Article

Attention guided multi-scale CNN Network for Cervical Vertebral Maturation Assessment from Lateral Cephalometric Radiography

Hamideh Manoochehri [1], Seyed Ahmad Motamedi [2,]* Ali Mohammad-Djafari [3,]*Masrour Makaremi[4] and Alireza Vafaie sadr[5]

[1] Department of Electrical Engineering, Amirkabir University of Technology, Tehran, Iran; h_manoochehri@aut.ac.ir
[2] Department of Electrical Engineering, Amirkabir University of Technology, Tehran, Iran; motamedi@aut.ac.ir
[3] International Science Consulting and Training (ISCT), 91440 Bures-sur-Yvette, France; djafari@free.fr
[4] Departement dentofacial orthopedics, UFR des Sciences Odontologiques, 146, rue Le´o-Saignat, 33076 Bordeaux cedex, France; masrour@makaremi.fr
[5] Institute of Pathology, RWTH Aachen University Hospital, Aachen, Germany; asadr@ukaachen.de

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 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 create networks that resemble how medical doctors are trained, mimic their diagnostic patterns, or focus on the features or areas they pay particular attention to. We proposed a novel supervised learning method with multi-scale attention mechanism and also, we incorporated the general diagnostic patterns of medical doctors to classify lateral x-ray images as six cervical vertebrae maturation (CVM) classes. The proposed network highlights the important features, surpasses the irrelevant part of the image and efficiently models long-range feature dependencies. attention mechanism improves both their performance and interpretability in visual tasks including image classification. In this work, we used additive spatial and channel attention modules. Our proposed network consists of three branches. The first branch extracts local features, creates attention maps and related mask, the second branch uses this mask to extract discriminate features for classification. And the third branch fuses local and global features.

The result shows that proposed method can represent more discriminative features therefore the accuracy of image classification in compare backbone and some attention-based state of art networks

Keywords: Machine Learning; Deep learning; attention mechanism; convolutional neural network; cervical vertebra maturation; supervised learning

1. Introduction

Accurate determination of skeletal maturation indicator is crucial. As Chronologic age is not a reliable indicator for skeletal maturation, physicians use bone age indicator. Generally, bone age assessment in the classical radiographic manual methods is done in two main ways: Hand-wrist radiograph method (HWM)[1-2] and cervical vertebra maturation (CVM) degree[3]. The first method has been used as a gold standard in the assessment of skeletal maturation for many decades, but presented several issues as: the additional x-ray exposure, the time spending and experience required and a 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 morphological description of vertebrae spines (C2, C3 and C4). CVM stages have been described into 6 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 stages estimation can decrease diagnosis time and treatment cost.

Deep learning methods have attracted much attention in both industrial and educational fields and are used for many medical image processing, clustering and classification. Recent advances in computer vision and neural networks have demonstrated that automatic feature learning using deep neural networks are more successful than hand-engineered features. Handcrafted features are not generalizable and often fail to capture the extensive structural diversity found in images. Specifically, convolutional neural networks (CNN) have been extensively used to produce state-of-the-art results in different computer vision and pattern recognition problems. Image classification deep learning-based methods have achieved state of art results with natural images. But in the medical field there are more challenges that are due to three main reasons: 1- in comparison with popular natural image datasets like ImageNet, medical image datasets size is too small. This problem can cause overfitting. 2- medical images are noisy, their boundaries are ambiguous and ROI is located in the small part of the image. 3- The same body organs have a variety of anatomical shapes.

To solve the above problems, we incorporate domain knowledge and utilize attention mechanisms. Our proposed network, simulates 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 chest X-ray images: first browsing the whole image, then concentrating on the local lesion areas, and finally combining the global and local information to make decisions. This pattern is incorporated in the architecture 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 be multiped with the input image in the local branch. we zoom in 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 and output more accurate prediction.

Neural networks are generally applied as opaque black box models and often the network's decisions are difficult to interpret. Making the decision process transparent, and hence reliable is important for a computer-assisted diagnosis (CAD) system. Moreover, it is crucial that the network's decision be based on morphological features that are in agreement with a human expert. attention mechanism improves both their performance and interpretability in visual tasks including image classification. For this purpose, our proposed method used spatial and channel soft attention.

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 is introduced in 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. It significantly improves the classification results by modeling rich contextual dependencies over local and global features.
(3) We achieve 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, in this section in addition to introducing the methods applied to the CVM classification, we introduce attention mechanisms.

2.1 CVM classification methods:

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

2.1.1 Classical machine learning methods for CVM stage classification

In [4] nineteen reference points were defined on second, third, and 4th cervical vertebrae, and 20 different linear measurements were taken. Seven algorithms of artificial intelligence that are 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 confusion matrices decision tree, CSV1 (97.1%)–CSV2 (90.5%), SVM: CVS3 (73.2%)–CVS4(58.5%), and KNN: CVS 5 (60.9%)–CVS 6 (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] extracts 54 features from 24 points were defined on second, third, 4th and 5th cervical vertebrae, the five classical frequently used ML algorithms (artificial neural network (ANN), logistic regression (LR), decision tree (DT), random forest (RF), support vector machine (SVM)) are used and among the CVM stage classifier models, the best result was achieved using the artificial neural network model (κ = 0.926). Among cervical vertebrae morphology classifier models, the best result was achieved using the logistic regression model (κ = 0.968) for the presence of concavity, and the decision tree model (κ = 0.949) for vertebral body shapes.

2.1.2 Deep learning methods for CVM stage classification

[7] used a convolution deep neural network and different preprocessing filters for CVM stage classification and achieved high accuracy results.

[8] utilized transfer learning techniques for six different pre-trained network architecture and compared the results. The results show 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.

[9] propose a stepwise segmentation-based model that focuses on the C2–C4 regions. They propose three convolutional neural network-based classification models: a one-step model with only CVM classification, a two-step model with 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 six 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.

2.2 Attention mechanism

It is well known that attention plays an important role in human perception. Naturally, Humans can and effectively find salient regions in complex scenes. One important property of a human visual system is that one does not attempt to process a whole scene at once. Instead, humans exploit a sequence of partial glimpses and selectively focus on salient parts in order to capture visual structure. Better attention mechanisms were inspired by this observation and introduced into computer vision with the aim of imitating this aspect of the human visual system. attention mechanism can be regarded as a dynamic weight adjustment process based on features of the input image. Attention mechanisms have achieved great success in many visual tasks, including image classification, object detection, semantic segmentation, video understanding, image generation, 3D vision, multimodal tasks, and self-supervised learning. Attention mechanisms highlight the most important regions of an image and disregard irrelevant parts of the image. Attention not only tells where to focus, it also improves the representation of interests. Existing attention methods, in image classification include three basic categories: channel attention (what to pay attention to [10]), spatial attention (where to pay attention) and branch channel (which to pay attention to), along with one hybrid combined categories: channel & spatial attention.

Attention in deep neural networks is traditionally implemented in two main forms known as hard and soft attention. The implementation of hard (or stochastic) attention is nondifferentiable, the training procedure is based on a sampling technique, and as a consequence, the models are difficult to optimize [11]. Soft (or deterministic) attention models are differentiable and trained with backpropagation; because of these properties, they have been the preferred form of implementation. This methods are more recently applied to image analysis [12-]

[12] utilized soft attention mechanism for medical image segmentation. AG_CNN model [13] proposed for medical image classification and segmentation and automatically learns to focus on target structures of varying shapes and sizes. This model improves model sensitivity and accuracy for global and dense label predictions.

[14] first uses a low-capacity, yet memory efficient, network on the whole image to identify the most informative regions. It then applies another higher-capacity network to collect details from chosen regions. Finally, it employs a fusion module that aggregates global and local information to make a prediction.

2. Materials and Methods

2.1 Basic idea

The aim is classification of lateral cephalometric images into 6 classes. Our dataset is too small and ROI (C2, C3 and C4) 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. attention mechanism improves both their performance and interpretability in 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, creates attention maps and related mask, 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 use spatial and channel attention in different backbones. And compare them with attention and without attention blocks.

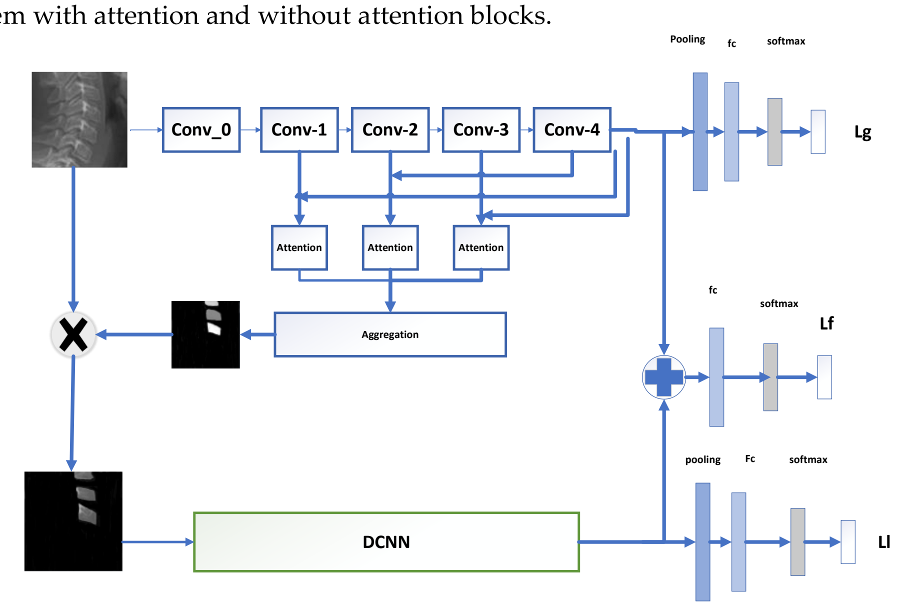

Figure 1 overview of proposed method

2.2 Global branch

As the proposed classification framework resembles the diagnostic procedure of radiologist, we first use a global branch to extract a relevant mask Mg from input image x, i.e., we compute:

$M_g = F_g(x)$

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 denoted as Fs, where s indicates the level in the architecture (Fig.1). Since features come at different resolutions for each level s, they are upsampled to a common resolution by employing bilinear interpolation, leading to enlarged feature maps Fs. Then, F's from the last dens block is all concatenated with all scales and create Fcs feature map. Fcs encodes low-level detail information from shallow layers as well as high-level semantics learned in deeper layers. then Fcs feature maps are fed to the attention block.

2.3 attention mechanism

We use a soft attention block that contains spatial and channel attention that focus on modelling position and channel feature dependencies, respectively. There are two commonly used attention types: Multiplicative and additive attention. The former is faster to compute and more memory-efficient in practice since it can be implemented as a matrix multiplication. However, additive attention is experimentally shown to be performing better for large dimensional input features [16]. In this work we use additive attention module (Fig. 2)

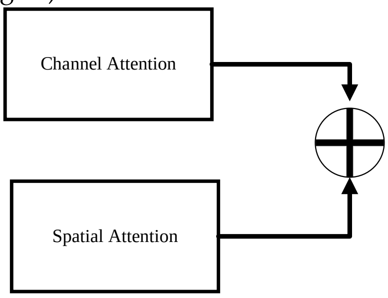

Figure 2 the proposed attention module

We use the multi-scale approach that generates stacks at different resolutions containing different semantics. While lower-level stacks focus on local appearance, higher-level stacks will encode global representations. This multi-scale strategy encourages that attention maps generated at different resolutions encode different semantic information. Then, at each scale, a stack of attention modules will gradually remove noisy areas and emphasize those regions that are more relevant to the semantic descriptions of the targets.

2.3.1 channel attention(CA)

channel maps can be considered as class-specific responses, where different semantic responses are associated with each other. Thus, another strategy to enhance the feature representation of specific semantics is to improve the dependencies between channel maps [17]. The CA will assign larger weight to channels which show high response to salient objects and determine what to pay attention in our novel channel attention network is depicted in Figure 3.

The proposed channel attention module consists of two parts. The first part is a pyramid model that used to weighted multi-scale multi-receptive field features. The second part captures relationship between channels and assigns larger weight to channels which show high response to salient objects.

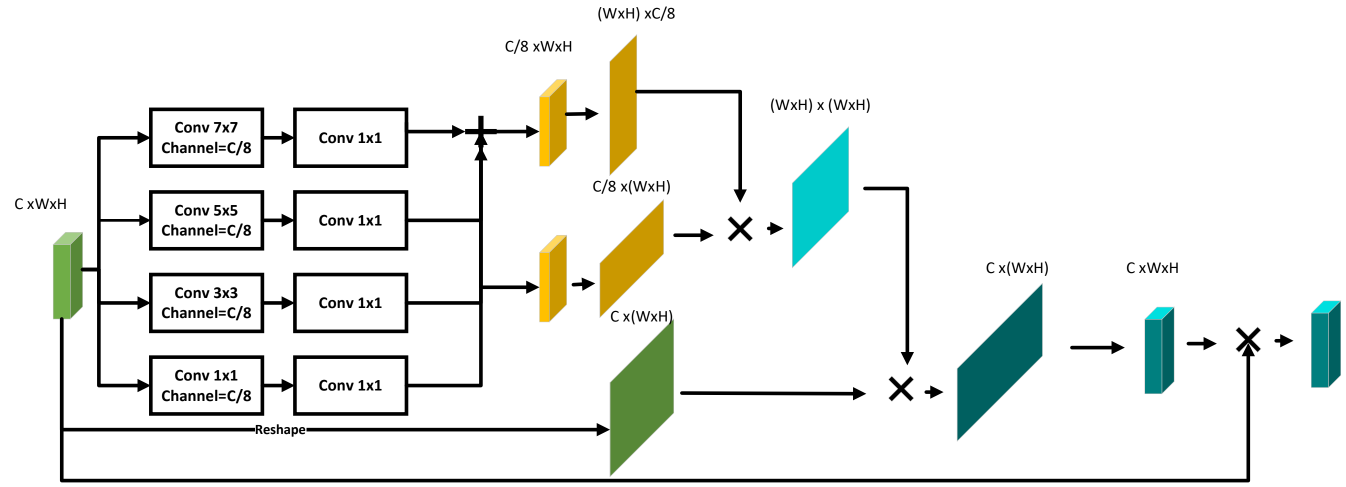

Figure 3 Channel attention module

2.3.2 spatial attention

The spatial attention module selectively aggregates the feature at each position by a weighted sum of the features at all positions. Any two positions with similar features can contribute mutual improvement regardless of their distance in spatial dimension. The saliency map from low-level features contains a lot of details which easily brings bad results. In saliency detection, we want to obtain detailed boundaries between salient objects and background without other texture which can distract human attention. Therefore, instead of considering all spatial positions equally, we adopt spatial attention to focus more on the foreground regions, which helps to generate effective features for saliency prediction.

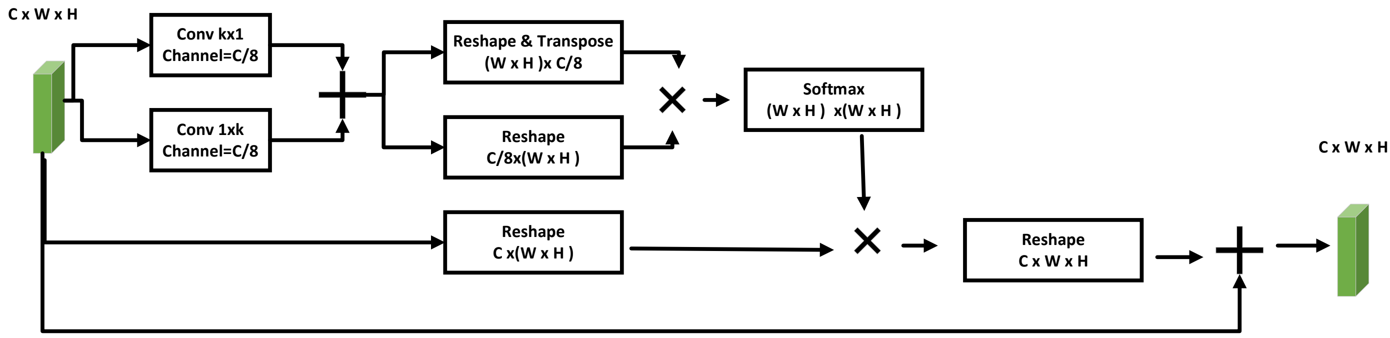

Figure 4 Spatial attention module

The proposed spatial attention module is shown in Fig.4 .For increasing receptive field and getting global information but not increasing parameters, similar to [18], we apply two convolution layers ,one's kernel is 1×k and the other's is k×1, for high-level feature to capture spatial concerns

2.3.3 local and fusion branch

The created salient mask is multiplied with original image to highlight the salient region and suppress the irrelevant part of image. This branch is a pretrained Densnet169 network.

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

3. Experiments

3.1 Dataset

According to table 1 Our dataset consists of 1870 grayscale x-ray image of lateral cephalometric that 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

3.2 Implementation Details

We employ a pretrained DensNet169 network as the backbone. The result shows that proposed method can represent more discriminative features therefore the accuracy of image classification can be increased. We implement our method based on Pytorch. We train all the networks using Adam optimizer with a mini-batch of size 8, and with β1 and β2 set to 0.9 and 0.99, respectively. While most of the networks converged during the first 250 epochs. The learning rate is 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.

3.3 Results

To validate the individual contribution of different components to the CVM stage classification performance, we perform an ablation experiment under different settings. Compared to the baseline (i.e., transfer learning with DensNet169), we observe that by integrating either a Spatial (SAM) or channel attention module (CAM) at each scale in the baseline architecture the performance improves between 6-7% in terms of accuracy and 3-5% in terms of F1-score.

Conclusions

In this paper, we showed that the domain knowledge and mimic of radiologist's behavior in CVM stage classification can be integrated into deep neural networks to improve their performance. In particular, we utilized a novel multi-scale attention module to combine semantic information at different levels and highlight the ROI and suppress the irrelevant part of the image.

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, which may be explained by the enhanced ability to model rich contextual dependencies over local and global features.

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.
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.
3. Hassel, B.; Farman, A.G. Skeletal maturation evaluation using cervical vertebrae. *Am. J. Orthod. Dentofac. Orthop.* **1995**, *107*, 58–66.
4. Seo, Hyejun, et al. "Comparison of Deep Learning Models for Cervical Vertebral Maturation Stage Classification on Lateral Cephalometric Radiographs." Journal of Clinical Medicine 10.16 (2021): 3591.

5. H. Amasya, D. Yildirim, T. Aydogan, N. Kemaloglu, and K. Orhan, "Cervical vertebral maturation assessment on lateral cephalometric radiographs using artificial intelligence: comparison of machine learning classifier models," *Dentomaxillofacial Radiology*, vol. 49, no. 5, p. 20190441, 2020.
6. R. S. Baptista *et al.*, "A semi-automated method for bone age assessment using cervical vertebral maturation," *The Angle* Orthodontist, vol. 82, no. 4, pp. 658-662, 2012.
7. M. Makaremi, C. Lacaule, and A. Mohammad-Djafari, "Deep learning and artificial intelligence for the determination of the cervical vertebra maturation degree from lateral radiography," Entropy, vol. 21, no. 12, p. 1222, 2019.
8. H. Kök, A. M. Acilar, and M. S. İzgi, "Usage and comparison of artificial intelligence algorithms for determination of growth and development by cervical vertebrae stages in orthodontics," Progress in orthodontics, vol. 20, no. 1, pp. 1-10, 2019.
9. Kim, Eun-Gyeong, et al. "Estimating Cervical Vertebral Maturation with a Lateral Cephalogram Using the Convolutional Neural Network." Journal of Clinical Medicine 10.22 (2021): 5400.
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, 6298–6306, 2017
11. Mnih, Volodymyr, Nicolas Heess, and Alex Graves. "Recurrent models of visual attention." Advances in neural information processing systems 27 (2014).
12. Schlemper, Jo, et al. "Attention gated networks: Learning to leverage salient regions in medical images." Medical image analysis 53 (2019): 197-207.
13. Shen, Yiqiu, et al. "An interpretable classifier for high-resolution breast cancer screening images utilizing weakly supervised localization." Medical image analysis 68 (2021): 101908.
14. Guan, Qingji, et al. "Diagnose like a radiologist: Attention guided convolutional neural network for thorax disease classification." arXiv preprint arXiv:1801.09927 (2018).
15. Sinha, Ashish, and Jose Dolz. "Multi-scale self-guided attention for medical image segmentation." IEEE journal of biomedical and health informatics 25.1 (2020): 121-130.
16. Britz, Denny, et al. "Massive exploration of neural machine translation architectures." arXiv preprint arXiv:1703.03906 (2017).
17. L. Chen et al., “SCA-CNN: Spatial and channel-wise attention in convolutional networks for image captioning,” in Proceedings of theIEEE conference on computer vision and pattern recognition, 2017, pp. 5659–5667
18. Peng, Chao, et al. "Large kernel matters--improve semantic segmentation by global convolutional network." Proceedings of the IEEE conference on computer vision and pattern recognition. 2017.