NON-LOCAL ATTENTION LEARNING FOR MEDICAL IMAGE CLASSIFICATION

Yang Wen[1], *Leiting Chen*[1,2], *Haisheng Chen*[1], *Ximan Tang*[1], *Yu Deng*[3], *Yongbiao Chen*[4], *Chuan Zhou*[*1]

[1]Key Laboratory of Digital Media Technology of Sichuan Province, School of Computer Science and Engineering, [2]Institute of Electronic and Information Engineering in Guangdong University of Electronic Science and Technology of China
[3]Dept. Biomedical Engineering, King's College London, [4]Shanghai Jiao Tong University
young.wen@foxmail.com, richardchen@uestc.edu.cn, hesse.chen@foxmail.com, tangxm@std.uestc.edu.cn, malikteng@foxmail.com, chenyongbiao0319@sjtu.edu.cn, zhouchuan@uestc.edu.cn

2021 IEEE International Conference on Multimedia and Expo (ICME) | 978-1-6654-3864-3/20/$31.00 ©2021 IEEE | DOI: 10.1109/ICME51207.2021.9428267

ABSTRACT

In recent years, deep convolutional neural networks (CNNs) have been used with great success in medical image classification. However, within CNNs, the convolutional operation only considers localized regions and the stacking of pooling layers can lead to the loss of information about tiny lesions. In this paper, we propose a non-local attention learning method that models long-range dependencies between pixels to preserve global information and help CNNs better identify the tiny lesions. It consists of two main parts, the non-local attention module and the non-local visual context fusion module, one for improving global understanding of the visual scene and one for aggregating non-local visual features. These two modules are used in parallel with the CNN backbone as an auxiliary branch. We conducted extensive experiments on two public medical image datasets and showed that our model achieves the best performance in terms of accuracy, precision, and sensitivity on both datasets compared to other recent models.

Index Terms— Deep learning, Convolutional neural network, Non-local attention learning, Medical image classification, Tiny lesions

1. INTRODUCTION

Medical image classification is a fundamental task of computer aided diagnosis aimed at the detection of diseases and lesions in different types of medical images. Recently, deep learning-based convolutional neural networks (CNNs) methods have been gaining increasing success in the field of medical image classification, such as dermatologist-level classification of skin cancer [1], tumor detection and classification in breast ammography [2], and classification of breast cancer [3]. These diseases are classified by identifying large regional pathological changes.

*Corresponding author

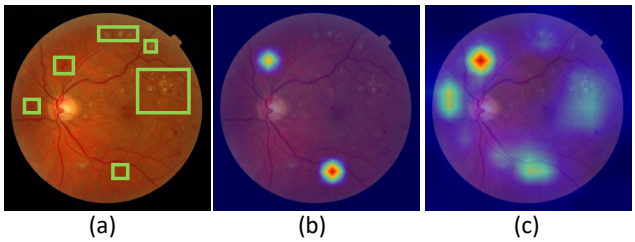

Fig. 1. Example of medical image and lesions detected by CNN models. (a) an image from MESSIDOR data set [4] with lesions marked by green boxes; (b) and (c) are lesions (better visualized in color) detected by CNN models that without and with our method.

However, when it comes to classification tasks that require the identification of tiny lesions, such as locating areas of infection in X-ray images of the lungs, CNN-based models typically fail to achieve excellent results. The main reason is that convolutional operations can only extract features from local neighborhoods, whereas accurate classification requires the network to focus on numerous tiny lesions that are unevenly distributed in the image. The pooling layers, commonly used in CNNs to downsample feature sizes and aggregate global information, making it difficult for the model to retain all the features of tiny lesions, as shown in Fig. 1.

To address this issue, the non-local network [5] was proposed to model the long-range dependencies via self-attention mechanism [6]. Through calculating the relative corresponding importance of all pixels, the self-attention mechanism can easily capture long-range dependencies and thus provide a global understanding of the visual scene. However, since the previous methods usually embed self-attention blocks into the CNN backbone after the downsampling operation in order to reduce the computational cost, it still fails to retain information about the tiny lesions. For medical image classification, it is important to preserve both the high-level visual semantics

978-1-6654-3864-3/21/$31.00 ©2021 IEEE

generated by CNNs and the global dependencies generated by self-attention blocks.

Therefore, in this paper, we propose a non-local attention learning specifically for medical image classification. Specifically, it consists of a dual-branch architecture with a CNN backbone branch and a non-local attention module branch to extract visual semantics and global dependencies in a parallel manner. In addition, a novel non-local visual context fusion module is proposed that aggregate global dependencies without causing information loss. Such methods maximizes the ability of our model on detecting the tiny lesions. To evaluate the proposed method, we conducted extensive experiments on two public medical data sets, namely, MESSIDOR [4] and COVID-19 [7] data sets. The results showed that model with our method surpasses all previous models.

Our main contributions are as follows.

- We propose a novel non-local attention learning approach using a dual-branch architecture that exploits visual semantics and global dependencies of CNNs and self-attentive blocks to preserve information about tiny lesions.
- We propose a non-local visual context fusion module to aggregate global dependencies without causing information loss.
- We demonstrate the effectiveness of the proposed methods through a series of ablation and qualitative experiments on two public MESSIDOR and COVID-19 data sets. The results show that our method allows the model to pay more attention to tiny lesions and achieves the highest classification accuracy compared to other CNN-based models.

2. RELATED WORK

Non-local attention. Human observation of data tends to prioritize the important parts of the data rather than treating all data equally. Such a process, also known as the attention mechanism, is widely used in a variety of computer vision tasks [8]. Hu *et al.* presented a squeeze-and-excitation block that uses the local context of the image to calibrate the weights of different channels and adaptively adjust channel importance for image classification [9]. Woo *et al.* proposed a convolutional block attention module (CBAM) [10] that divides the attention process into two separate parts, the channel attention and the spatial attention, to emphasize the most important spatial area in the image. Such approaches, *i.e.*, local attention, usually select the most important entity globally. However, in many scenarios, there are multiple important entities, so that some of them are discarded or treated equally by local attention, resulting in poor performance.

Therefore, non-local attention, which computes a weighted mean of all pixels, is recently proposed to allows distant pixels to contribute to final prediction [5]. Through reinforing long-range dependencies of visual features, the non-local attention has been successfully used in improving the performance for natural image recognition [11] and semantic segmentation [12]. Since the non-local attention module can be integrated with the existing architecture to enhance its ability to discover global dependencies, it is also very suitable for medical image classification.

Attention-based medical image classification. In recent years, deep architectures (*e.g.*, CNNs) and attention model have made great breakthroughs in the field of medical image analysis [13–15], while studies on the application of non-local attention is still very rare. Wang *et al.* proposed a non-local U-Net model to equip the non-local attention module as a flexible global feature aggregator for biomedical image segmentation [16]. To the best of our knowledge, the application of non-local attention remains unexplored in medical classification.

3. METHOD

In this section, we first describe the overall framework of non-local attention learning and specify its dual-branch architecture in detail. Then we introduce the non-local visual context fusion module.

3.1. Dual-branch Architecture

To best exploit visual information from tiny lesions, we propose a deep learning-based framework consists a dual-branch architecture, as shown in Fig. 2. Unlike previous studies which embed non-local block into CNN backbone [5], we use a non-local attention module as a second encoder to extract visual features in parallel with the CNN. Specifically, we use the CNN backbone as branch A for local feature extraction and a non-local attention learning subnetwork of NAM and non-local visual context fusion module as branch B for global feature extraction. Since the non-local attention module is applied at the beginning of the network, no information is lost due to the pooling layer, so the global information is preserved intact.

Branch A: CNN Backbone. We use two popular CNN architectures, namely, ResNet18 [17] and ShuffleNetV2 [18], as fundamental CNN models. To keep the computational complexity relatively low and avoid overfitting during training, we intercepted the first four encoder modules of ResNet18 and ShuffleNetV2, which have been pre-trained on ImageNet data set [19], as the backbone of our model to generate four feature maps of e1, e2, e3, and e4.

Branch B: Non-local Attention Learning. To obtain global understanding of the input image, we introduces a non-local attention module (NAM) which collects contextual information to enchance pixel-wise representative capability of our

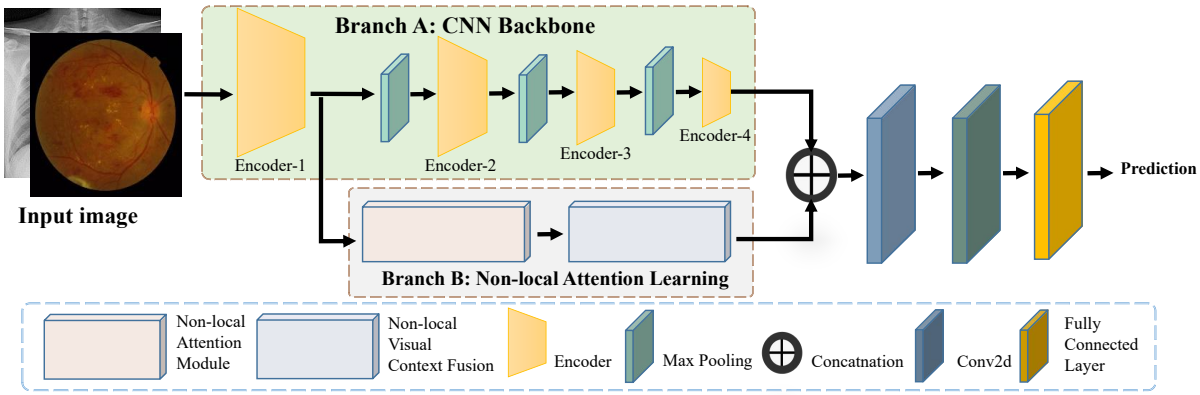

Fig. 2. Overall framework of the proposed non-local attention learning for medical image classification.

model. Let $\mathrm{X} = \{\mathrm{x}_i\}_{i=1}^{N_p}$ as the input feature maps, where N_p denotes the number of pixels in X. In practice, we regard the feature maps generated by the first convolutional layer of the backbone as X, thus N_p is of size $\mathrm{H} \times \mathrm{W}$, where H and W denote height and width, respectively. Following the previous study [5], we define the non-local operation in NAM as

$$\mathrm{y}_i = \frac{1}{\mathcal{N}(\mathrm{x})} \sum_{j=1}^{N_p} f\left(\mathrm{x}_i, \mathrm{x}_j\right) g\left(\mathrm{x}_j\right), \tag{1}$$

where x and y denote the input and output of the NAM, i is the index of the output position and j is the index of all positions to be calculated, $\mathcal{N}$ is a normalization factor. The pairwise function $f\left(\mathrm{x}_i, \mathrm{x}_j\right)$ computes the ralationship between position i and j. It can be defined as a dot-product similarity:

$$f\left(\mathrm{x}_i, \mathrm{x}_j\right) = \theta\left(\mathrm{x}_i\right)^T \phi\left(\mathrm{x}_j\right). \tag{2}$$

And the normalization factor is set as:

$$\mathcal{N}(\mathrm{x}) = \sum_{\forall j} f\left(\mathrm{x}_i, \mathrm{x}_j\right). \tag{3}$$

Given x_i, $\frac{1}{\mathcal{N}(\mathrm{x})} f\left(\mathrm{x}_i, \mathrm{x}_j\right)$ can be calculated using a softmax activation along the dimension j. Besides, a unary function g is used to compute the representation of the input. To alleviate computational complexity without compromising performance, we consider g in the form of a linear embedding:

$$g\left(\mathrm{x}_j\right) = W_g \mathrm{x}_j \tag{4}$$

where W_g are learnable parameters. In practice, we use a 1×1 convolutional layer as g.

The detailed structure of non-local attention module is shown in Fig. 3. The pairwise computation in Equation 2 can be simply done by the matrix multiplication between feature maps produced from the θ and ϕ, where θ and ϕ are 1×1 convolutional layers. After the feature maps pass through θ or ϕ the width and height remain the same while the channel is halved. Unlike the g in Non-local Block [5] which halves the channel, the g in our module keeps the original channel. This adjustment was made for the following two reasons: (1) we intend to keep more useful visual information; (2) the non-local block is directly embedded in an existing model, and our non-local focus module would be part of another parallel branch. In order to guarantee a high performance and to keep the computational complexity low , we place the non-local focus module after the first encoder of the backbone network. This allows us to build a rich hierarchy that combines both non-local and local information. We will continue the discussion of the placement of this module in Experiment 4.5.

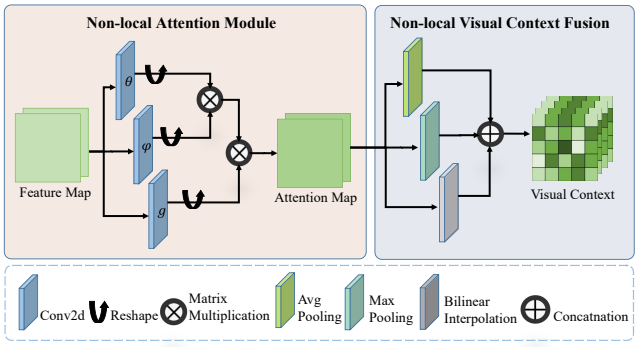

Fig. 3. Take an input feature map X of size $[C, H, W]$ as an example, where C, H and W denote the number of channels, height, and width, respectively. Through θ and ϕ, the channels of X are halved and X is transformed into X_θ and X_ϕ with size of $[HW, \frac{C}{2}]$ and $[\frac{C}{2}, HW]$, respectively. Then, X_θ and X_ϕ are multiplied to get the weight map $X_{\theta\phi}$ of size $[HW, HW]$. Finally, The $X_{\theta\phi}$ is multiplied by the reshaped map after g of size $[C, HW]$ to obtain the final attention map.

3.2. Non-local Visual Context Fusion

To aggregate non-local visual features from global dependencies, we propose a non-local visual context fusion (NVCF)

module, of which pipeline is shown in Fig. 3. Specifically, we downsampled the attention maps obtained from the NAM with three different layers: an average pooling layer, a max pooling layer, and a bilinear interpolation layer. We then concatenated the visual features of these three results as the final non-local visual context. The reasons for this design is to extract features from three different perspectives that preserve the most intact and informative global features, and to match the non-local visual context with the local feature size obtained from the CNN backbone. In this way, local and non-local features from the two branches can be concatenated together and then fed into the subsequent fully connected layer to generate the final classification prediction results.

4. EXPERIMENTS

4.1. Experimental Settings

Data set. The MESSIDOR data set [4] consists of 1,200 fundus images, for each of which results is given corresponding to the stage of diabetic retinopathy and macular edema symptom. The COVID-19 data set [7] consists of 2,905 CT pictures of the lungs, consisting 219 images of COVID-19, 1,345 images of viral pneumonia, and 1,341 normal images. During training and testing, we uniformly reshaped the images to a resolution of 386 pixels $\times$ 386 pixels, using 90% of the images for training and the remaining 10% for testing.

Metrics. We use the metrics of accuracy, sensitivity and area under curve (AUC) to evaluate the classification performance. The accuracy and sensitivity are defined as

$$accuracy\ (ACC) = \frac{tp + tn}{tp + fp + fn + tn} \tag{5}$$

$$sensitivity\ (SEN) = \frac{tp}{tp + fn}, \tag{6}$$

where tp, fp, tn, fn denote true positive, false positive, true negative and false negative, respectively. And AUC is the area under the ROC curve enclosed by the coordinate axis.

Implementation Details. To avoid overfitting, all images were randomly flipped (horizontally, vertically, or diagonally) and randomly rotated (by 30° to 270°) to enlarge the data set. We normalized the input image with the contrast limited adaptive histogram equalization [20] algorithm for contrast enhancement and brightness balancing. All our experiments were conducted with the PyTorch deep learning framework using the Adam optimizer [21] with a fixed learning rate of 2e-4 and a batch size of four. All experiments were performed on an Ubuntu 16.04 system employing an NVIDIA GeForce RTX 2060 with 6.0 GB memory. Every experiment was repeated for five times, and the metrics were obtained in the form of mean with standard deviation.

Table 1. Results of the ablation experiments on MESSIDOR data set. NAM, DB and NVCF denote non-local attention module, dual branches architecture and non-local visual context fusion module, respectively.

Method	AUC	SEN	ACC
ShuffleNetV2 [18]	89.88 ± 2.04	72.27 ± 2.41	88.13 ± 1.36
+ NAM	90.09 ± 2.19	72.81 ± 1.85	89.69 ± 0.86
+ NAM + DB	91.67 ± 1.64	73.93 ± 2.06	93.88 ± 0.62
+ NAM + DB + NVCF	**92.33 ± 1.60**	**74.37 ± 2.83**	**94.82 ± 0.53**
ResNet18 [17]	92.09 ± 1.39	77.50 ± 2.59	90.51 ± 1.21
+ NAM	92.89 ± 1.72	77.46 ± 1.98	90.23 ± 0.94
+ NAM + DB	94.10 ± 1.18	79.20 ± 2.21	94.13 ± 1.14
+ NAM + DB + NVCF	**95.39 ± 0.94**	**80.94 ± 1.84**	**95.62 ± 0.60**

4.2. Ablation Study

The main contributions of our method lie in the network architecture of dual branches which extracting local and non-local visual features with CNN backbone and NAM in a parallel manner, and the aggregation of non-local information using the NVCF module. To demonstrate the effectiveness of our method, we conduct ablation experiments on our network architecture and NVCF, and report the results on the MESSIDOR dataset. As shown in Table 1, compared to the baseline model, ShuffleNetV2, the NAM can improve the AUC, accuracy and sensitivity by 0.21%, 0.54% and 1.56%, respectively. Combining the DB, the AUC, accuracy and sensitivity can be further improved by 1.58%, 1.12% and 4.19%, respectively. And the best performance is obtained after adding the NVCF, reaching 92.33% for AUC, 74.37% for sensitivity and 94.82% for accuracy. It can be easily seen that our approach greatly improve classification performance. Similar results can be observed for ResNet18, which demonstrates the effective and validity of the proposed methods.

4.3. Comparison Analysis

We select a variety of the latest attention classification models (namely, SE-Net [9], CBAM [10] , Non-local Block [5], and CCNet [12]) for comparison on the MESSIDOR and COVID-19 data sets. We embedded the attentional modules of these models into the ResNet18 backbone in the manner of their papers. In addition, we tested two backbone networks, ShuffleNet-V2 and ResNet18. As shown in Table 2, compared to ResNet18, our model improves the AUC, sensitivity and accuracy by 3.3%, 4.44% and 1.97% on MESSIDOR data set, and by 1.29%, 1.92% and 1.61% on COVID-19 data set, respectively. Furthermore, compared to the latest attentional models, our model achieved an improvement of AUC, sensitivity and accuracy by at least 1.3%, 1.74%, and 0.88% on MESSIDOR data set, and by at least 1.22%, 1.72%, and 1.47% on COVID-19 data set. Evidently, the proposed method achieved the best performance than other attentional counterparts.

Table 2. Comparison of our method with baseline networks and other attention models on MESSIDOR dataset (M) and COVID-19 dataset (C-19). Ours denotes the ResNet18 network trained with our method. The best results are shown in bold.

Datasets	Method	AUC	SEN	ACC
M	ShuffleNet-V2 [18]	89.88 ± 2.04	72.27 ± 2.41	88.13 ± 1.36
	ResNet18 [17]	92.09 ± 1.39	76.50 ± 2.59	90.51 ± 1.21
	CBAM [10]	93.44 ± 1.09	78.86 ± 2.39	91.37 ± 1.13
	SE-Net [9]	93.48 ± 1.34	78.21 ± 2.59	91.32 ± 1.20
	Non-local Block [5]	93.89 ± 1.13	78.04 ± 2.16	90.95 ± 1.02
	CCNet [12]	94.09 ± 0.81	79.20 ± 1.83	91.60 ± 0.74
	HyNet [22]	94.17 ± 1.03	80.31 ± 1.62	91.92 ± 0.88
	Ours	**95.39 ± 0.94**	**80.94 ± 1.84**	**92.48 ± 0.91**
C-19	ShuffleNet-V2 [18]	98.21 ± 0.92	92.83 ± 1.69	96.89 ± 0.86
	ResNet18 [17]	98.38 ± 0.60	94.09 ± 1.57	97.31 ± 0.82
	CBAM [10]	98.55 ± 0.93	94.31 ± 1.26	97.80 ± 1.24
	SE-Net [9]	98.42 ± 0.97	94.20 ± 0.80	97.57 ± 1.20
	Non-local Block [5]	98.37 ± 0.79	94.09 ± 1.82	97.33 ± 0.87
	CCNet [12]	98.45 ± 0.88	94.29 ± 1.35	97.45 ± 0.70
	HyNet [22]	98.76 ± 1.13	94.60 ± 1.12	97.86 ± 0.92
	Ours	**99.67 ± 0.44**	**96.01 ± 0.77**	**98.92 ± 0.61**

4.4. Qualitative Results

To qualitatively demonstrate the effectiveness of the proposed method, we visualized saliency maps of the ResNet18 [17] trained with and without our method by using grad-CAM [23]. By comparing Fig. 4(b) with Fig. 4(c), it can be seen that our method refines the salient regions on the MESSIDOR and COVID-19 datasets and generates more focused attentional signals on the lesions. Our approach helps the model to capture global features from images and more visual characteristics on tiny lesions, thus improving the classification performance of the network.

4.5. Placement of Non-local Attention Module

In this section, we explore the placement of the NAM. The feature map of a CNN gets smaller and less computationally intensive as it is downsampled multiple times, but it also loses more global visual information due to pooling operations. Technically, using NAM directly on the input image preserves the most global information, but due to the complexity of the Equation 4, $O\left(CH^2W^2\right)$, it is too large and can lead to an overwhelming amount of computation. Therefore, in our implementation, we employed NAM after the first encoding block of the backbone, as it is in practice the best choice for balancing performance and computational cost. We also report the performance while NAM is employed after different levels of encoders. As shown in Table 3, using ResNet18 as the backbone network, our implementation yielded the best results on both MESSIDOR and COVID-19 data sets.

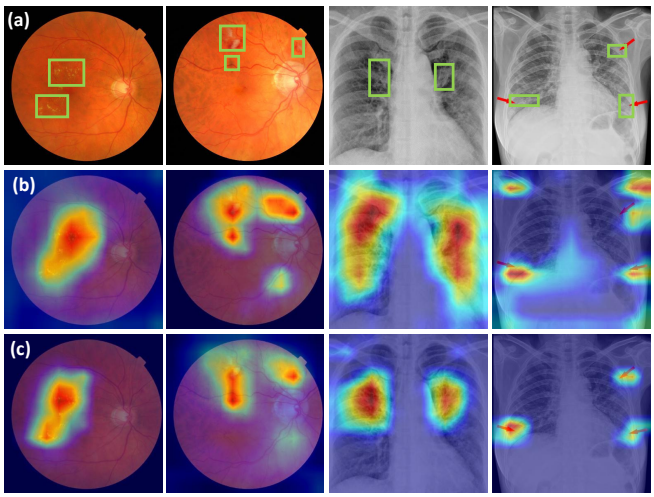

Fig. 4. Qualitative results of the proposed method. (a) input images with lesions marked by green boxes; (b) and (c) are visualization results produced by grad-CAM [23] from a network that without and with our method.

Table 3. The results when NAM used on outputs produced by different encoders.

NAM on features from	AUC	SEN	ACC
MESSIDOR			
Encoder-1	**95.39 ± 0.94**	**80.94 ± 1.84**	**92.48 ± 0.91**
Encoder-2	94.46 ± 1.25	80.25 ± 2.26	91.44 ± 1.14
Encoder-3	94.16 ± 1.49	78.92 ± 2.74	91.00 ± 1.37
Encoder-4	92.35 ± 1.23	74.26 ± 2.21	89.35 ± 0.85
COVID-19			
Encoder-1	**99.67 ± 0.44**	**96.01 ± 0.77**	**98.92 ± 0.61**
Encoder-2	97.39 ± 1.46	88.81 ± 1.03	95.62 ± 1.38
Encoder-3	96.31 ± 1.82	86.34 ± 1.41	94.40 ± 1.64
Encoder-4	95.61 ± 2.24	85.15 ± 1.60	93.85 ± 1.96

5. CONCLUSION

In this paper, we proposed non-local attention learning for medical image classification. It helps the network capture non-local information and perform visual context fusion, overcoming the problem of insufficient understanding of the global imgaes and insufficient focus on tiny lesions. We have demonstrated the validity of the proposed method through quantitative and qualitative studies on two data sets.

6. ACKNOWLEDGMENT

This study was supported by Sichuan Science and Technology Program (No.2019YJ0176 / 2019YJ0177 / 2019YFQ0005).

7. REFERENCES

[1] Andre Esteva, Brett Kuprel, Roberto A Novoa, Justin Ko, Susan M Swetter, Helen M Blau, and Sebastian Thrun, "Dermatologist-level classification of skin cancer with deep neural networks," pp. 115–118, 2017.

[2] Ayelet Akselrod-Ballin, Leonid Karlinsky, Sharon Alpert, Sharbell Hasoul, and Ella Barkan, "A region based convolutional network for tumor detection and classification in breast mammography," in *Expert Label Synthesis International Workshop on Deep Learning in Medical Image Analysis*, 2016.

[3] Dalal Bardou, Kun Zhang, and Sayed Mohammad Ahmad, "Classification of breast cancer based on histology images using convolutional neural networks," *IEEE Access*, vol. 6, pp. 24680–24693, 2018.

[4] Etienne Decencière, Xiwei Zhang, Guy Cazuguel, Bruno Lay, Béatrice Cochener, Caroline Trone, Philippe Gain, Richard Ordonez, Pascale Massin, Ali Erginay, Béatrice Charton, and Jean-Claude Klein, "Feedback on a publicly distributed database: the messidor database," *Image Analysis & Stereology*, vol. 33, no. 3, pp. 231–234, Aug. 2014.

[5] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He, "Non-local neural networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7794–7803.

[6] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin, "Attention is all you need," in *Advances in neural information processing systems*, 2017, pp. 5998–6008.

[7] Chowdhury, "Can ai help in screening viral and covid-19 pneumonia?," *IEEE Access*, vol. 8, pp. 132665–132676, 2020.

[8] X Liu and M Milanova, "Visual attention in deep learning: a review," *Int. Robot. Automat. J*, vol. 4, pp. 154–155, 2018.

[9] Jie Hu, Li Shen, and Gang Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7132–7141.

[10] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon, "Cbam: Convolutional block attention module," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 3–19.

[11] Hengshuang Zhao, Jiaya Jia, and Vladlen Koltun, "Exploring self-attention for image recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 10076–10085.

[12] Zilong Huang, Xinggang Wang, Lichao Huang, Chang Huang, Yunchao Wei, and Wenyu Liu, "Ccnet: Criss-cross attention for semantic segmentation," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 603–612.

[13] Fouzia Altaf, Syed MS Islam, Naveed Akhtar, and Naeem Khalid Janjua, "Going deep in medical image analysis: Concepts, methods, challenges, and future directions," *IEEE Access*, vol. 7, pp. 99540–99572, 2019.

[14] Qingji Guan and Yaping Huang, "Multi-label chest x-ray image classification via category-wise residual attention learning," *Pattern Recognition Letters*, vol. 130, 2018.

[15] L. Fang, C. Wang, S. Li, H. Rabbani, X. Chen, and Z. Liu, "Attention to lesion: Lesion-aware convolutional neural network for retinal optical coherence tomography image classification," *IEEE Transactions on Medical Imaging*, vol. 38, no. 8, pp. 1959–1970, 2019.

[16] Zhengyang Wang, Na Zou, Dinggang Shen, and Shuiwang Ji, "Non-local u-nets for biomedical image segmentation.," in *AAAI*, 2020, pp. 6315–6322.

[17] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[18] Ningning Ma, Xiangyu Zhang, Hai-Tao Zheng, and Jian Sun, "Shufflenet v2: Practical guidelines for efficient cnn architecture design," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 116–131.

[19] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 2009, pp. 248–255.

[20] Ali M Reza, "Realization of the contrast limited adaptive histogram equalization (clahe) for real-time image enhancement," *Journal of VLSI signal processing systems for signal, image and video technology*, vol. 38, no. 1, pp. 35–44, 2004.

[21] Diederik P Kingma and Jimmy Ba, "Adam, a method for stochastic optimization," in *Proceedings of the 3rd International Conference on Learning Representations (ICLR)*, 2015, vol. 1412.

[22] Yang Wen, Leiting Chen, Lifeng Qiao, Chuan Zhou, Shuo Xi, Rui Guo, and Yu Deng, "An efficient weakly-supervised learning method for optic disc segmentation," in *2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE, 2020, pp. 835–842.

[23] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 618–626.