Automated Cephalometric Landmark Detection through Deep Learning and Computer Vision

Final Report

Cailyn Smith, Navdeep Bhusri, Yingshuang Yang

Problem Statement

Detection and annotation of anatomic craniofacial landmarks on a lateral cephalometric X-ray is an arduous task requiring a steep learning curve. Millimeter precision is imperative for detection of the anatomic landmarks since these identified landmarks are further used to conduct multiple Cephalometric analyses, which are one of the most important diagnostic tools available to orthodontists and oral-maxillofacial surgeons in treatment planning and decision process. However, existing learning-based approaches for anatomic landmark detection are time consuming and do not achieve the precision required to be utilized by medical professionals.

Solution

To address this challenge, we propose an automated system based on Convolutional Neural Networks (CNNs) for landmark detection. Our model accepts a lateral cephalometric X-ray as input and outputs a vector of (x, y) coordinates for all anatomic landmarks. Model performance is evaluated using Mean Squared Error (MSE) by comparing predicted coordinates against ground truth annotations. For our model with the ResNet18 backbone with pre-trained weight, we got an average validation loss of 0.0005 which translates to 0.0001762 mm with cephalometric x-ray images at 72 DPI. We also tried implementing a custom CNN model as a comparison, but that had a much higher loss. See Figure 1 for our loss curve over training epochs and Figure 2 for example inference outputs on our trained model.

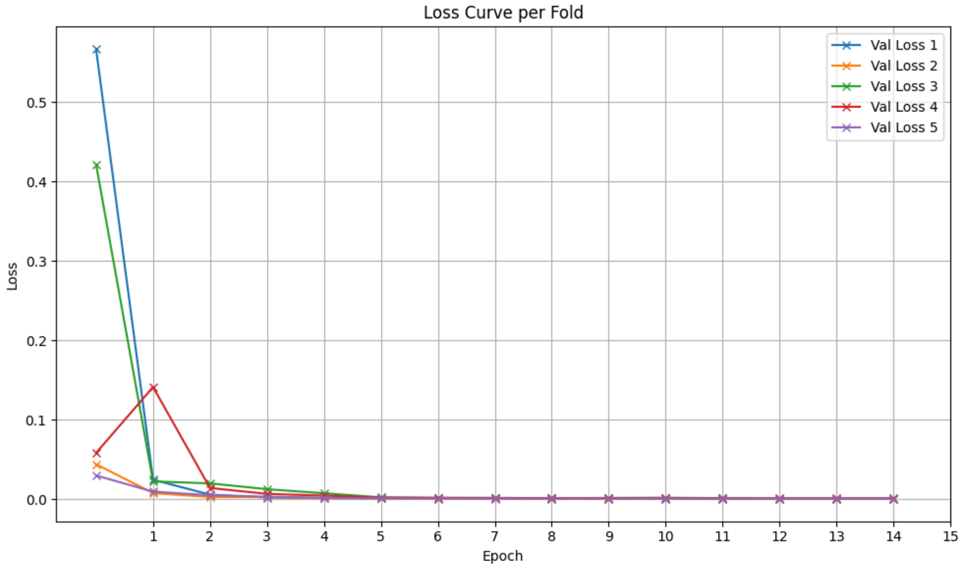

Figure 1: Loss curve for our ResNet Model

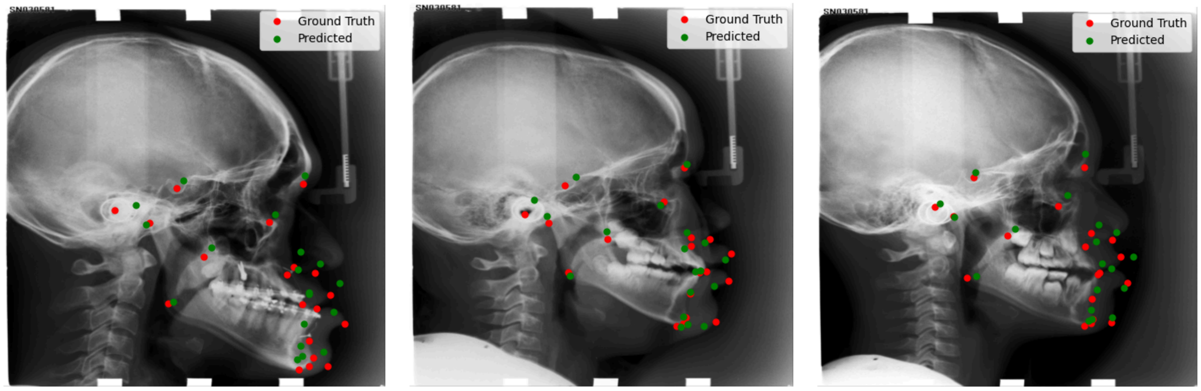

Figure 2: Example outputs from the model

Since the application area of this neural network is the medical field, it is very important that doctors are able to understand why the model is making certain decisions and whether they should trust it. To help provide this information behind the model's decision making, we incorporated saliency maps and Local Interpretable Model-Agnostic Explanations (LIME) when our model performed inference on an image. Saliency maps identify which pixels in the image contributed the most to the model's prediction based on its gradients [1]. On the other hand, LIME perturbes the input image to create a simple interpretable model that approximates the neural network and demonstrates which clusters of pixels are most important for the model's prediction [2]. Figure 3 shows an example output of a saliency map and LIME explanations for one of the model's predictions. In the saliency map, red values indicate that pixels were more important for the model's prediction. In the LIME output, green areas are ones that contributed positively to the model's certainty in its prediction while red areas contributed negatively.

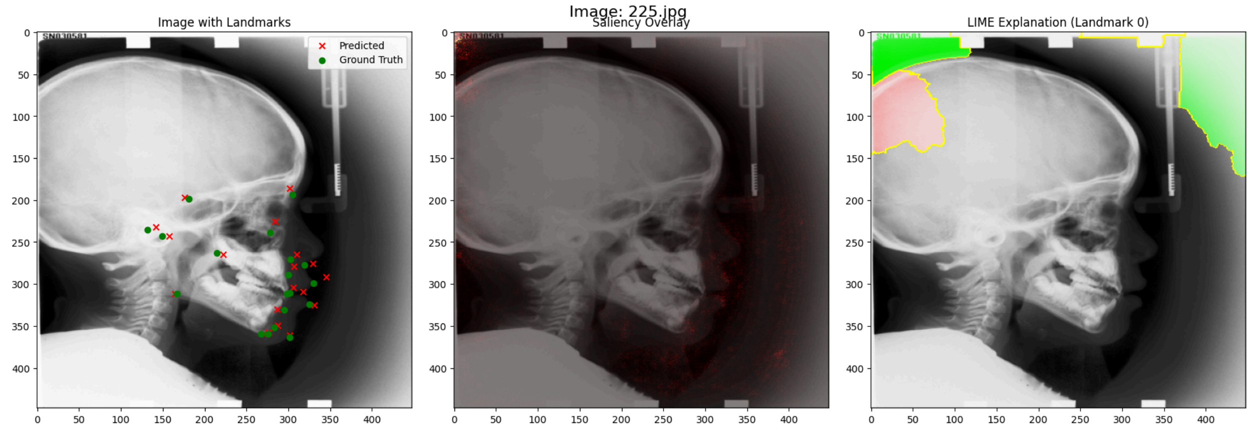

Figure 3: Example of a saliency map and LIME explanations for a prediction

Demo

A demo video for our code can be found here: https://drive.google.com/file/d/1HI4tz4ZCzJxkUnxmwfDtRBnQemzq0YRJ/view?usp=drive_link

Assumptions, Constraints & Implications

- Assumptions

Landmark definitions and annotation standards are consistent across the dataset and clinical practice. Since the landmark annotations were done by specialist orthodontists with good intra/inter operator reliability, it could be trusted.

- Constraints

Limited dataset size (400 images) may restrict the model's ability to fully capture anatomical variability across different populations. A multi-center trial could solve this constraint by aptly capturing varied population features.

The model relies on the accuracy of ground truth annotations. Inaccurate landmarks in the dataset would lead to prediction errors. Validation of annotation accuracy could be very taxing in terms of expert human resource and man hours.

This model was trained on grayscale X-ray images. The model would require retraining if applied on colorful medical images.

- Implications

This model significantly decreases the time required for manual landmark identification by doctors, enabling them to focus on diagnosis and treatment planning.

The model's dependency on standardized annotations underscores the need for a harmonized landmark definition across imaging protocols and further validation against gold standard human measurements.

CNNs could be trained on geometric features for pixel level precision for robust anatomical landmark identification.

Methods (How the solution was built)

- Data

We used a dataset for cephalometric x-rays, which includes 400 images and their associated 19 landmarks, from Kaggle [3]. The current dataset was split into an 80% training set and a 20% testing set. Each image was read in grayscale, as there was no color information present.

- Feature learning

To encourage model generalization across variations in pose and anatomy, we performed data augmentation during training. Techniques included random horizontal flipping and resizing all images to a consistent resolution. Landmark coordinates were normalized relative to image dimensions to maintain scale invariance.

To learn features in the data, we used a ResNet18 deep learning model, which consists of 18 total layers of which 17 are convolutional neural networks (CNNs) [4]. We used starting weights for ResNet18 that have been pre-trained on 1000 images from the ImageNet dataset [5].

- Model

We implemented a CNN-based regression model using ResNet18, where the final fully connected layer outputs 38 values (19 landmarks × 2 coordinates). The model was trained with 5-fold cross-validation, which allows for using multiple train-test splits to reduce any impact caused by randomly splitting the data. The loss function used is Mean Squared Error (MSELoss), and model performance is assessed based on the Root Mean Squared Error (RMSE) averaged across all landmarks per image. The Adam optimizer is used with an initial learning rate of 0.001.

Summary

Accurate identification of craniofacial landmarks in lateral cephalometric X-rays remains critical yet time-consuming for orthodontic diagnosis, requiring sub-millimeter precision. This work presents an automated system using a modified ResNet18 architecture pre-trained on ImageNet, achieving a validation loss of 0.0005 MSE (equivalent to ~0.0002 mm error at 72 DPI). The model outperforms custom CNNs, demonstrating transfer learning's value in medical imaging.

The key components for this work being a ResNet18 architecture adapted for coordinate regression (19 landmarks). Augmented training with flipping/resizing validated via a 5-fold cross validation. Furthermore integration of saliency maps and LIME to clarify predictions for clinicians.

A few limitations/issues are that firstly, 400 images may insufficiently capture anatomical and population variability. Second, the performance dependency on annotation consistency where human error can reach 3-4mm. Lastly, generalization is difficult since the model remains untested on multi-center data or varying imaging equipment.

References

[1] Z. Keita, “Explainable AI, Lime & Shap for model interpretability: Unlocking AI’s decision-making,” DataCamp, https://www.datacamp.com/tutorial/explainable-ai-understanding-and-trusting-machine-learning-models (accessed Apr. 28, 2025).

[2] B. Subhash, “Explainable AI: Saliency maps,” Medium, https://medium.com/@bijil.subhash/explainable-ai-saliency-maps-89098e230100 (accessed Apr. 28, 2025).

[3] Jiahong, “Cephalometric landmarks,” Kaggle, https://www.kaggle.com/datasets/jiahongqian/cephalometric-landmarks, 2020.

[4] K. He, X. Zhang, S. Ren, and J. Sun, “Deep Residual Learning for Image Recognition,” Dec. 10, 2015, arXiv: arXiv:1512.03385. doi: 10.48550/arXiv.1512.03385.

[5] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, “ImageNet: A large-scale hierarchical image database,” in 2009 IEEE Conference on Computer Vision and Pattern Recognition, Jun. 2009, pp. 248–255. doi: 10.1109/CVPR.2009.5206848.