physical sciences forum

MDPI

Proceeding Paper

Attention-Guided Multi-Scale CNN Network for Cervical Vertebral Maturation Assessment from Lateral Cephalometric Radiography †

Hamideh Manoochehri [1], Seyed Ahmad Motamedi [1,*], Ali Mohammad-Djafari [2,*], Masrour Makaremi [3] and Alireza Vafaie Sadr [4]

[1] Department of Electrical Engineering, Amirkabir University of Technology, Tehran 1591634311, Iran
[2] International Science Consulting and Training (ISCT), 91440 Bures-sur-Yvette, France
[3] Departement Dentofacial Orthopedics, UFR des Sciences Odontologiques, 146, Rue Léo-Saignat, CEDEX, 33076 Bordeaux, France
[4] Institute of Pathology, RWTH Aachen University Hospital, 52074 Aachen, Germany
* Correspondence: motamedi@aut.ac.ir (S.A.M.); djafari@free.fr (A.M.-D.)
† Presented at the 41st International Workshop on Bayesian Inference and Maximum Entropy Methods in Science and Engineering, Paris, France, 18–22 July 2022.

Abstract: Accurate determination of skeletal maturation indicators is crucial in the orthodontic process. Chronologic age is not a reliable skeletal maturation indicator, thus physicians use bone age. In orthodontics, the treatment timing depends on Cervical Vertebral Maturation (CVM) assessment. Determination of CVM degree remains challenging due to the limited annotated dataset, the existence of significant irrelevant areas in the image, the huge intra-class variances, and the high degree of inter-class similarities. To address this problem, researchers have started looking for external information beyond current available medical datasets. This work utilizes the domain knowledge from radiologists to train neural network models that can be utilized as a decision support system. We proposed a novel supervised learning method with a multi-scale attention mechanism, and we incorporated the general diagnostic patterns of medical doctors to classify lateral X-ray images as six CVM classes. The proposed network highlights the important regions, surpasses the irrelevant part of the image, and efficiently models long-range dependencies. Employing the attention mechanism improves both the performance and interpretability. In this work, we used additive spatial and channel attention modules. Our proposed network consists of three branches. The first branch extracts local features, and creates attention maps and related masks, the second branch uses the masks to extract discriminative features for classification, and the third branch fuses local and global features. The result shows that the proposed method can represent more discriminative features, therefore, the accuracy of image classification is greater in comparison to in backbone and some attention-based state-of-the-art networks.

Keywords: machine learning; deep learning; attention mechanism; convolutional neural network; cervical vertebra maturation; supervised learning

Citation: Manoochehri, H.; Motamedi, S.A.; Mohammad-Djafari, A.; Makaremi, M.; Vafaie Sadr, A. Attention-Guided Multi-Scale CNN Network for Cervical Vertebral Maturation Assessment from Lateral Cephalometric Radiography. *Phys. Sci. Forum* **2022**, *5*, 26. https://doi.org/10.3390/psf2022005026

Academic Editors: Frédéric Barbaresco, Frank Nielsen and Martino Trassinelli

Published: 12 December 2022

1. Introduction

Accurate determination of skeletal maturation indicators is crucial. As chronologic age is not a reliable indicator for skeletal maturation, physicians use bone age as an indicator. Generally, bone age assessment using the classical radiographic manual methods is performed in two main ways: the hand-wrist radiograph method (HWM) [1,2], and cervical vertebral maturation (CVM) degree [3]. The first method has been used as a gold standard in the assessment of skeletal maturation for many decades, but presents several issues such as the additional X-ray exposure, the time spent and experience required, and the sexual dimorphism and ethnic polymorphism in morphological modifications. Since

cephalometric radiography usually is used in orthodontic processes, by using the second method, the radiation dose can be reduced and the cost and time can be decreased.

CVM stages can be estimated by the morphological description of vertebrae spines (C2, C3, and C4). CVM stages have been described into six stages correlating with morphological modifications of the vertebral shapes and estimated time lapse from the mandibular growth peak. Manual analysis is time-consuming and demanding for expert graders, which is also prone to yield subjective results. Consequently, an automatic and reliable CVM classification is required for efficient diagnosis. Automatic CVM stage estimation can decrease diagnosis time and treatment cost.

In the medical field there are more challenges due to three main reasons: 1—In comparison to popular natural image datasets such as ImageNet, the medical image dataset's size is too small. This problem can cause overfitting. 2—Medical images are noisy, their boundaries are ambiguous, and ROI is located in a small part of the image. 3—The same body organs have a variety of anatomical shapes.

We incorporate domain knowledge and attention mechanisms to solve the above problems. Our proposed network simulates a radiologist's diagnosis that focuses on specific local regions when analyzing the lateral cephalometric radiographs. Radiologists generally follow a three-staged approach when they read X-ray images: they first browse the whole image, then concentrate on the local lesion areas, and finally combine the global and local information to make decisions. This pattern is incorporated into the architectural design of our network. Specifically, we first exploit a global branch to make a mask for ROI detection. This mask is a soft attention map from the input image, and then the created mask is multiplied with the input image in the local branch. We zoom into the most discriminative region with a higher resolution. Then, the obtained local image is applied to the local branch for extracting more fine-grained features for CVM classification. Finally global and local feature maps are integrated directly into the final classification layer to output more accurate prediction.

The main contributions of the proposed method are summarized as follows:

(1) We propose a multi-scale attention-based CNN network for CVM classification. To the best of our knowledge, this is the first time that an attention model has been introduced in the field of CVM analysis.
(2) A novel spatial attention module is proposed to learn the spatial interdependencies of features and a channel attention module is designed to model channel interdependencies. They significantly improve the classification results by modeling rich contextual dependencies over local and global features.
(3) We achieved new state-of-the-art results on our lateral cephalometric radiology dataset.

2. Related Works

To the best of our knowledge, there are a few works on CVM classification, which we discuss in this section in addition to introducing the methods applied to the CVM classification.

2.1. CVM Classification Methods

Some researchers [4–7] used classical machine learning methods and hand-crafted features for CVM analysis, while some other researchers utilized deep learning methods.

2.1.1. Classical Machine Learning Methods for CVM Stage Classification

In [4], nineteen reference points were defined on the second, third, and fourth cervical vertebrae, and twenty different linear measurements were taken. Seven artificial intelligence algorithms frequently used in the field of classification were selected and compared. These algorithms are k-nearest neighbors (k-NN), Naive Bayes (NB), decision tree (Tree), artificial neural networks (ANN), support vector machine (SVM), random forest (RF), and logistic regression (LR) algorithms. According to the confusion matrices, decision tree: CSV1

(97.1%)—CSV2 (90.5%), SVM: CVS3 (73.2%)—CVS4(58.5%), and KNN: CVS5 (60.9%)—CVS6 (78.7%) were the algorithms with the highest accuracy in determining cervical vertebrae stages. The ANN algorithm was observed to have the second highest accuracy values (93%, 89.7%, 68.8%, 55.6%, and 78%, respectively) in determining all stages except CVS5 (47.4%: third highest accuracy value). According to the average rank of the algorithms in predicting the CSV classes, ANN was the most stable algorithm with its 2.17 average rank.

In [5], 54 extracted features from 24 points were defined on second, third, fourth, and fifth cervical vertebrae, the 5 classical frequently used ML algorithms (artificial neural network (ANN), logistic regression (LR), decision tree (DT), random forest (RF), and support vector machine (SVM)) were used, and among the CVM stage classifier models, the best result was achieved using the artificial neural network model ($\kappa = 0.926$). Among cervical vertebrae morphology classifier models, the best result was achieved using the logistic regression model ($\kappa = 0.968$) for the presence of concavity, and the decision tree model ($\kappa = 0.949$) for vertebral body shapes.

2.1.2. Deep Learning Methods for CVM Stage Classification

Makaremi used a convolution deep neural network and different pre-processing filters for CVM stage classification and achieved high accuracy [7].

Kök utilized transfer learning techniques for six different pre-trained network architectures and compared the results. The results showed that all deep learning models demonstrated more than 90% accuracy, with Inception-ResNet-v2 performing the best, relatively. In addition, visualizing each deep learning model using Grad-CAM led to a primary focus on the cervical vertebrae and surrounding structures [8].

Kim proposed a stepwise segmentation-based model that focuses on the C2–C4 regions. They proposed three convolutional neural network-based classification models: a one-step model with only CVM classification, a two-step model with a region of interest (ROI) detection and CVM classification, and a three-step model with ROI detection, cervical segmentation, and CVM classification. Our dataset contains 600 lateral cephalogram images, comprising 6 classes with 100 images each. The three-step segmentation-based model produced the best accuracy (62.5%) compared to the models that were not segmentation-based [9].

3. Materials and Methods

3.1. Basic Idea

The aim is a classification of lateral cephalometric images into 6 classes. Our dataset is too small and ROI (C2, C3, and C4) are located in the small portion of images. To address these issues, we exploit domain knowledge. The proposed framework is shown in Figure 1. We used the multi-scale attention mechanism to highlight the important features, surpass the irrelevant part of the image, and efficiently model long-range feature dependencies. The attention mechanism improves both the performance and interpretability of visual tasks, including image classification. In this work, we used spatial and channel soft attention. Our proposed network consists of three branches. The first branch extracts global features, and creates attention maps and related masks, the second branch uses the mask to extract discriminate local features for classification, and the third branch fuses local and global features. The network's backbone is DensNet169. We used spatial and channel attention in different backbones, and compared them with attention blocks and without attention blocks.

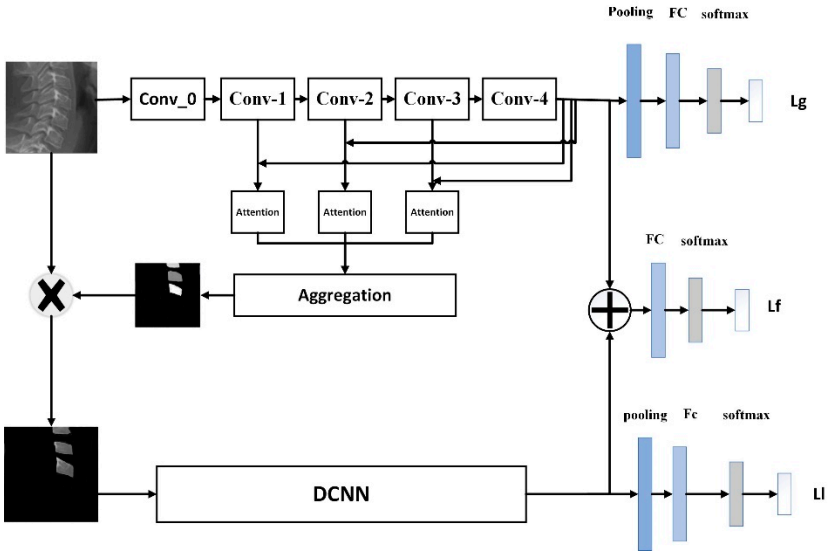

Figure 1. Overview of proposed method.

3.2. Global Branch

When the radiologist looks at the medical images, they first find the important region of the image. Inspired by this, we first used a global branch to extract a relevant mask from the input image.

The relevant mask multiplies to the original image to highlight the important spins and suppress the irrelevant part of the image. In this branch, features at multiple scales are used. Since features come at different resolutions for each level, they are upsampled to a common resolution by employing bilinear interpolation, leading to enlarged feature maps. Then, the feature maps from the last den's block are all concatenated with all scales and create a new feature map. This feature map encodes low-level detail information from shallow layers as well as high-level semantics learned in deeper layers. Then, these feature maps are fed to the attention block.

3.3. Attention Mechanism

We used a soft attention block that contains spatial and channel attention that focuses on modeling position (where to pay attention) and channel feature (what to pay attention) dependencies, respectively. There are two commonly used attention types: multiplicative and additive attention. Although, multiplicative attention can be implemented as matrix multiplication, therefore it is faster and memory-efficient. In this work we used an additive attention module (Figure 2) because it has been experimentally shown to perform better for large dimensional input features [10].

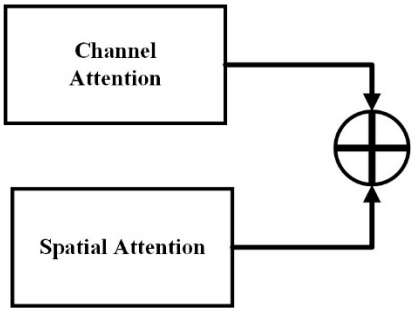

Figure 2. The proposed attention module.

We used the multi-scale approach that combines different semantics from different layers, and thus considers long-range dependencies. In the classic CNN networks, pooling

layers increase the receptive field, but remove many details. In this approach we used four outputs of dense blocks and concatenated them with the last layer to consider both low/middle feature maps with high-level feature maps. Furthermore, we fed each combination directly to the attention module to find the appropriate mask. While lower-level layers focus on local representations, higher-level layers encode global representations. This multi-scale strategy encourages attention maps generated at different layers to represent different semantic information.

3.3.1. Channel Attention (CA)

Channel maps can be considered as class-specific responses Thus, another strategy to enhance the feature representation of specific semantics is to improve the dependencies between channel maps [11]. The CA will assign larger weight to channels that show high response to salient objects and determine what to pay attention to our novel channel attention network is depicted in Figure 3.

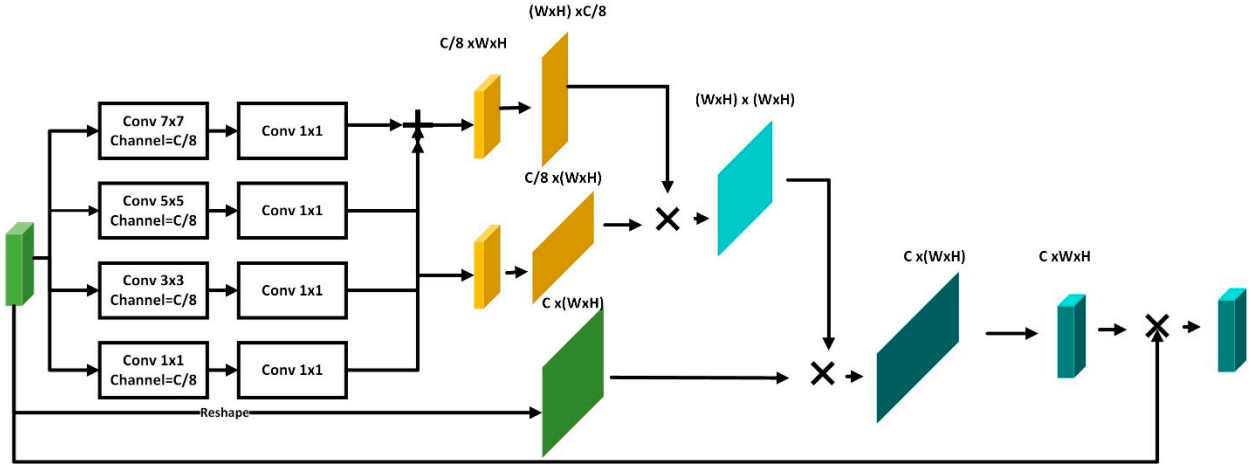

Figure 3. Channel attention module.

The proposed channel attention module consists of two parts. The aim of CA is to extract informative features to weight channels according to the importance of them in a specific class. The first part of the module is a pyramid model that is used to weigh multi-scale and multi-receptive field features. The second part captures the relationship between channels and assigns a larger weight to channels that show a high response to salient objects.

3.3.2. Spatial Attention

The spatial attention module highlights the important region of the image. The proposed spatial attention module is shown in Figure 4. We applied two convolution layers, one's kernel is $1 \times k$ and the other's is $k \times 1$, to increase receptive field and represent global features without increasing parameters. We used a correlation matrix that represents the relationship of features in any two positions, with similar features being able to contribute to mutual improvement regardless of their distance in the spatial dimension. We proposed a multi-scale attention module that contains a lot of details from low-level features and global information from high-level feature maps.

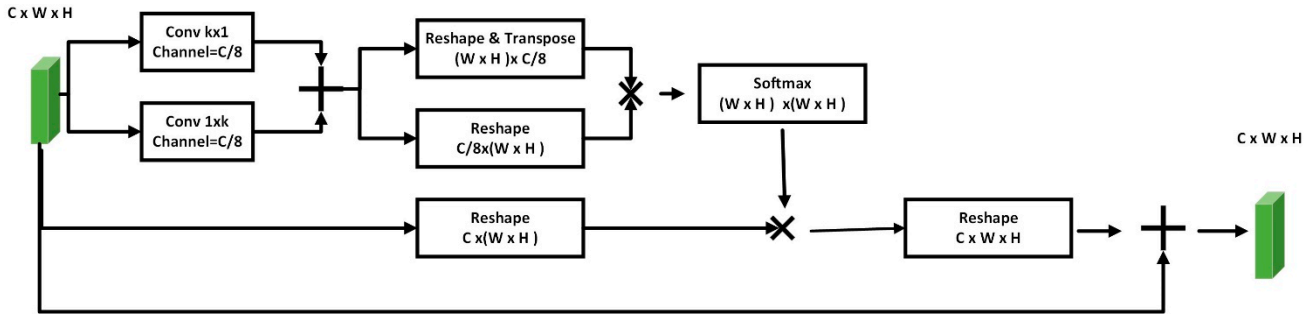

Figure 4. Spatial attention module.

3.3.3. Local and Fusion Branch

The created salient mask is multiplied with the original image to highlight the salient region and suppress the irrelevant part of image. This branch is a pre-trained Densnet169 architecture.

The fusion branch is an ensemble model that aggregates local and corresponding global features to extract discriminative features and improve the classification accuracy.

4. Experiments

4.1. Dataset

According to Table 1, our dataset consists of 1870 grayscale X-ray images of lateral cephalograms that were clinically acquired.

Table 1. Lateral cephalometric image dataset.

Class Name	Number of Images
CVS1	199
CVS2	184
CVS3	825
CVS4	300
CVS5	200
CVS6	162
Total number	1870

4.2. Implementation Details

We employed a pretrained DensNet169 network as the backbone. The result shows that the proposed method can represent more discriminative features, therefore the accuracy of image classification can be increased. We implemented our method based on Pytorch. We trained all the networks using Adam optimizer with a mini-batch of size 8, and with $\beta 1$ and $\beta 2$ set to 0.9 and 0.99, respectively. Most of the networks converged during the first 250 epochs. The learning rate was initially set to 0.001 and multiplied by 0.5 after 50 epochs, without improvement on the validation set. The optimal values of these parameters were found empirically.

In all cases, the multi-class cross-entropy between the network prediction and the ground truth labels was employed as classification loss. The final objective function to optimize becomes:

$$L_{Total} = \alpha L_{Global} + \beta L_{Local} + \gamma L_{fusion}$$

where α, β, and γ control the importance of each term in the main loss function.

4.3. Results

We performed an ablation experiment under different settings to validate the individual contribution of different components to the CVM stage classification performance. Compared to the baseline (i.e., transfer learning with DensNet169), we observed that by integrating either a spatial (SAM) or channel attention module (CAM) at each scale in the baseline architecture, the performance improved between 3–5% in terms of accuracy and 2–4% in terms of F1 score.

The provision of salient maps helps the model to focus on target regions with high saliency and suppress irrelevant regions, leading to higher accuracy. In the proposed network, by using pyramid architecture, long-range feature dependencies are extracted that represent global features efficiently.

5. Conclusions

In this paper, we showed that the combination of the domain knowledge, mimicking of radiologists' behavior in CVM stage classification, and deep neural networks can improve the network accuracy. In particular, we utilized a novel multi-scale soft attention module to combine semantic information and long-range dependencies at different levels and create a mask for useful regions. In the second branch of network for ROI selection, we do not resize the image, and thus avoid detail information loss in ROI. To validate our approach, we conducted experiments on our dataset and compared the results with some state-of-the-art methods. Experiment results showed that the proposed model outperformed all previous approaches.

Author Contributions: Conceptualization, H.M.; methodology, H.M.; software, H.M.; validation, H.M.; formal analysis, H.M.; investigation, H.M.; resources, M.M.; data curation, M.M.; writing original draft preparation, H.M.; writing review and editing, A.V.S.; visualization, H.M.; supervision, S.A.M. and A.M.-D.; project administration, S.A.M.; project advisor, A.M.-D. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: The study did not require ethical approval.

Informed Consent Statement: Not applicable.

Data Availability Statement: The study did not report any data.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Krisztina, M.I.; Ogodescu, A.; Réka, G.; Zsuzsa, B. Evaluation of the Skeletal Maturation Using Lower First Premolar Mineralisation. *Acta Med. Marisiensis* **2013**, *59*, 289–292. [CrossRef]
2. Pyle, S.I.; Waterhouse, A.M.; Greulich, W.W. Attributes of the radiographic standard of reference for the National Health Examination Survey. *Am. J. Phys. Anthropol.* **1971**, *35*, 331–337. [CrossRef] [PubMed]
3. Hassel, B.; Farman, A.G. Skeletal maturation evaluation using cervical vertebrae. *Am. J. Orthod. Dentofac. Orthop.* **1995**, *107*, 58–66. [CrossRef] [PubMed]
4. Seo, H.; Hwang, J.; Jeong, T.; Shin, J. Comparison of Deep Learning Models for Cervical Vertebral Maturation Stage Classification on Lateral Cephalometric Radiographs. *J. Clin. Med.* **2021**, *10*, 3591. [CrossRef] [PubMed]
5. Amasya, H.; Yildirim, D.; Aydogan, T.; Kemaloglu, N.; Orhan, K. Cervical vertebral maturation assessment on lateral cephalometric radiographs using artificial intelligence: Comparison of machine learning classifier models. *Dentomaxillofac. Radiol.* **2020**, *49*, 20190441. [CrossRef] [PubMed]
6. Baptista, R.S.; Quaglio, C.L.; Mourad, L.M.; Hummel, A.D.; Caetano CA, C.; Ortolani CL, F.; Pisa, I.T. A semi-automated method for bone age assessment using cervical vertebral maturation. *Angle* **2012**, *82*, 658–662. [CrossRef] [PubMed]
7. Makaremi, M.; Lacaule, C.; Mohammad-Djafari, A. Deep learning and artificial intelligence for the determination of the cervical vertebra maturation degree from lateral radiography. *Entropy* **2019**, *21*, 1222. [CrossRef]
8. Kök, H.; Acilar, A.M.; İzgi, M.S. Usage and comparison of artificial intelligence algorithms for determination of growth and development by cervical vertebrae stages in orthodontics. *Prog. Orthod.* **2019**, *20*, 1–10. [CrossRef] [PubMed]
9. Kim, E.G.; Oh, I.S.; So, J.E.; Kang, J.; Le VN, T.; Tak, M.K.; Lee, D.W. Estimating Cervical Vertebral Maturation with a Lateral Cephalogram Using the Convolutional Neural Network. *J. Clin. Med.* **2021**, *10*, 5400. [CrossRef] [PubMed]

10. Chen, L.; Zhang, H.W.; Xiao, J.; Nie, L.Q.; Shao, J.; Liu, W.; Chua, T. SCA-CNN: Spatial and channelwise attention in convolutional networks for image captioning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 6298–6306.
11. Peng, C.; Zhang, X.; Yu, G.; Luo, G.; Sun, J. Large kernel matters–improve semantic segmentation by global convolutional network. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017.