with the project experiments can be found in Readme file submitted along with this report.

II. Data Preparation

A. Dataset Description

The dataset used for this project is publicly available on Kaggle [1] and consists of 12,611 X-ray images in the training set, each labeled with bone age (in months) and gender. The images are grayscale and vary in size. While the dataset also includes a separate 'boneage-test-dataset', it was not used in this project because it lacks the boneage (target) column included in the training set. Instead, only the training dataset was utilized, with a portion of it reserved as a validation set to evaluate model performance.

Figure 1 shows example of X-ray images from the dataset with corresponding labels.

Fig. 1. Example X-ray Images With Bone Age Labels

B. Data Cleaning

The following data cleaning techniques were applied to the dataset before starting the work:

- Missing Data: Rows with missing labels or image paths were removed.
- Duplicate Removal: Duplicates were identified and dropped.

C. Data Preprocessing

For the purpose of training optimisations with limited compute resources (limited usage or GPU in Google Colab) images were pre-resized to 224x224 and stored on Google Drive. This reduced computational overhead during training while preserving sufficient detail for accurate bone age prediction. The size was selected to match image size used during ResNet18 model pretraining [2].

D. Data Splitting

The dataset was split into training (80%), validation (10%), and testing (10%) subsets using stratified sampling to ensure balanced distributions.

E. Data Augmentation

1) Baseline Model

The baseline model was trained using images without augmentations (resize only).

2) Improved Model

For the improved model, the following extensive augmentations were applied:

- Rotation: Random rotations (±15 degrees) simulate slight variations in imaging angles while preserving anatomical structures.
- Shear: Random shearing in both x and y directions (±20%) mimics distortions from uneven imaging perspectives.
- Scaling: Random scaling in both x and y directions (90% to 120%) introduces variability in image size, helping the model generalize across differently scaled inputs.
- Contrast: Adjustments to contrast levels (70% to 150%) mimic variations in image quality caused by differences in X-ray machine settings.
- Brightness: Random brightness shifts (70% to 130%) simulate varying imaging conditions,

increasing the model's robustness to lighting variability.

- Blur: Occasional blurring (20% probability, medium-level blur) introduces realistic imperfections to mimic slightly out-of-focus images.
- Salt-and-Pepper Noise: Noise added with a gain range of 3% to 8% helps the model handle textual or minor imaging artifacts effectively.
- HSV Adjustment: Slight hue and saturation variations (±5) mimic color changes in grayscale images caused by scanning or digital processing.
- Cropping: Random cropping to a fixed size (224x224 pixels) encourages the model to focus on specific regions of the hand X-ray, improving its attention to critical bone structures.

Figure 2 shows examples of augmented X-rays, demonstrating random rotations, brightness shifts, contrast adjustments, cropping, and added noise. These augmentations enhanced model robustness.

Fig. 2. Example Augmented X-Ray Images

III. METHODOLOGY

A. Model Architecture

- Baseline Model: For the baseline model ResNet18 architecture was selected, pretrained on ImageNet. Due to ResNet18 was initially designed for image classification tasks on datasets like ImageNet, it was modified by replacing the final fully connected layer with a single-output neuron to adapt model for bone age regression task.
- Improved Model: For the improved model the same backbone ResNet18 architecture was selected, with the improvement in the final fully connected layer. The final layer was replaced with a sequential block, including a 256-neuron linear layer, ReLU activation, dropout (0.3), and a final linear layer for single-value output. This allows the model to predict continuous bone age values while leveraging pretrained weights for feature extraction.

B. Loss Function and Optimization

- Loss Function: Mean Squared Error (MSE) for regression. MSE is a standard loss function for regression tasks, penalizing large errors more heavily than smaller ones, which aligns with the goal of minimizing prediction deviations.
- Optimizer: Adam combines the benefits of RMSprop and momentum optimization, making it well-suited for non-stationary objectives like bone age prediction.

C. Training Configuration

- Batch size: 8
- Epochs: 30

The training process utilized a *Trainer* class adapted from Assignment 3, which provided a structured framework for initializing, training, and validating models. Minor modifications were made to accommodate this project's specific requirements, such as regression outputs and data augmentation.

D. Hyperparameter Selection and Justification

Several hyperparameters were selected based on experimentation and prior knowledge. Below is a detailed explanation of each parameter and its chosen value:

- Learning Rate: 0.001: This is a standard starting point for models trained with the Adam optimizer. It balances convergence speed and stability. Lower learning rates (e.g., 0.0001) resulted in slower convergence without significant accuracy improvements.
- Weight Decay: 1e-4: Weight decay prevents overfitting by penalizing large weights. The value 1e-4 was selected based on its success in prior experiments with similar architectures.
- Batch Size: 8: Due to hardware constraints (Google Colab), a smaller batch size was necessary. Larger batch sizes (e.g., 16 or 32) caused memory overflow errors.
- Number of Epochs: 30: The majority of training converged within 30 epochs.

By selecting these hyperparameters, we aimed to balance computational efficiency with model accuracy and robustness.

IV. RESULTS

A. Training and Validation Loss Curves

The training and validation loss curves for both models are shown in Figure 3.

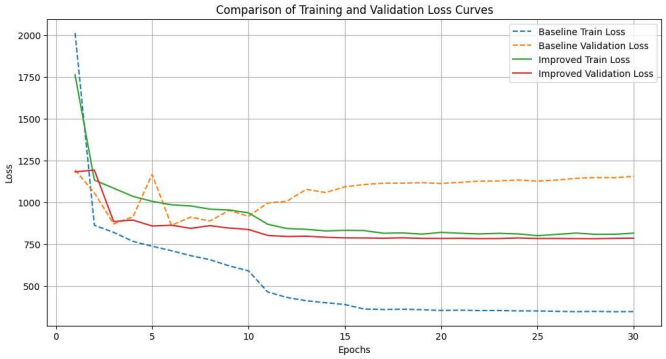

Fig. 3. Training and validation curves for the loss

The baseline model showed noticeable overfitting after approximately 15 epochs, indicating that prolonged training could harm its generalization. However, the improved model architecture demonstrated its ability to leverage the full 30 epochs without significant overfitting, leading to better convergence and overall performance.

In addition, the improved model shows a significant reduction in both training and validation loss, demonstrating better learning and generalization capabilities. The improved model's curves are smoother and more stable, indicating less overfitting and better training process optimization.

B. Model Performance and Statistical analysis

The improved model demonstrated lower mean MAE ower 10 epochs of Bootstrap conparison with the baseline model:

Mean MAE Difference (Improved - Baseline): -3.243.

Bootstrap analysis confirmed the improvement was statistically significant (95% CI: [-3.857, -2.386]).

Figure 4 compares predicted and actual bone ages for the baseline and improved models. Points closer to the red dashed line indicate better predictions.

The scatter plot for the baseline model shows a higher dispersion of predicted values around the perfect prediction line, indicating less accurate predictions and higher error.

In contrast, the scatter plot for the improved model displays points that are more tightly clustered along the diagonal line, showing predictions closer to the actual bone age and significantly reduced errors.

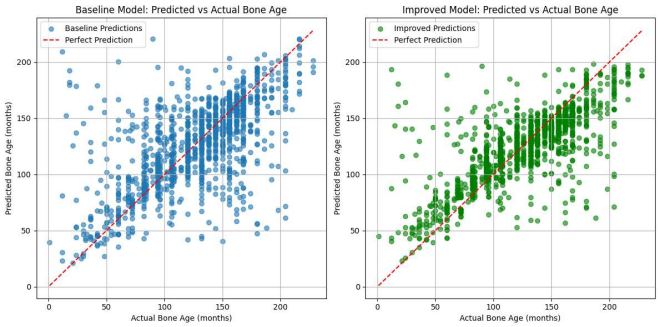

Fig. 4. Predicted vs. Actual Bone Age for Baseline and Improved Models

This visual evidence supports the numerical evaluation (Mean Absolute Error), where the improved model likely achieved a lower MAE, aligning with the observed better performance.

V. CONCLUTION

The improved model, which incorporates advanced data augmentations, enhanced model architecture and hyperparameter tuning, outperforms the baseline model in terms validation loss and prediction accuracy, while training loss was lower in the baseline model due to noticeable overfitting after approximately 15 epochs of training. The enhancements mentioned above significantly improve the model's generalization capabilities, enabling more reliable predictions of bone age from X-ray images. This highlights the critical role of preprocessing, data augmentation, and model architecture optimization in achieving superior performance in deep learning tasks.

Future improvements could be achieved by utilizing larger and more diverse datasets, which would better capture the variability seen in real-world scenarios. Additionally, exploring ensemble methods or leveraging larger pretrained architectures could further enhance model accuracy and robustness.

REFERENCES

[1] K. Mader, "RSNA Bone Age," Kaggle, 2017. [Online]. Available: https://www.kaggle.com/kmader/rsna-bone-age [Accessed: Dec. 12, 2024].

[2] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2016, pp. 770–778.