JOINT ATTENTION FOR MEDICAL IMAGE SEGMENTATION

Mo Zhang[1,2,3], *Bin Dong*[4,1], *Quanzheng Li*[5]

[1]Peking University, Center for Data Science, China;
[2]Peking University, Center for Data Science in Health and Medicine, China;
[3]Beijing Institute of Big Data Research, Laboratory for Biomedical Image Analysis, China;
[4]Peking University, Beijing International Center for Mathematical Research (BICMR), China;
[5]Harvard Medical School, Massachusetts General Hospital, MGH/BWH Center for Clinical Data Science, Center for Advanced Medical Computing and Analysis, Department of Radiology, USA.

ABSTRACT

Medical image segmentation is crucial for computer aided diagnosis. In recent years, spatial attention mechanisms have leaded to breakthroughs in the task of image segmentation. In this work, we firstly present a unified formula for spatial attention mechanisms. Within this framework, we find that point-wise attention has better localization while self-attention can learn more global features. Motivated by this observation, we then propose a new joint attention module, which jointly leverages the advantages of point-wise attention and self-attention. Moreover, by integrating joint attention with DenseUNet, we conduct image segmentation experiments on two public datasets. The proposed method outperforms recent state-of-the-art models, verifying the superiority of joint attention. Additionally, ablation studies demonstrate that our joint attention obtains more balanced results compared to the previous point-wise attention and self-attention. The design of joint attention provides a novel insight into understanding spatial attention mechanisms.

Index Terms— Spatial attention, medical image segmentation, self-attention.

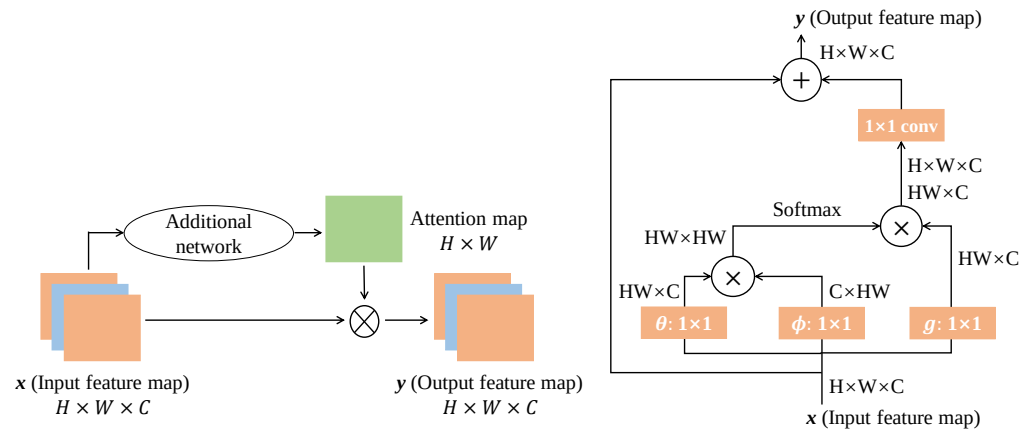
1. INTRODUCTION

Image segmentation is a fundamental yet challenging task in biomedical image analysis. It aims at segmenting clinically relevant objects from images. In the past decade, convolutional neural networks (CNNs) have exhibited promising performance on medical image segmentation. Currently, in this field, encoder-decoder based networks are the most widely used architectures, such as U-Net [1], DenseUNet [2], U-Net++ [3], etc. To further improve the representational power, some researchers attempted to incorporate the attention mechanism into CNNs for image recognition [4, 5, 6]. It enables neural networks to focus more on salient features while suppressing irrelevant ones. In computer vision, there are mainly two types of attention mechanisms: channel attention and spatial attention. Channel attention aims to model the relationship between channels, which assigns different weights to features of different channels [7]. The motivation of spatial attention is to model the dependencies among features at different spatial positions. In this paper, we primarily consider the application of spatial attention.

Spatial attention mechanism can be further categorized into point-wise attention and self-attention. As shown in Fig.1(a), point-wise attention firstly uses a small additional network to learn an attention map from data. Then the input feature map is multiplied (element-wise product) by the attention map to generate the final output feature map. In practice, the additional network usually includes convolutional layers with local receptive fields, so the attention map encodes local structural features. The spatial attention modules in CBAM [4], BAM [8] and attention U-Net [9] are typical examples of point-wise attention. On the other hand, there are also many applications of self-attention, such as nonlocal layer [5], dual attention [6] and nonlocal U-Net [10]. The structure of nonlocal layer is shown in Fig.1(b). It calculates the response of a pixel as a weighted mean of the features at all positions in the feature map, thus capturing long-range contextual information. Moreover, the nonlocal block contains a residual connection to fuse the original features and new features.

Although both the above attention mechanisms belong to spatial attention, self-attention is quite different from point-wise attention. Point-wise attention can learn rich local features while ignoring global context. On the contrary, self-attention is good at capturing long-range dependencies, while it ignores local position information. Motivated by this difference, in this work, we integrate point-wise attention with self-attention to form a new joint attention module, which attends jointly to both spatial locations and global features. In order to evaluate the proposed module, we incorporate it into DenseUNet and perform segmentation experiments on ISIC-

Bin Dong is supported in part by the NSFC under Grant 12090022, 12090020, 11831002.



(a) Point-Wise attention

(b) Nonlocal block (self-attention)

Fig. 1. Illustrations of spatial attention mechanisms. (a) The structure of point-wise attention. The additional network always involves convolutional layers. (b) The structure of nonlocal block. It consists of a residual connection for feature aggregation.

2017 and DRIVE datasets. The results show that our proposal achieves the state-of-the-art performance. The contributions of this paper are as follows:

1) We present a unified formula for spatial attention mechanisms. In this framework, we analyze the key difference between point-wise attention and self-attention.

2) We propose a new module called joint attention for medical image segmentation. It leverages the advantages of point-wise attention and self-attention simultaneously.

3) The proposed module achieves excellent performance on two datasets, demonstrating its potential to benefit more image analysis tasks even clinical practice.

2. METHOD

2.1. Spatial Attention Mechanism

Spatial attention has been widely used in deep leaning. It can be formulated as certain information aggregation [11, 12]:

$$\mathbf{y}_i = \sum_{m=1}^{M} \left[\sum_{\forall j \in \Omega_m(i)} A_m\left(\mathbf{x}_i, \mathbf{x}_j, \Delta_{ij}\right) \cdot \mathbf{W}_m \mathbf{x}_j \right]. \quad (1)$$

We denote M as the number of attention head. $\mathbf{x}_j \in R^{n_1}$ is the feature vector at position j of the input feature map, and $\mathbf{y}_i \in R^{n_2}$ is the output feature representation at position i. $A_m\left(\mathbf{x}_i, \mathbf{x}_j, \Delta_{ij}\right)$ represents the attention weight in the m-th attention head, where Δ_{ij} denotes the relative location between position i and j. $\mathbf{W}_m \in R^{n_2 \times n_1}$ is a learnable weight matrix, while $\Omega_m(i)$ enumerates all positions related to position i.

Point-Wise attention. As mentioned before and shown in Fig.1(a), point-wise attention consists of two steps: 1) An additional network is applied to learn the attention map, whose values represent attention weights of the corresponding positions; 2) The input feature vectors are multiplied by their attention weights, allowing networks to pay more attention to important features. In practice, applying convolutional layers to learn attention weights is always efficient, so here we let the additional network be a regular 3×3 convolution to simplify the analysis. The 3×3 convolution operation can also be defined as information aggregation:

$$w_i = \sum_{m=1}^{9} \left[\sum_{\forall j} A_m\left(\mathbf{x}_i, \mathbf{x}_j, \Delta_{ij}\right) \cdot \mathbf{W}_m \mathbf{x}_j \right], \quad (2)$$

$$A_m(\mathbf{x}_i, \mathbf{x}_j, \Delta_{ij}) = \begin{cases} 1 & \text{if } j = i + k_m \\ 0 & \text{else} \end{cases}. \quad (3)$$

In the above formula, the 3×3 convolution can be regarded as an attention mechanism with 9 attention heads. w_i is the learned attention weight at position i. Note that w_i is a scalar so that the dimension of $\mathbf{W}_m$ degenerates into $1 \times n_1$ here. $k_m \in \{(-1,-1),(-1,0),\cdots,(1,0),(1,1)\}$ denotes the offset corresponding to the m-th sampling location. Based on this expression, point-wise attention can be formulated as:

$$\mathbf{y}_i = \mathbf{x}_i \cdot w_i = \sum_{m=1}^{9} \left[\sum_{\forall j} A_m\left(\mathbf{x}_i, \mathbf{x}_j, \Delta_{ij}\right) \cdot \mathbf{x}_i \mathbf{W}_m \mathbf{x}_j \right]. \quad (4)$$

$A_m(\mathbf{x}_i, \mathbf{x}_j, \Delta_{ij})$ is the same as Eq.3, which encodes the relative position information between location i and j.

Self-Attention. Self-attention is also called intra-attention, as it models intra-sample relations. One sample can represent a sentence in natural language processing (NLP) or an image in computer vision. In CNNs, the essence of self-attention is a weighted average of all positions on the input feature

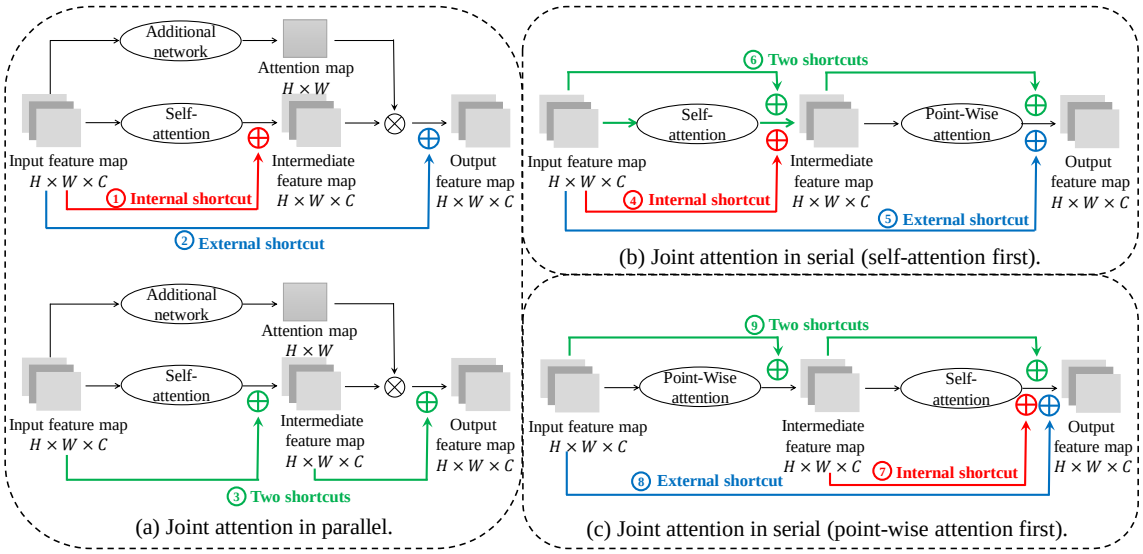


Fig. 2. The structures of joint attention. (a) Joint attention in parallel. (b) Joint attention in serial with self-attention first. (c) Joint attention in serial with point-wise attention first.

map. More specifically, it can also be expressed as the form of Eq.1:

$$\mathbf{y}_i = \sum_{\forall j} A_1\left(\mathbf{x}_i, \mathbf{x}_j, \Delta_{ij}\right) \cdot \mathbf{W}_1 \mathbf{x}_j. \tag{5}$$

It can be observed that self-attention only contains one attention head. The attention weight $A_1\left(\mathbf{x}_i, \mathbf{x}_j, \Delta_{ij}\right)$ denotes the correlation between features on the position i and j. Further, $A_1\left(\mathbf{x}_i, \mathbf{x}_j, \Delta_{ij}\right)$ can be written as $A_1\left(\mathbf{x}_i, \mathbf{x}_j\right)$, as it is independent of the relative position Δ_{ij}. As designed in the nonlocal block, there are many choices of the function $A_1\left(\mathbf{x}_i, \mathbf{x}_j\right)$, such as Gaussian $e^{\mathbf{x}_i^T \mathbf{x}_j}$, embedded Gaussian $e^{\theta(\mathbf{x}_i)^T \phi(\mathbf{x}_j)}$ and dot product $\theta\left(\mathbf{x}_i\right)^T \phi\left(\mathbf{x}_j\right)$. In this work, we use $\theta\left(\mathbf{x}_i\right)^T \phi\left(\mathbf{x}_j\right)$, where $\theta(\cdot)$ and $\phi(\cdot)$ are implemented via 1×1 convolution. In this case, a normalization factor is often imposed on the attention weight, so the final weight is $A_1\left(\mathbf{x}_i, \mathbf{x}_j\right) = \frac{\theta(\mathbf{x}_i)^T \phi(\mathbf{x}_j)}{N}$. N is the number of positions on the feature map.

2.2. Joint Attention

Motivation. As discussed in Section 2.1, both point-wise attention and self-attention can be expressed by the unified formula Eq.1. However, their attention weights are quite different. The attention weights of point-wise attention (Eq.3) are relevant to spatial locations. In fact, this setting defines a region of interest associated with position i in its neighborhood. Hence, point-wise attention can provide good localization while losing global dependencies. Oppositely, the attention weight of self-attention is $A_1\left(\mathbf{x}_i, \mathbf{x}_j\right) = \frac{\theta(\mathbf{x}_i)^T \phi(\mathbf{x}_j)}{N}$, which ignores the relative position Δ_{ij}. This kind of attention weight only takes the similarity between features into account. Therefore, it is easy to capture global information but neglects some important spatial details. This motivates us to integrate these two attention techniques to benefit from both merits.

The structures of joint attention. We propose a new attention module, i.e. joint attention. It is a combination of point-wise attention and self-attention. Further, these two attention mechanisms can be integrated in three different ways: 1) Arrangement in parallel (Fig.2(a)); 2) Arrangement in serial with self-attention first and point-wise attention second (Fig.2(b)); 3) Arrangement in serial with point-wise attention first and self-attention second (Fig.2(c)). In addition, self-attention is always used in conjunction with a residual connection, so it is necessary to consider where to place this shortcut in the joint attention. In this work, we consider three options: 1) Inside the point-wise attention (eg. 1,4,7-shortcut in Fig.2); 2) Outside the point-wise attention (eg. 2,5,8-shortcut in Fig.2); 3) To use another shortcut specifically for the point-wise attention (eg. 3,6,9-shortcut in Fig.2). In view of these two aspects, there are totally 9 feasible structures of joint attention, which are illustrated in Fig.2. In the experiments, we will perform ablation studies to find the best structure.

CNN backbone. To evaluate the performance of joint attention, we insert it into the classic DenseUNet. As shown in Fig.1 of the supplementary material, DenseUNet is an encoder-decoder framework which incorporates the dense block [17] into UNet [1]. More details on DenseUNet can be found in the original work [2]. In this paper, joint at-

Table 1. Quantitative results of different methods on ISIC-2017 dataset.

Model	Dice	Precision	Recall	Accuracy
Abhishek [13]	0.8384	-	0.8130	0.9217
Abhishek [14]	0.8386	-	**0.8706**	0.9220
Kaul [15]	0.8404	0.8002	0.8222	**0.9349**
Zhang [16]	0.8463	0.9076	0.8432	0.9319
DenseUNet	0.8363	**0.9335**	0.8028	0.9300
DenseUNet-SA	0.8408	0.9131	0.8304	0.9321
DenseUNet-PWA	0.8396	0.9006	0.8377	0.9322
DenseUNet-JA	**0.8500**	0.9187	0.8358	0.9338

tention is inserted into the skip connection part to boost the representation ability of low-level features.

3. EXPERIMENTS

To demonstrate the effectiveness of joint attention, we compare four models: 1) The baseline DenseUNet; 2) DenseUNet-SA, which incorporates self-attention into DenseUNet; 3) DenseUNet-PWA, which integrates point-wise attention into DenseUNet; 4) DenseUNet-JA, which inserts joint attention into DenseUNet. Note that all the attention modules are placed in the skip connections of DenseUNet. For model evaluation, we calculate several common metrics including Dice, Precision, Recall, TNR, Accuracy and AUC (the area under the ROC curve).

Data Description. We validate the proposed module on two public datasets: DRIVE and ISIC-2017. The DRIVE dataset contains 40 fundus images with a dimension of 565×584, which aims to segment retinal vessels. Following the previous works, the whole dataset is split into 20 training images and 20 testing images. For ISIC-2017 dataset, its full name is International Skin Imaging Collaboration dataset. It consists of 2750 dermatoscope images and corresponding annotations of skin lesions. Moreover, ISIC-2017 dataset is divided into training set (2000 images), validation set (150 images) and testing set (600 images). As images in ISIC-2017 dataset have various resolutions, we resize all the images into 256×256. In regard to image preprocessing, we apply normalization and Contrast Limited Adaptive Histogram Equalization (CLAHE) to both datasets. More implementation details are in the supplementary material.

4. RESULTS

Evaluation of the performance of joint attention. Table 1 lists the experimental results on ISIC-2017 dataset, DenseUNet-JA obtains the best Dice coefficient compared to the baseline DenseUNet, DenseUNet-SA, DenseUNet-PWA and recent state-of-the-art methods. Accuracy (0.9338) of DenseUNet-JA is only slightly lower than the optimal one (0.9349). Moreover, predicted segmentation maps are depicted in Fig.2 of the supplementary material. As shown in Table 1 of the supplementary material, DenseUNet-JA performs best in Accuracy (0.9690) and AUC (0.9845) on DRIVE dataset. Its Dice (0.8164) and TNR (0.9864) are also very close to the maximum ones (0.8171 and 0.9881). To sum up, all of the above results demonstrate the excellent performance of joint attention on image segmentation. More ablation studies are in the supplementary material.

5. CONCLUSION

In this work, firstly, we propose a unified framework for spatial attention mechanisms. In this frame, we observe that point-wise attention maintains more position-related information while self-attention can encode the long-range context. Secondly, we present the joint attention module, which exploits the superiorities of point-wise attention and self-attention simultaneously. Next, we incorporate joint attention into DenseUNet to perform image segmentation experiments on two datasets. The results demonstrate the superior ability of our joint attention. Finally, through ablation studies, we demonstrate that the proposed joint attention indeed gets the advantages of point-wise attention and self-attention, which leads to its more balanced and comprehensive segmentation results.

6. REFERENCES

[1] Olaf Ronneberger, Philipp Fischer, and Thomas Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241.

[2] Xiaomeng Li, Hao Chen, Xiaojuan Qi, Qi Dou, Chi-Wing Fu, and Pheng-Ann Heng, "H-denseunet: hybrid

densely connected unet for liver and tumor segmentation from ct volumes," *IEEE transactions on medical imaging*, vol. 37, no. 12, pp. 2663–2674, 2018.

[3] Zongwei Zhou, Md Mahfuzur Rahman Siddiquee, Nima Tajbakhsh, and Jianming Liang, "Unet++: A nested u-net architecture for medical image segmentation," in *Deep learning in medical image analysis and multimodal learning for clinical decision support*, pp. 3–11. Springer, 2018.

[4] 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.

[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] Jun Fu, Jing Liu, Haijie Tian, Yong Li, Yongjun Bao, Zhiwei Fang, and Hanqing Lu, "Dual attention network for scene segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 3146–3154.

[7] 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.

[8] Jongchan Park, Sanghyun Woo, Joon-Young Lee, and In So Kweon, "Bam: Bottleneck attention module," *arXiv preprint arXiv:1807.06514*, 2018.

[9] Ozan Oktay, Jo Schlemper, Loic Le Folgoc, Matthew Lee, Mattias Heinrich, Kazunari Misawa, Kensaku Mori, Steven McDonagh, Nils Y Hammerla, Bernhard Kainz, et al., "Attention u-net: Learning where to look for the pancreas," *arXiv preprint arXiv:1804.03999*, 2018.

[10] Zhengyang Wang, Na Zou, Dinggang Shen, and Shuiwang Ji, "Non-local u-nets for biomedical image segmentation," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020, vol. 34, pp. 6315–6322.

[11] Xizhou Zhu, Dazhi Cheng, Zheng Zhang, Stephen Lin, and Jifeng Dai, "An empirical study of spatial attention mechanisms in deep networks," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 6688–6697.

[12] Hengshuang Zhao, Yi Zhang, Shu Liu, Jianping Shi, Chen Change Loy, Dahua Lin, and Jiaya Jia, "Psanet: Point-wise spatial attention network for scene parsing," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 267–283.

[13] Kumar Abhishek and Ghassan Hamarneh, "Matthews correlation coefficient loss for deep convolutional networks: Application to skin lesion segmentation," *arXiv preprint arXiv:2010.13454*, 2020.

[14] Kumar Abhishek, Ghassan Hamarneh, and Mark S Drew, "Illumination-based transformations improve skin lesion segmentation in dermoscopic images," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2020, pp. 728–729.

[15] Chaitanya Kaul, Nick Pears, and Suresh Manandhar, "Divided we stand: A novel residual group attention mechanism for medical image segmentation," *arXiv preprint arXiv:1912.02079*, 2019.

[16] Mo Zhang, Bin Dong, and Quanzheng Li, "Deep active contour network for medical image segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2020, pp. 321–331.

[17] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger, "Densely connected convolutional networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4700–4708.