Medical Transformer: Gated Axial-Attention for Medical Image Segmentation

Jeya Maria Jose Valanarasu[1], Poojan Oza[1], Ilker Hacihaliloglu[2], and Vishal M. Patel[1]

[1] Johns Hopkins University, Baltimore, MD, USA
[2] Rutgers, The State University of New Jersey, NJ, USA

Abstract. Over the past decade, deep convolutional neural networks have been widely adopted for medical image segmentation and shown to achieve adequate performance. However, due to inherent inductive biases present in convolutional architectures, they lack understanding of long-range dependencies in the image. Recently proposed transformer-based architectures that leverage self-attention mechanism encode long-range dependencies and learn representations that are highly expressive. This motivates us to explore transformer-based solutions and study the feasibility of using transformer-based network architectures for medical image segmentation tasks. Majority of existing transformer-based network architectures proposed for vision applications require large-scale datasets to train properly. However, compared to the datasets for vision applications, in medical imaging the number of data samples is relatively low, making it difficult to efficiently train transformers for medical imaging applications. To this end, we propose a gated axial-attention model which extends the existing architectures by introducing an additional control mechanism in the self-attention module. Furthermore, to train the model effectively on medical images, we propose a Local-Global training strategy (LoGo) which further improves the performance. Specifically, we operate on the whole image and patches to learn global and local features, respectively. The proposed Medical Transformer (MedT) is evaluated on three different medical image segmentation datasets and it is shown that it achieves better performance than the convolutional and other related transformer-based architectures. Code: https://github.com/jeya-maria-jose/Medical-Transformer

Keywords: Transformers · Medical Image Segmentation · Self-Attention.

1 Introduction

Developing automatic, accurate, and robust medical image segmentation methods have been one of the principal problems in medical imaging as it is essential for computer-aided diagnosis and image-guided surgery systems. Segmentation of organs or lesion from a medical scan helps clinicians make an accurate diagnosis, plan the surgical procedure, and propose treatment strategies. Following the

popularity of deep convolutional neural networks (ConvNets) in computer vision, ConvNets were quickly adopted for medical image segmentation. Networks like U-Net [17], V-Net [15], 3D U-Net [4], Res-UNet [27], Dense-UNet [13], Y-Net [14], U-Net++ [31], KiU-Net [22,21] and U-Net3+ [8] have been proposed specifically for performing image and volumetric segmentation for various medical imaging modalities. These methods achieve impressive performance on many difficult datasets, proving the effectiveness of ConvNets in learning discriminative features to segment the organ or lesion from a medical scan.

ConvNets are currently the basic building blocks of most methods proposed for image segmentation. However, they lack the ability to model long-range dependencies present in an image. More precisely, in ConvNets each convolutional kernel attends to only a local-subset of pixels in the whole image and forces the network to focus on local patterns rather than the global context. There have been works that have focused on modeling long-range dependencies for ConvNets using image pyramids [29], atrous convolutions [3] and attention mechanisms [9]. However, it can be noted that there is still a scope of improvement for modeling long-range dependencies as the majority of previous methods do not focus on this aspect for medical image segmentation tasks.

(a) (b) (c) (d) (e)

Fig. 1. (a) Input Ultrasound of in vivo preterm neonatal brain ventricle. Predictions by (b) U-Net, (c) Res-UNet, (d) MedT, and (e) Ground Truth. The red box highlights the region which are miss-classified by ConvNet based methods due to lack of learned long-range dependencies. The ground truth here was segmented by an expert clinician. Although it shows some bleeding inside the ventricle area, it does not correspond to the segmented area. This information is correctly captured by transformer-based models.

To first understand why long-range dependencies matter for medical images, we visualize an example ultrasound scan of a preterm neonate and segmentation predictions of brain ventricles from the scan in Fig 1. For a network to provide an efficient segmentation, it should be able to understand which pixels correspond to the mask and which to the background. As the background of the image is scattered, learning long-range dependencies between the pixels corresponding to the background can help in the network to prevent miss-classifying a pixel as the mask leading to reduction of false positives (considering 0 as background and 1 as segmentation mask). Similarly, whenever the segmentation mask is large, learning long-range dependencies between the pixels corresponding to the mask is also helpful in making efficient predictions. In Fig 1 (b) and (c), we can see that the convolutional networks miss-classify the background as a brain ventricle while the proposed transformer-based method does not make that mistake. This happens as our proposed method learns long-range dependencies of the pixel regions with that of the background.

In many natural language processing (NLP) applications, transformers [5] have shown to be able to encode long-range dependencies. This is due to the self-attention mechanism which finds the dependency between given sequential input. Following their popularity in NLP applications, transformers have been adopted to computer vision applications very recently [6,20]. With regard to transformers for segmentation tasks, Axial-Deeplab [24] utilized the axial attention module [7], which factorizes 2D self-attention into two 1D self-attentions and introduced position-sensitive axial attention design for segmentation. In Segmentation Transformer (SETR) [30], a transformer was used as encoder which inputs a sequence of image patches and a ConvNet was used as decoder resulting in a powerful segmentation model. In medical image segmentation, transformer-based models have not been explored much. The closest works are the ones that use attention mechanisms to boost the performance [16,26]. However, the encoder and decoder of these networks still have convolutional layers as the main building blocks.

It was observed that that the transformer-based models work well only when they are trained on large-scale datasets [6]. This becomes problematic while adopting transformers for medical imaging tasks as the number of images, with corresponding labels, available for training in any medical dataset is relatively scarce. Labeling process is also expensive and requires expert knowledge. Specifically, training with fewer images causes difficulty in learning positional encoding for the images. To this end, we propose a gated position-sensitive axial attention mechanism where we introduce four gates that control the amount of information the positional embedding supply to key, query, and value. These gates are learnable parameters which make the proposed mechanism to be applied to any dataset of any size. Depending on the size of the dataset, these gates would learn whether the number of images would be sufficient enough to learn proper position embedding. Based on whether the information learned by the positional embedding is useful or not, the gate parameters either converge to 0 or to some higher value. Furthermore, we propose a Local-Global (LoGo) training strategy, where we use a shallow global branch and a deep local branch that operates on the patches of the medical image. This strategy improves the segmentation performance as we do not only operate on the entire image but focus on finer details present in the local patches. Finally, we propose Medical Transformer (MedT), which uses our gated position-sensitive axial attention as the building blocks and adopts our LoGo training strategy.

In summary, this paper (1) proposes a gated position-sensitive axial attention mechanism that works well even on smaller datasets, (2) introduces Local-Global (LoGo) training methodology for transformers which is effective, (3) proposes Medical-Transformer (MedT) which is built upon the above two concepts proposed specifically for medical image segmentation, and (4) successfully improves the performance for medical image segmentation tasks over convolutional networks and fully attention architectures on three different datasets.

2 Medical Transformer (MedT)

2.1 Self-Attention Overview

Let us consider an input feature map $x \in \mathbb{R}^{C_{in} \times H \times W}$ with height H, weight W and channels C_{in}. The output $y \in \mathbb{R}^{C_{out} \times H \times W}$ of a self-attention layer is computed with the help of projected input using the following equation:

$$y_{ij} = \sum_{h=1}^{H} \sum_{w=1}^{W} \text{softmax}\left(q_{ij}^{T} k_{hw}\right) v_{hw}, \tag{1}$$

where queries $q = W_Q x$, keys $k = W_K x$ and values $v = W_V x$ are all projections computed from the input x. Here, q_{ij}, k_{ij}, v_{ij} denote query, key and value at any arbitrary location $i \in \{1, \ldots, H\}$ and $j \in \{1, \ldots, W\}$, respectively. The projection matrices $W_Q, W_K, W_V \in \mathbb{R}^{C_{in} \times C_{out}}$ are learnable. As shown in Eq. 1, the values v are pooled based on global affinities calculated using softmax$(q^T k)$. Hence, unlike convolutions the self-attention mechanism is able to capture non-local information from the entire feature map. However, computing such affinities are computationally very expensive and with increased feature map size it often becomes infeasible to use self-attention for vision model architectures. Moreover, unlike convolutional layer, self-attention layer does not utilize any positional information while computing the non-local context. Positional information is often useful in vision models to capture structure of an object.

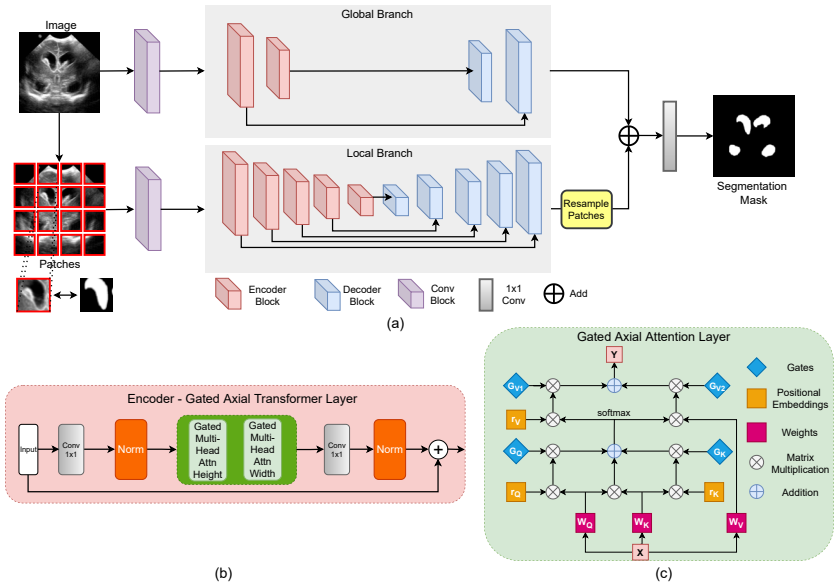

Fig. 2. (a) The main architecture diagram of MedT which uses LoGo strategy for training. (b) The gated axial transformer layer which is used in MedT. (c) Gated Axial Attention layer which is the basic building block of both height and width gated multi-head attention blocks found in the gated axial transformer layer.

Axial-Attention To overcome the computational complexity of calculating the affinities, self-attention is decomposed into two self-attention modules. The first module performs self-attention on the feature map height axis and the second one operates on the width axis. This is referred to as axial attention [7]. The axial attention consequently applied on height and width axis effectively model original self-attention mechanism with much better computational efficacy. To add positional bias while computing affinities through self-attention mechanism, a position bias term is added to make the affinities sensitive to the positional information [18]. This bias term is often referred to as relative positional encodings. These positional encodings are typically learnable through training and have been shown to have the capacity to encode spatial structure of the image. Wang *et al.* [24] combined both the axial-attention mechanism and positional encodings to propose an attention-based model for image segmentation. Additionally, unlike previous attention model which utilizes relative positional encodings only for queries, Wang *et al.* [24] proposed to use it for all queries, keys and values. This additional position bias in query, key and value is shown to capture long-range interaction with precise positional information [24]. For any given input feature map x, the updated self-attention mechanism with positional encodings along with width axis can be written as:

$$y_{ij} \;=\; \sum_{w=1}^{W} \text{softmax}\left(q_{ij}^{T} k_{iw} + q_{ij}^{T} r_{iw}^{q} + k_{iw}^{T} r_{iw}^{k}\right)\left(v_{iw} + r_{iw}^{v}\right), \tag{2}$$

where the formulation in Eq. 2 follows the attention model proposed in [24] and $r^q, r^k, r^v \in \mathbb{R}^{W \times W}$ for the width-wise axial attention model. Note that Eq. 2 describes the axial attention applied along the width axis of the tensor. A similar formulation is also used to apply axial attention along the height axis and together they form a single self-attention model that is computationally efficient.

2.2 Gated Axial-Attention

We discussed the benefits of using the axial-attention mechanism proposed in [24] for visual recognition. Specifically, the axial-attention proposed in [24] is able to compute non-local context with good computational efficiency, able to encode positional bias into the mechanism and enables the ability to encode long-range interaction within an input feature map. However, their model is evaluated on large-scale segmentation datasets and hence it is easier for the axial-attention to learn positional bias at key, query and value. We argue that for experiments with small-scale datasets, which is often the case in medical image segmentation, the positional bias is difficult to learn and hence will not always be accurate in encoding long-range interactions. In the case where the learned relative positional encodings are not accurate enough, adding them to the respective key, query and value tensor would result in reduced performance. Hence, we propose a modified axial-attention block that can control the influence positional bias can exert in the encoding of non-local context. With the proposed modification the self-

attention mechanism applied on the width axis can be formally written as:

$$y_{ij} = \sum_{w=1}^{W} \text{softmax}\left(q_{ij}^T k_{iw} + G_Q q_{ij}^T r_{iw}^q + G_K k_{iw}^T r_{iw}^k\right)\left(G_{V1} v_{iw} + G_{V2} r_{iw}^v\right), \quad (3)$$

where the self-attention formula closely follows Eq. 2 with added gating mechanism. Also, $G_Q, G_K, G_{V1}, G_{V2} \in \mathbb{R}$ are learnable parameters and together they create gating mechanism which control influence of the learned relative positional encodings have on encoding non-local context. Typically, if a relative positional encoding is learned accurately, the gating mechanism will assign it high weight compared to the ones which are not learned accurately. Fig 2 (c) illustrates the feed-forward in a typical gated axial attention layer.

2.3 Local-Global Training

It is evident that a transformer on patches is faster but patch-wise training alone is not sufficient for the tasks like medical image segmentation. Patch-wise training restricts the network in learning any information or dependencies for inter-patch pixels. To improve the overall understanding of the image, we propose to use two branches in the network, i.e., a global branch which works on the original resolution of the image, and a local branch which operates on patches of the image. In the global branch, we reduce the number of gated axial transformer layers as we observe that the first few blocks of the proposed transformer model is sufficient to model long range dependencies. In the local branch, we create 16 patches of size $I/4 \times I/4$ of the image where I is the dimensions of the original image. In the local branches, each patch is feed forwarded through the network and the output feature maps are re-sampled based on their location to get the output feature maps. The output feature maps of both of the branches are then added and passed through a 1×1 convolution layer to produce the output segmentation mask. This strategy improves the performance as the global branch focuses on high-level information and the local branch can focus on finer details. The proposed Medical Transformer (MedT) uses gated axial attention layer as the basic building block and uses LoGo strategy for training. It is illustrated in Fig 2 (a). More details on the architecture and an ablation study with regard to the architecture can be found in the supplementary file.

3 Experiments and Results

3.1 Dataset details

We use Brain anatomy segmentation (ultrasound) [25,23], Gland segmentation (microscopic) [19] and MoNuSeg (microscopic) [11,12] datasets for evaluating our method. More details about the datasets can be found in the supplementary.

3.2 Implementation details

We use binary cross-entropy (CE) loss between the prediction and the ground truth to train our network and can be written as:

$$\mathcal{L}_{CE(p,\hat{p})} = -\left(\frac{1}{wh}\sum_{x=0}^{w-1}\sum_{y=0}^{h-1}(p(x,y)\log(\hat{p}(x,y))) + (1-p(x,y))\log(1-\hat{p}(x,y))\right)$$

where w and h are the dimensions of the image, $p(x,y)$ corresponds to the pixel in the image and $\hat{p}(x,y)$ denotes the output prediction at a specific location (x,y). The training details are provided in the supplementary document.

For baseline comparisons, we first run experiments on both convolutional and transformer-based methods. For convolutional baselines, we compare with fully convolutional network (FCN) [1], U-Net [17], U-Net++ [31] and Res-Unet [27]. For transformer-based baselines, we use Axial-Attention U-Net with residual connections inspired from [24]. For our proposed method, we experiment with all the individual contributions. In gated axial attention network, we use axial attention U-Net with all its axial attention layers replaced with the proposed gated axial attention layers. In LoGo, we perform local global training for axial attention U-Net without using the gated axial attention layers. In MedT, we use gated axial attention as the basic building block for global branch and axial attention without positional encoding for local branch.

3.3 Results

Table 1. Quantitative comparison of the proposed methods with convolutional and transformer based baselines in terms of F1 and IoU scores.

Type	Network	Brain US		GlaS		MoNuSeg	
		F1	IoU	F1	IoU	F1	IoU
Convolutional Baselines	FCN [1]	82.79	75.02	66.61	50.84	28.84	28.71
	U-Net [17]	85.37	79.31	77.78	65.34	79.43	65.99
	U-Net++ [31]	86.59	79.95	78.03	65.55	79.49	66.04
	Res-UNet [27]	87.50	79.61	78.83	65.95	79.49	66.07
Fully Attention Baseline	Axial Attention U-Net [24]	87.92	80.14	76.26	63.03	76.83	62.49
Proposed	Gated Axial Attn.	88.39	80.7	79.91	67.85	76.44	62.01
	LoGo	88.54	80.84	79.68	67.69	79.56	66.17
	MedT	**88.84**	**81.34**	**81.02**	**69.61**	**79.55**	**66.17**

For quantitative analysis, we use F1 and IoU scores for comparison. The quantitative results are tabulated in Table 1. It can be noted that for datasets with relatively more images like Brain US, fully attention (transformer) based baseline performs better than convolutional baselines. For GlaS and MoNuSeg datasets, convolutional baselines perform better than fully attention baselines as it is difficult to train fully attention models with less data [6]. The proposed method is able to overcome such issue with the help of gated axial attention and LoGo both individually perform better than the other methods. Our final architecture MedT performs better than Gated axial attention, LoGo and all the previous methods. The improvements over fully attention baselines are 0.92

%, 4.76 % and 2.72 % for Brain US, GlaS and MoNuSeg datasets, respectively. Improvements over the best convolutional baseline are 1.32 %, 2.19 % and 0.06 %. All of these values are in terms of F1 scores. For the ablation study, we use the Brain US data for all our experiments. The results for the same has been tabulated in Table 2.

Furthermore, we visualize the predictions from U-Net [17], Res-UNet [27], Axial Attention U-Net [24] and our proposed method MedT in Fig 3. It can be seen that the predictions of MedT captures the long range dependencies really well. For example, in the second row of Fig 3, we can observe that the small segmentation mask highlighted on red box goes undetected in all the convolutional baselines. However, as fully attention model encodes long range dependencies, it learns to segment well thanks to the encoded global context. In the first and fourth row, other methods make false predictions at the highlighted regions as those pixels are in close proximity to the segmentation mask. As our method takes into account pixel-wise dependencies that are encoded with gating mechanism, it is able to learn those dependencies better than the axial attention U-Net. This makes our predictions more precise as they do not miss-classify pixels near the segmentation mask.

Table 2. Ablation Study

Network	U-Net [17]	Res-UNet [27]	Axial UNet [24]	Gated Axial UNet	Global only	Local only	LoGo	MedT
F1 Score	85.37	87.5	87.92	88.39	87.67	77.55	88.54	88.84

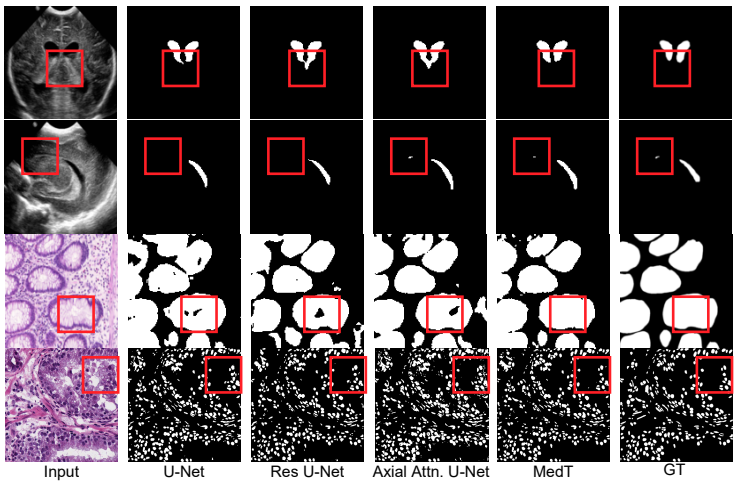

Fig. 3. Qualitative results on sample test images from Brain US, Glas and MoNuSeg datasets. The red box highlights regions where exactly MedT performs better than the other methods in comparison making better use of long range dependencies.

4 Conclusion

In this work, we explored the use of transformer-based architectures for medical image segmentation. Specifically, we propose a gated axial attention layer

which is used as the building block for multi-head attention models. We also proposed a LoGo training strategy to train the image in both full resolution as well in patches. The global branch helps learn global context features by modeling long-range dependencies, where as the local branch focus on finer features by operating on patches. Using these, we propose MedT (Medical Transformer) which has gated axial attention as its main building block for the encoder and uses LoGo strategy for training. Unlike other transformer-based model the proposed method does not require pre-training on large-scale datasets. Finally, we conduct extensive experiments on three datasets where we achieve a good performance for MedT over ConvNets and other related transformer-based architectures.

Acknowledgment

This work was supported by the NSF grant 1910141.

References

1. Badrinarayanan, V., Kendall, A., Cipolla, R.: Segnet: A deep convolutional encoder-decoder architecture for image segmentation. IEEE transactions on pattern analysis and machine intelligence **39**(12), 2481–2495 (2017)
2. Chen, J., Lu, Y., Yu, Q., Luo, X., Adeli, E., Wang, Y., Lu, L., Yuille, A.L., Zhou, Y.: Transunet: Transformers make strong encoders for medical image segmentation. arXiv preprint arXiv:2102.04306 (2021)
3. Chen, L.C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L.: Semantic image segmentation with deep convolutional nets and fully connected crfs. arXiv preprint arXiv:1412.7062 (2014)
4. Çiçek, Ö., Abdulkadir, A., Lienkamp, S.S., Brox, T., Ronneberger, O.: 3d u-net: learning dense volumetric segmentation from sparse annotation. In: International conference on medical image computing and computer-assisted intervention. pp. 424–432. Springer (2016)
5. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)
6. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
7. Ho, J., Kalchbrenner, N., Weissenborn, D., Salimans, T.: Axial attention in multidimensional transformers. arXiv preprint arXiv:1912.12180 (2019)
8. Huang, H., Lin, L., Tong, R., Hu, H., Zhang, Q., Iwamoto, Y., Han, X., Chen, Y.W., Wu, J.: Unet 3+: A full-scale connected unet for medical image segmentation. In: ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 1055–1059. IEEE (2020)
9. Huang, Z., Wang, X., Huang, L., Huang, C., Wei, Y., Liu, W.: Ccnet: Criss-cross attention for semantic segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 603–612 (2019)
10. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)

11. Kumar, N., Verma, R., Anand, D., Zhou, Y., Onder, O.F., Tsougenis, E., Chen, H., Heng, P.A., Li, J., Hu, Z., et al.: A multi-organ nucleus segmentation challenge. IEEE transactions on medical imaging **39**(5), 1380–1391 (2019)
12. Kumar, N., Verma, R., Sharma, S., Bhargava, S., Vahadane, A., Sethi, A.: A dataset and a technique for generalized nuclear segmentation for computational pathology. IEEE transactions on medical imaging **36**(7), 1550–1560 (2017)
13. Li, X., Chen, H., Qi, X., Dou, Q., Fu, C.W., Heng, P.A.: H-denseunet: hybrid densely connected unet for liver and tumor segmentation from ct volumes. IEEE transactions on medical imaging **37**(12), 2663–2674 (2018)
14. Mehta, S., Mercan, E., Bartlett, J., Weaver, D., Elmore, J.G., Shapiro, L.: Y-net: joint segmentation and classification for diagnosis of breast biopsy images. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 893–901. Springer (2018)
15. Milletari, F., Navab, N., Ahmadi, S.A.: V-net: Fully convolutional neural networks for volumetric medical image segmentation. In: 2016 fourth international conference on 3D vision (3DV). pp. 565–571. IEEE (2016)
16. Oktay, O., Schlemper, J., Folgoc, L.L., Lee, M., Heinrich, M., Misawa, K., Mori, K., McDonagh, S., Hammerla, N.Y., Kainz, B., et al.: Attention u-net: Learning where to look for the pancreas. arXiv preprint arXiv:1804.03999 (2018)
17. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical image computing and computer-assisted intervention. pp. 234–241. Springer (2015)
18. Shaw, P., Uszkoreit, J., Vaswani, A.: Self-attention with relative position representations. In: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers). pp. 464–468 (2018)
19. Sirinukunwattana, K., Pluim, J.P., Chen, H., Qi, X., Heng, P.A., Guo, Y.B., Wang, L.Y., Matuszewski, B.J., Bruni, E., Sanchez, U., et al.: Gland segmentation in colon histology images: The glas challenge contest. Medical image analysis **35**, 489–502 (2017)
20. Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., Jégou, H.: Training data-efficient image transformers & distillation through attention. arXiv preprint arXiv:2012.12877 (2020)
21. Valanarasu, J.M.J., Sindagi, V.A., Hacihaliloglu, I., Patel, V.M.: Kiu-net: Overcomplete convolutional architectures for biomedical image and volumetric segmentation. arXiv preprint arXiv:2010.01663 (2020)
22. Valanarasu, J.M.J., Sindagi, V.A., Hacihaliloglu, I., Patel, V.M.: Kiu-net: Towards accurate segmentation of biomedical images using over-complete representations. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 363–373. Springer (2020)
23. Valanarasu, J.M.J., Yasarla, R., Wang, P., Hacihaliloglu, I., Patel, V.M.: Learning to segment brain anatomy from 2d ultrasound with less data. IEEE Journal of Selected Topics in Signal Processing **14**(6), 1221–1234 (2020)
24. Wang, H., Zhu, Y., Green, B., Adam, H., Yuille, A., Chen, L.C.: Axial-deeplab: Stand-alone axial-attention for panoptic segmentation. arXiv preprint arXiv:2003.07853 (2020)
25. Wang, P., Cuccolo, N.G., Tyagi, R., Hacihaliloglu, I., Patel, V.M.: Automatic real-time cnn-based neonatal brain ventricles segmentation. In: 2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018). pp. 716–719. IEEE (2018)

26. Wang, X., Han, S., Chen, Y., Gao, D., Vasconcelos, N.: Volumetric attention for 3d medical image segmentation and detection. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 175–184. Springer (2019)
27. Xiao, X., Lian, S., Luo, Z., Li, S.: Weighted res-unet for high-quality retina vessel segmentation. In: 2018 9th international conference on information technology in medicine and education (ITME). pp. 327–331. IEEE (2018)
28. Zhang, Y., Liu, H., Hu, Q.: Transfuse: Fusing transformers and cnns for medical image segmentation. arXiv preprint arXiv:2102.08005 (2021)
29. Zhao, H., Shi, J., Qi, X., Wang, X., Jia, J.: Pyramid scene parsing network. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2881–2890 (2017)
30. Zheng, S., Lu, J., Zhao, H., Zhu, X., Luo, Z., Wang, Y., Fu, Y., Feng, J., Xiang, T., Torr, P.H., et al.: Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. arXiv preprint arXiv:2012.15840 (2020)
31. Zhou, Z., Siddiquee, M.M.R., Tajbakhsh, N., Liang, J.: 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)

Supplementary Material for Medical Transformer: Gated Axial-Attention for Medical Image Segmentation

Jeya Maria Jose Valanarasu[1], Poojan Oza[1], Ilker Hacihaliloglu[2], and Vishal M. Patel[1]

[1] Johns Hopkins University, Baltimore, MD, USA
[2] Rutgers, The State University of New Jersey, NJ, USA

In this supplementary material, we describe more details about the datasets that we used; provide more intricate details on our proposed architecture and training strategy; conduct an ablation study for our proposed methods; conduct an analysis on the number of parameters and present some more results.

1 Dataset details

In this section, we describe the datasets that we use in this paper in detail.

1.1 Brain US Dataset

Intraventricular hemorrhage (IVH) which results in the enlargement of brain ventricles is one of the main causes of preterm brain injury. The main imaging modality used for diagnosis of brain disorders in preterm neonates is cranial US because of its safety and cost-effectiveness. Also, absence of septum pellucidum is an important biomarker for septo-optic dysplasia diagnosis. Automatic segmentation of brain ventricles and septum pellucidum from these US scans is essential for accurate diagnosis and prognosis of these ailments. After obtaining institutional review board (IRB) approval, US scans were collected from 20 different premature neonates (age < 1 year). The total number of images collected were 1629 with annotations out of which 1300 were allocated for training and 329 for testing. We resize the images to 128×128 for all our experiments.

1.2 GLAS Dataset

GLAnd Segmentation (GLAS) datatset [19] contains microscopic images of Hematoxylin and Eosin (H&E) stained slides and the corresponding ground truth annotations by expert pathologists. It contains a total of 165 images which are split into 85 images for training and 80 for testing. Since the images in the dataset are of different sizes, we resize every image to a resolution of 128×128 for all our experiments.

1.3 MoNuSeg Dataset

MoNuSeg dataset [11,12] was created using H&E stained tissue images captured at 40x magnification. This dataset is diverse as it contains images across multiple organs and patients. The training data contains 30 images with around 22000 nuclear boundary annotations. The test data contains 14 images which have over 7000 nuclear boundary annotations. We resize the images to 512×512 for all our experiments.

2 MedT details

Medical Transformer (MedT) uses gated axial attention layer as the basic building block and uses LoGo strategy for training. MedT has two branches - a global branch and local branch. The input to both of these branches are the feature maps extracted from an initial conv block. This block has 3 conv layers, each followed by a batch normalization and ReLU activation. In the encoder of both branches, we use our proposed transformer layer while in the decoder, we use a conv block. The encoder bottleneck contains a 1×1 conv layer followed by normalization and two layers of multi-head attention layers where one operates along height axis and the other along width axis. Each multi-head attention block is made up of the proposed gated axial attention layer. Note that each multi-head attention block has 8 gated axial attention heads. The output from the multi-head attention blocks are concatenated and passed through another 1×1 conv which are added to residual input maps to produce the output attention maps. In each decoder block, we have a conv layer followed by an upsampling layer and ReLU activation. We also have skip connections between each encoder and decoder blocks in both the branches.

In the global branch of MedT, we have 2 blocks of encoder and 2 blocks of decoder. In the local branch, we have 5 blocks of encoder and 5 blocks of decoder.

3 Training details

We use a batch size of 4, Adam optimizer [10] and a learning rate of 0.001 for our experiments. The network is trained for 400 epochs. While training the gated axial attention layer, we do not activate the training of the gates for the first 10 epochs. We use a Nvidia Quadro 8000 GPU for all our experiments.

4 Analysis

In this section, we present an analysis over some of the parameters and methods we used for our proposed method.

4.1 Ablation Study

For the ablation study, we use the Brain US data for all our experiments. We first start with a standard U-Net. Then, we add residual connections to the U-Net making it a Res-UNet. Now, we replace all the convolutional layers in the encoder of Res-UNet with axial attention layers. This configuration is Axial Attention UNet inspired from [24]. Note that in this configuration we have an additional conv block at the front for feature extraction. Next, we replace all the axial attention layers from the previous configuration with gated axial attention layers. This configuration is denoted as Gated Axial attention. We then experiment using only the global branch and local branch individually from LoGo strategy. This shows that using just 2 layers in the global branch is enough to get a decent performance. The local branch in this configuration is tested on the patches extracted from the image. Then, we combine both the branches to train the network in an end-to-end fashion which is denoted as LoGo. Note that in this configuration the attention layers used are just axial attention layers [24]. Finally, we replace the axial attention layers in LoGo with gated axial attention layers which leads to MedT. The ablation study shows that each individual components of MedT provides useful contribution to improve the performance.

Table 1. Ablation Study

Network	U-Net [17]	Res-UNet [27]	Axial UNet [24]	Gated Axial UNet	Global only	Local only	LoGo	MedT
F1 Score	85.37	87.5	87.92	88.39	87.67	77.55	88.54	88.84

4.2 Number of Parameters

Table 2. Comparison in terms of number of parameters between the proposed method with the existing methods.

Network	FCN [1]	U-Net [17]	U-Net [17] (mod)	Res-UNet [27]	Res UNet [27] (mod)	Axial UNet [24]	Gated Axial UNet	MedT
Parameters	12.5 M	3.13 M	1.3 M	5.32 M	1.34 M	1.3 M	1.3 M	1.4 M
F1 Score	82.79	87.71	85.37	87.73	87.5	87.92	88.39	**88.84**

Although MedT is a multi-branch network, we reduce the number of parameters by using only 2 layers of encoder and decoder in the global branch and making the local branch operate on only patches of image. Also, the proposed gated axial attention block adds only 4 more learnable parameters to the layer.

In Table 2, we compare the number of parameters with other methods. U-Net corresponds to the original implementation according to [17]. U-Net (mod) corresponds to the U-Net configuration with reduced number of filters so as to match the number of parameters in MedT. Similarly, Res-UNet and Res-UNet (mod) corresponds to configurations with more and less number of parameters by adjusting the number of filters. We do this to show that even with more number of parameters, the baselines do not exceed MedT in terms of performance indicating that the improvement is not due to slight change in the number of parameters.

5 Results

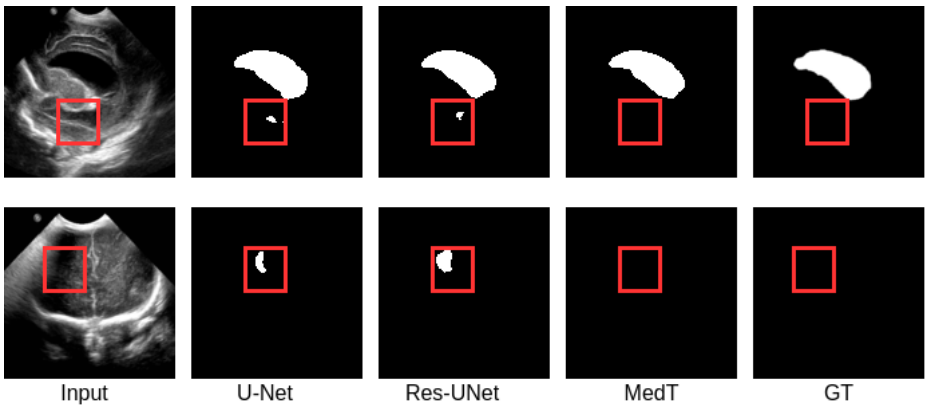

Fig. 1. Qualitative Results. The red box highlights the regions where our proposed method outperforms the convolutional baselines.

We present some additional qualitative results on top of the qualitative results presented in the main paper. In Fig 1, we visualize the predictions for our proposed method MedT along with the predictions for baselines UNet and Res-UNet for a couple of US scans. In both the samples, it can be seen that the regions that are highlighted in the red box are miss-classified to be brain ventricles for the convolutional baselines. However, our proposed attention based MedT does not make the same mistake.

6 Concurrent works

Very recently, TransUNet [2] was proposed which uses a transformer-based encoder operating on sequences of image patches and a convolutional decoder with skip connections for medical image segmentation. As TransUNet is inspired by

ViT, it is still dependent on pretrained weights obtained by training on a large image corpus. TransFuse [28] was recently proposed for polyp segmentation tasks using a parallel CNN branch and transformer branch fused using a BiFusion module. Unlike these works, we explore the feasibility of applying transformers working on only self-attention mechanisms as an encoder for medical image segmentation and without any need for pre-training.

References

1. Badrinarayanan, V., Kendall, A., Cipolla, R.: Segnet: A deep convolutional encoder-decoder architecture for image segmentation. IEEE transactions on pattern analysis and machine intelligence **39**(12), 2481–2495 (2017)
2. Chen, J., Lu, Y., Yu, Q., Luo, X., Adeli, E., Wang, Y., Lu, L., Yuille, A.L., Zhou, Y.: Transunet: Transformers make strong encoders for medical image segmentation. arXiv preprint arXiv:2102.04306 (2021)
3. Chen, L.C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L.: Semantic image segmentation with deep convolutional nets and fully connected crfs. arXiv preprint arXiv:1412.7062 (2014)
4. Çiçek, Ö., Abdulkadir, A., Lienkamp, S.S., Brox, T., Ronneberger, O.: 3d u-net: learning dense volumetric segmentation from sparse annotation. In: International conference on medical image computing and computer-assisted intervention. pp. 424–432. Springer (2016)
5. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)
6. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
7. Ho, J., Kalchbrenner, N., Weissenborn, D., Salimans, T.: Axial attention in multidimensional transformers. arXiv preprint arXiv:1912.12180 (2019)
8. Huang, H., Lin, L., Tong, R., Hu, H., Zhang, Q., Iwamoto, Y., Han, X., Chen, Y.W., Wu, J.: Unet 3+: A full-scale connected unet for medical image segmentation. In: ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 1055–1059. IEEE (2020)
9. Huang, Z., Wang, X., Huang, L., Huang, C., Wei, Y., Liu, W.: Ccnet: Criss-cross attention for semantic segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 603–612 (2019)
10. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
11. Kumar, N., Verma, R., Anand, D., Zhou, Y., Onder, O.F., Tsougenis, E., Chen, H., Heng, P.A., Li, J., Hu, Z., et al.: A multi-organ nucleus segmentation challenge. IEEE transactions on medical imaging **39**(5), 1380–1391 (2019)
12. Kumar, N., Verma, R., Sharma, S., Bhargava, S., Vahadane, A., Sethi, A.: A dataset and a technique for generalized nuclear segmentation for computational pathology. IEEE transactions on medical imaging **36**(7), 1550–1560 (2017)
13. Li, X., Chen, H., Qi, X., Dou, Q., Fu, C.W., Heng, P.A.: H-denseunet: hybrid densely connected unet for liver and tumor segmentation from ct volumes. IEEE transactions on medical imaging **37**(12), 2663–2674 (2018)

14. Mehta, S., Mercan, E., Bartlett, J., Weaver, D., Elmore, J.G., Shapiro, L.: Y-net: joint segmentation and classification for diagnosis of breast biopsy images. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 893–901. Springer (2018)
15. Milletari, F., Navab, N., Ahmadi, S.A.: V-net: Fully convolutional neural networks for volumetric medical image segmentation. In: 2016 fourth international conference on 3D vision (3DV). pp. 565–571. IEEE (2016)
16. Oktay, O., Schlemper, J., Folgoc, L.L., Lee, M., Heinrich, M., Misawa, K., Mori, K., McDonagh, S., Hammerla, N.Y., Kainz, B., et al.: Attention u-net: Learning where to look for the pancreas. arXiv preprint arXiv:1804.03999 (2018)
17. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical image computing and computer-assisted intervention. pp. 234–241. Springer (2015)
18. Shaw, P., Uszkoreit, J., Vaswani, A.: Self-attention with relative position representations. In: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers). pp. 464–468 (2018)
19. Sirinukunwattana, K., Pluim, J.P., Chen, H., Qi, X., Heng, P.A., Guo, Y.B., Wang, L.Y., Matuszewski, B.J., Bruni, E., Sanchez, U., et al.: Gland segmentation in colon histology images: The glas challenge contest. Medical image analysis **35**, 489–502 (2017)
20. Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., Jégou, H.: Training data-efficient image transformers & distillation through attention. arXiv preprint arXiv:2012.12877 (2020)
21. Valanarasu, J.M.J., Sindagi, V.A., Hacihaliloglu, I., Patel, V.M.: Kiu-net: Over-complete convolutional architectures for biomedical image and volumetric segmentation. arXiv preprint arXiv:2010.01663 (2020)
22. Valanarasu, J.M.J., Sindagi, V.A., Hacihaliloglu, I., Patel, V.M.: Kiu-net: Towards accurate segmentation of biomedical images using over-complete representations. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 363–373. Springer (2020)
23. Valanarasu, J.M.J., Yasarla, R., Wang, P., Hacihaliloglu, I., Patel, V.M.: Learning to segment brain anatomy from 2d ultrasound with less data. IEEE Journal of Selected Topics in Signal Processing **14**(6), 1221–1234 (2020)
24. Wang, H., Zhu, Y., Green, B., Adam, H., Yuille, A., Chen, L.C.: Axial-deeplab: Stand-alone axial-attention for panoptic segmentation. arXiv preprint arXiv:2003.07853 (2020)
25. Wang, P., Cuccolo, N.G., Tyagi, R., Hacihaliloglu, I., Patel, V.M.: Automatic real-time cnn-based neonatal brain ventricles segmentation. In: 2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018). pp. 716–719. IEEE (2018)
26. Wang, X., Han, S., Chen, Y., Gao, D., Vasconcelos, N.: Volumetric attention for 3d medical image segmentation and detection. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 175–184. Springer (2019)
27. Xiao, X., Lian, S., Luo, Z., Li, S.: Weighted res-unet for high-quality retina vessel segmentation. In: 2018 9th international conference on information technology in medicine and education (ITME). pp. 327–331. IEEE (2018)
28. Zhang, Y., Liu, H., Hu, Q.: Transfuse: Fusing transformers and cnns for medical image segmentation. arXiv preprint arXiv:2102.08005 (2021)

29. Zhao, H., Shi, J., Qi, X., Wang, X., Jia, J.: Pyramid scene parsing network. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2881–2890 (2017)
30. Zheng, S., Lu, J., Zhao, H., Zhu, X., Luo, Z., Wang, Y., Fu, Y., Feng, J., Xiang, T., Torr, P.H., et al.: Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. arXiv preprint arXiv:2012.15840 (2020)
31. Zhou, Z., Siddiquee, M.M.R., Tajbakhsh, N., Liang, J.: 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)