자동 뼈 연령 평가를 위한 비전 트랜스포머와 손 X선 영상 분석

정경희 [1], Sammy Yap Xiang Bang [1], Nguyen Duc Toan [1], 추현승 [2]
[1] 성균관대학교 수퍼인텔리전스학과
[2] 성균관대학교 전자전기컴퓨터공학과

datakira@g.skku.edu, sammyyap98@skku.edu, austin47@g.skku.edu, choo@skku.edu

Unleashing the Potential of Vision Transformer for Automated Bone Age Assessment in Hand X-rays

Kyunghee Jung [1], Sammy Yap Xiang Bang [1], Nguyen Duc Toan [1], Hyunseung Choo [2]
[1] Dept. of Superintelligence, Sungkyunkwan University
[2] Dept. of Electrical and Computer Engineering, Sungkyunkwan University

Abstract

Bone age assessment is a crucial task in pediatric radiology for assessing growth and development in children. In this paper, we explore the potential of Vision Transformer, a state-of-the-art deep learning model, for bone age assessment using X-ray images. We generate heatmap outputs using a pre-trained Vision Transformer model on a publicly available dataset of hand X-ray images and show that the model tends to focus on the overall hand and only the bone part of the image, indicating its potential for accurately identifying the regions of interest for bone age assessment without the need for pre-processing to remove background noise. We also suggest two methods for extracting the region of interest from the heatmap output. Our study suggests that Vision Transformer holds great potential for bone age assessment using X-ray images, as it can provide accurate and interpretable output that may assist radiologists in identifying potential abnormalities or areas of interest in the X-ray image.

1. Introduction

Bone age assessment is a crucial procedure in pediatric endocrinology that helps in determining the level of skeletal maturity of a child. Fig 1 shows the process of ossification in pediatric patients and demonstrates how the degree of ossification in certain bones can be used as a proxy for assessing the bone age of a child. As children grow and develop, their bones undergo a series of changes that can be visualized on X-ray images, making bone age assessment an important tool for assessing growth and development. Traditionally, this has been done by expert radiologists, who visually examine the X-rays and make subjective assessments. However, recent advancements in artificial intelligence (AI) have led to the development of automated systems that can accurately estimate bone age with a high level of precision. In particular, the use of deep learning models such as Vision Transformers (ViT) [1] has shown promising results in various image-based tasks, including medical image analysis. In this paper, we present the results of our experiment on applying Vision Transformers (ViT) for bone age assessment using X-ray images of the hand and wrist. Specifically, we focus on the heatmap output of the ViT, which provides a visualization of the regions in the input image that the model identifies as most important for making its prediction.

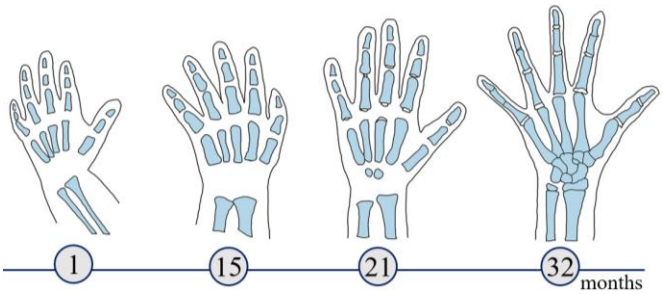

Fig 1. Ossification process of hand bones

2. Model

The ViT model consists of a sequence of self-attention layers, which allows the model to capture long-range dependencies in the input image. Unlike convolutional neural networks (CNNs), which use a fixed set of filters to extract features, the ViT model learns to attend to different parts of the image at different levels of abstraction. This makes it more adaptable to a wide range of image analysis tasks and enables it to learn complex patterns and features. In our proposed method, we use a pre-trained ViT model on a dataset of X-ray images of the hand and wrist. We use the ViT model to extract features from the input image, which are then used to predict the bone age of the patient. In

addition to the prediction, the ViT model also produces a heatmap output, which provides a visualization of the regions in the input image that the model identifies as most important for making its prediction. The heatmap output is particularly useful in medical image analysis tasks, as it provides an interpretable output that can help clinicians and radiologists identify potential abnormalities or areas of interest in the image. The heatmap can also be used as a diagnostic aid to identify potential developmental disorders or diseases that may affect bone growth.

After generating the heatmap output from the Vision Transformer model, the region of interest (ROI) for bone age assessment can be extracted using two different methods. The first method involves extracting only the specific part of the image that is attended by the heatmap output. This can be achieved by applying a threshold to the heatmap output and extracting the region that exceeds the threshold. This approach allows for precise extraction of the most important features of the image, as identified by the model. The second method involves cropping a rectangle box shape around the area of interest in the X-ray image. This method is less precise but can be more practical in cases where the heatmap output is not well defined or the attended region is too small to extract. Once the ROI is extracted using either method, it can be used for bone age assessment.

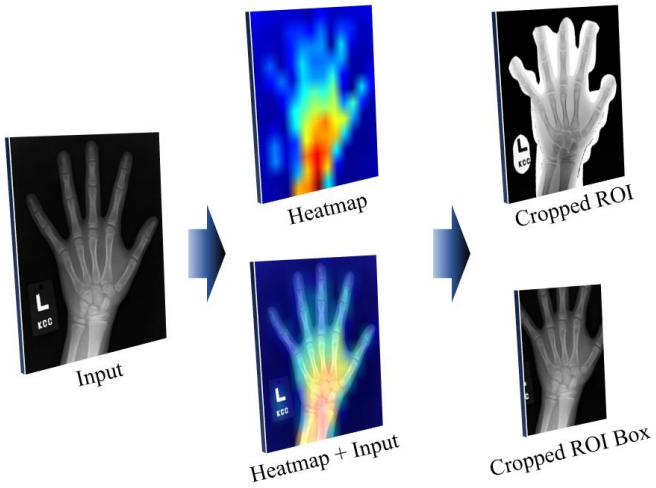

Figure 2. Two ROI cropping methods for bone age assessment

3. Result

We used RSNA hand bone X-ray public dataset [2]. It is a collection of hand X-ray images that was made publicly available by the Radiological Society of North America (RSNA) as part of its annual Machine Learning Challenge. The dataset contains over 14,000 hand X-ray images, each labeled with the corresponding bone age of the patient.

The results show that the Vision Transformer model tends to focus on the overall hand, especially wrist in X-ray images. Furthermore, the model appears to only focus on the bone part of the image while ignoring the background. These findings are significant as many previous studies have suggested the need for pre-processing X-ray images to remove background noise, which is typically considered to be irrelevant to the task of bone age assessment [3]. However, our results suggest that this may not be necessary when using Vision Transformers, as the model appears to inherently focus only on the bone and ignore the background. This could potentially save time and resources in the pre-processing step. Overall, our results suggest that this method can provide accurate and efficient bone age assessment results, while also providing an interpretable heatmap output that can assist radiologists in identifying potential abnormalities or areas of interest in the X-ray images.

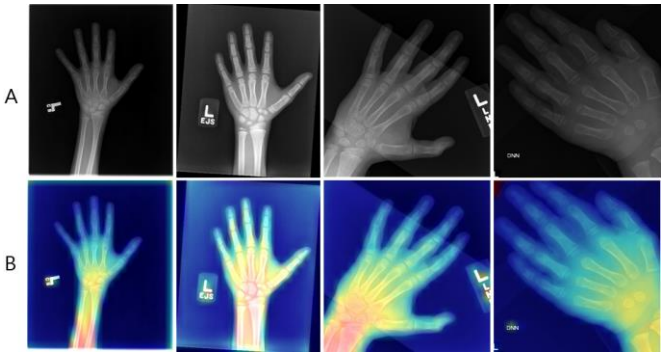

Figure 3. Input image (A) and generated heatmaps (B) using Vision Transformer for hand X-ray images

4. Conclusion

Our results show that the heatmap output of the ViT can accurately identify the bones and regions of interest in the X-ray images. Moreover, our experiments demonstrate that the heatmaps can be used as a diagnostic aid to identify potential abnormalities or developmental disorders in bone growth, which can assist radiologists in making more accurate and informed diagnoses. Overall, our study highlights the potential of using ViT for bone age assessment and suggests that the heatmap output can be a useful tool for radiologists and clinicians in this important diagnostic procedure. Further research is needed to evaluate the model's performance on a larger and more diverse dataset, as well as to explore its potential for other applications in medical image analysis.

Acknowledgement

This research was supported by the MSIT(Ministry of Science and ICT), Korea, under the ICT Creative Consilience Program(IITP-2023-2020-0-01821), High Potential Individuals Global Training Program) (RS-2022-00155415) (contribution rate:50%) supervised by the IITP(Institute for Information & communications Technology Planning & Evaluation) (No.2021-0-02068, Artificial Intelligence Innovation Hub)

Reference

[1] Dosovitskiy, Alexey, et al. "An image is worth 16x16 words: Transformers for image recognition at scale. arXiv 2020." arXiv preprint arXiv:2010.11929 (2010).

[2] Halabi, Safwan S., et al. "The RSNA pediatric bone age machine learning challenge." Radiology 290.2 (2019): 498-503.

[3] Jung, Kyunghee, et al. "Hand Bone X-rays Segmentation and Congregation for Age Assessment using Deep Learning." 2023 International Conference on Information Networking (ICOIN). IEEE, 2023.