2023 45th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC) | 979-8-3503-2447-1/23/$31.00 ©2023 IEEE | DOI: 10.1109/EMBC40787.2023.10340201

A novel and simple approach to regularise attention frameworks and its efficacy in segmentation

Srividya Tirunellai Rajamani[1*], Kumar Rajamani[2], Björn W. Schuller[1,3]

***Abstract*— Deep neural networks with attention mechanism have shown promising results in many computer vision and medical image processing applications. Attention mechanisms help to capture long range interactions. Recently, more sophisticated attention mechanisms like criss-cross attention have been proposed for efficient computation of attention blocks. In this paper, we introduce a simple and low-overhead approach of adding noise to the attention block which we discover to be very effective when using an attention mechanism. Our proposed methodology of introducing regularisation in the attention block by adding noise makes the network more robust and resilient, especially in scenarios where there is limited training data. We incorporate this regularisation mechanism in the criss-cross attention block. This criss-cross attention block enhanced with regularisation is integrated in the bottleneck layer of a U-Net for the task of medical image segmentation. We evaluate our proposed framework on a challenging subset of the NIH dataset for segmenting lung lobes. Our proposed methodology results in improving dice-scores by 2.5 % in this context of medical image segmentation.**

I. Introduction

Deep learning based medical image segmentation is a challenging and widely researched topic [1, 2]. Limited availability of labelled data for training continues to be a major challenge in the medical domain, especially for rare medical disorders. Hence, there is significant need for approaches that can capture sufficient spatial context without requiring too complex models that are hard to train with limited labelled data.

The attention mechanism is a major recent advancement that helps gather contextual information within deep networks [3, 4, 5, 6]. Attention mechanisms used with deep networks significantly benefit semantic segmentation tasks [7, 8]. The recent criss-cross-attention module [9] captures global self-attention while remaining memory and time efficient. Regularisation is yet another prominent approach that has been shown to aid in various ways in training deep neural networks. Utilising gradient noise in very deep networks has been shown to improve learning as well as avoid overfitting [10, 11]. However, to the best of our knowledge, the effect of adding regularisation in the attention module has not yet been explored.

II. METHODOLOGY

Our simple, yet novel approach is to introduce regularisation in the attention module in order to make the network robust and resilient. In this work, we improve upon the criss-cross attention [9] for semantic segmentation tasks. We modify the criss-cross attention module by including regularisation. A block diagram of our proposed Recurrent Criss-Cross Attention (RCCA) module with regularisation is shown in Figure 1. As depicted in the block diagram, the regularisation is realised after each pass of the criss-cross attention module. This approach of interspersing regularisation and criss-cross attention was empirically discovered by us to be one of the most promising network constellations.

The effectiveness of this regularised RCCA module for the task of medical image segmentation is evaluated by integrating it within the U-Net architecture [12]. This is introduced as an extension of the U-Net's bottleneck in order to capture non-local contextual information in a robust way.

In the following sub-sections, we discuss in more detail, (a). The criss-cross attention module [9], (b). The baseline U-Net + CCA architecture we use, and (c). Our proposed regularised attention module.

A. Criss-Cross Attention Module

The criss-cross attention module (CCA) proposed by Huang et al. [9] aggregates contextual information in horizontal and vertical directions for each pixel. The input image $\mathbf{X}$ is passed through a convolutional neural network (CNN) to generate the feature maps $\mathbf{H}$ of reduced dimension. The CCA module comprises of three convolutional layers applied on $\mathbf{H} \in \mathbb{R}^{C \times H \times W}$ with 1×1 as kernel size.

The contextual information is aggregated by

$$\mathbf{H}'_{\mathbf{u}} = \sum_{\mathbf{i} \in |\mathbf{\Phi_u}|} \mathbf{A_{i,u}} \mathbf{\Phi_{i,u}} + \mathbf{H_u}, \tag{1}$$

with $\mathbf{H}'_{\mathbf{u}}$ being a feature vector in the module's output feature maps $\mathbf{H}' \in \mathbb{R}^{C \times H \times W}$ at position u and $\mathbf{A_i}, \mathbf{u}$ being a scalar value at channel i and position u in the attention map $\mathbf{A}$. The set $\mathbf{\Phi_u}$ is a collection of feature vectors in the feature map V obtained for feature adaption by applying another convolutional layer with 1×1 filters on $\mathbf{H}$.

[1] Srividya Tirunellai Rajamani and Björn W. Schuller are with the Chair of Embedded Intelligence for Health Care & Wellbeing, University of Augsburg, Germany

[2] Kumar Rajamani is with the Marwadi University, Rajkot, Gujarat, India

[3] Björn Schuller is also with GLAM, the Group on Language, Audio, & Music, Imperial College London, London, UK

[*] Corresponding author: `srividya.tirunellai@uni-a.de`

979-8-3503-2447-1/23/$31.00 ©2023 IEEE

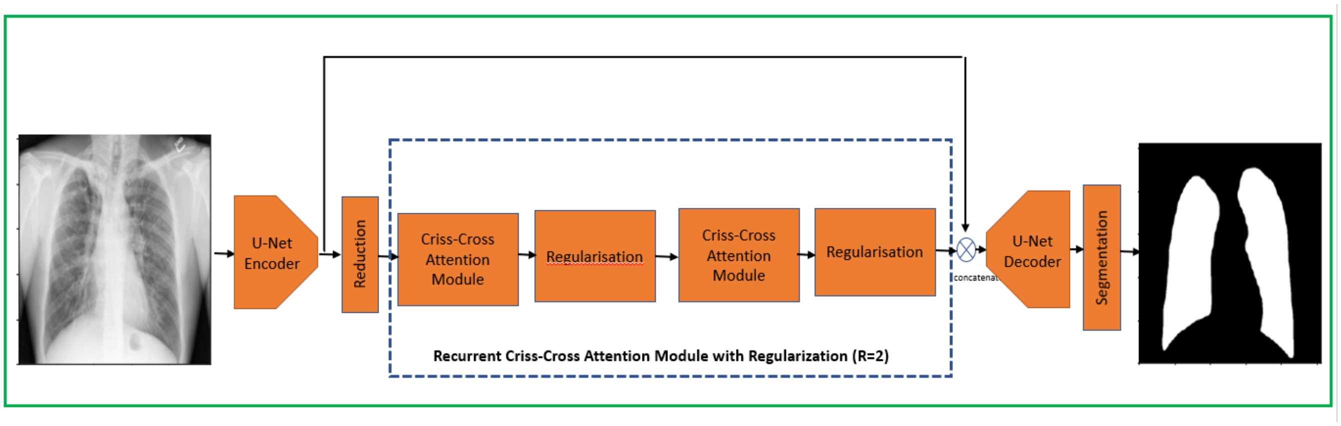
U-Net Encoder
Reduction
Criss-Cross Attention Module
Regularisation
Criss-Cross Attention Module
Regularisation
concatenat
U-Net Decoder
Segmentation
Recurrent Criss-Cross Attention Module with Regularization (R=2)

Fig. 1. Enclosed in dotted blue line is the block diagram of our proposed module, Recurrent Criss-Cross Attention (RCCA) module with regularisation. We utilise this proposed module in the bottleneck layer of the U-Net (enclosed in solid green line) for the task of medical image segmentation

B. *Baseline Network Architecture: U-Net + Criss-Cross Attention Module*

The baseline architecture we use is a U-Net structure from Oktay et al. [12, 13]. It consists of four blocks each in the downsampling and upsampling path. Each block consists of $2\times$(Batch Normalisation – 2D Convolution (kernel size 3×3, stride 1, padding 1) – ReLU). The last block consists of a 2D convolution with kernel size 1×1. In the downsampling path, in order to halve the spatial dimension of the feature maps after each block, max- pooling is applied. 2D transposed convolution is used in the upsampling path to double the size of the spatial dimension of the concatenated feature maps. The number of feature channels is increased as $(1-64-128-256-512)$ in the downsampling path and decreased again accordingly in the upsampling path. The U-Net's last layer outputs a number of feature channels that matches the number of label classes for semantic segmentation.

The local representation feature maps $\mathbf{H}$ being output from the U-Net's last block within the downsampling path serve as input of reduced dimension to the criss-cross module. The attention module is inserted in the bottleneck, as the feature maps are of reduced dimension, and hence the attention maps have smaller and therefore, more manageable time and space complexity. The CCNet [9] attention module gathers contextual information in the criss-cross path of each pixel leading to feature maps $\mathbf{H}'$. The contextual features $\mathbf{H}''$ obtained after $R=2$ loops through the attention module are concatenated with the feature maps $\mathbf{X}$ and merged by a convolution layer. The resulting feature maps are then passed through the U-Net's upsampling path.

C. *Our proposed regularised attention sampling*

The criss-cross attention operation gathers non-local information on a feature map of height H and width W. We compute a noise mask of similar dimension as the attention feature map. The noise is randomly sampled from a Gaussian distribution. The mean and variance values for the Gaussian noise yielding the best results were empirically determined through a grid search methodology. Such a noise mask is added to the attention feature map after each criss-cross attention module as seen in Figure 1.

III. Dataset

We conduct our experiments on the NIH chest X-ray dataset [14] that contains both posterior-anterior and anterior-posterior views. From this dataset, we used 100 abnormal chest X-ray images with various severity of lung diseases for which the lung masks were manually annotated[1] and were utilised by Tang et al. in their work [15]. The size of each chest X-ray image is 512×512.

IV. Experiments and Result

Training was done for 60 epochs using 60 images for training and the rest for validation and testing. The choice for usage of only 60 % for training was to simulate the scenario of training segmentation networks with limited training data. The mean dice score obtained by averaging over 5 runs are reported in Table I. The first row contains the mean dice score obtained when using U-Net + vanilla Recurrent Criss-Cross Attention (RCCA). The second row contains the mean dice score obtained when using U-Net with our proposed regularised RCCA. We empirically determine the mean and variance for the Gaussian-noise based regularisation. Our proposed approach results in a mean dice score of 0.955 which is an improvement of about 2.5 % over the baseline U-NET + vanilla RCCA.

In Figure 3, the train-loss curve as well as the validation mean dice score for one of the runs is shown. The blue curve is the plot obtained with U-Net + vanilla RCCA, while the orange curve is the one for the U-Net + the proposed regularised RCCA.

As demonstrated in Figure 2, our proposed regularised attention block estimates the lung masks much closer to the ground truth.

[1] Data: https://nihcc.app.box.com/s/r8kf5xcthjvvvf6r7l1an99e1nj4080m

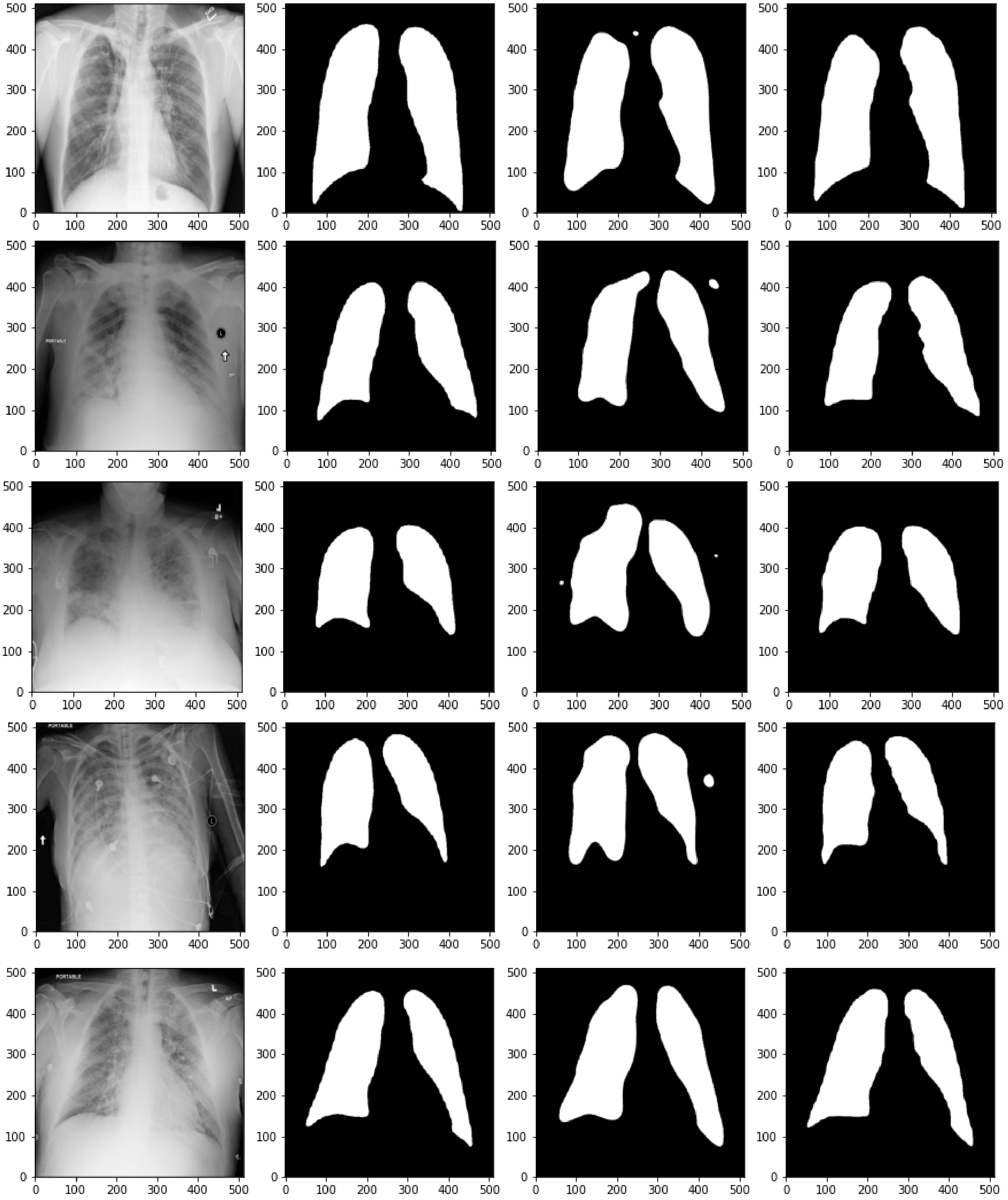

Fig. 2. Visual comparison of lung segmentation results. Results for 5 different test images are shown, one on each row. The 4 columns, from left to right, contain (a). the input image, (b). the ground-truth segmentation mask, (c). the segmentation mask predicted with U-Net + the vanilla RCCA, and (d). the segmentation mask predicted with U-Net + the regularised RCCA, respectively.

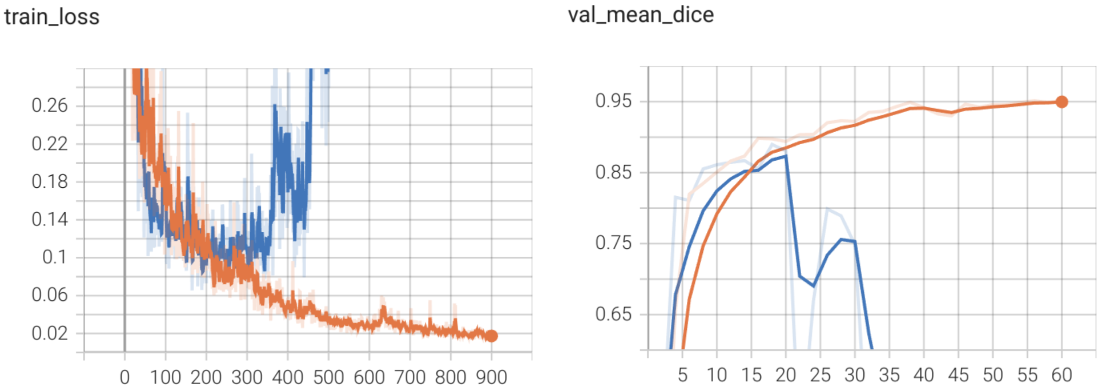

Fig. 3. Plots visualising the train-loss and the validation mean dice-score, respectively. The blue curve is the plot obtained with the U-Net + the vanilla RCCA, while the orange curve is for the U-Net + the proposed regularised RCCA.

Method	Mean Dice Score	% gain
U-Net + RCCA	0.931	-
U-Net + our proposed regularised RCCA	**0.955**	**2.5**

TABLE I

MEAN DICE SCORE VALUES AVERAGED OVER 5 RUNS

V. CONCLUSION

In this paper, we proposed a novel regularisation of the attention mechanism with additive Gaussian noise. This has been incorporated in the chosen U-Net + RCCA framework to improve the segmentation of lung lobes. Our experiments have shown that adding regularisation in a CCNet makes the segmentation network more robust and also generates better segmentation results as compared to the baseline.

To the best of our knowledge, this is the first time that a regularised attention mechanism is proposed in the CCNet framework for medical image segmentation. Our regularised attention helps the network to segment the objects of interest (in our case, the lung lobes) with an improved dice score. The loss curves in our regularised attention framework follow a well-defined trajectory as compared to that of the vanilla criss-cross attention, especially in scenarios where there is limited training data as visualised in Figure 3.

For our future experimentation, we plan to extend our hypothesis to attention-based classification tasks. Diverse regularisation approaches by changing the noise distribution could also be explored.

REFERENCES

[1] O. Ronneberger, P.Fischer, and T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," in *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, ser. LNCS, vol. 9351, Springer, 2015, pp. 234–241.

[2] G. Lin, A. Milan, C. Shen, and I. D. Reid, "RefineNet: Multi-path Refinement Networks for High-Resolution Semantic Segmentation," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 5168–5177.

[3] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proceedings of the 31st International Conference on Neural Information Processing Systems (NeurIPS)*, 2017.

[4] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local Neural Networks," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[5] S. T. Rajamani, K. Rajamani, A. Mallol-Ragolta, S. Liu, and B. Schuller, "A Novel Attention-Based Gated Recurrent Unit and its Efficacy in Speech Emotion Recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021.

[6] S. T. Rajamani, K. Rajamani, and B. Schuller, "Towards an Efficient Deep Learning Model for Emotion and Theme Recognition in Music," in *IEEE 23rd International Workshop on Multimedia Signal Processing (MMSP)*, 2021.

[7] K. Rajamani, S. D. Gowda, V. N. Tej, and S. T. Rajamani, "Deformable attention (DANet) for semantic image segmentation," in *2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, IEEE, 2022, pp. 3781–3784.

[8] K. Rajamani, P. Rani, H. Siebert, R. ElagiriRamalingam, and M. Heinrich, "Attention-augmented U-Net (AA-U-Net) for semantic segmentation," in *Signal, Image and Video Processing*, 2022.

[9] Z. Huang, X. Wang, L. Huang, C. Huang, Y. Wei, and W. Liu, "CCNet: Criss-Cross Attention for Semantic Segmentation," in *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019.

[10] A. Neelakantan, L. Vilnis, Q. Le, I. Sutskever, L. Kaiser, K. Kurach, and J. Martens, "Adding Gradient Noise Improves Learning for Very Deep Networks," in *arXiv, https://arxiv.org/abs/1511.06807*, 2015.

[11] B. Akilesh, T. Marwah, V. N. Balasubramanian, and K. Rajamani, "On the relevance of very deep networks for diabetic retinopathy diagnostics," in *Applications of Cognitive Computing Systems and IBM Watson : 8th IBM Collaborative Academia Research Exchange*, 2017.

[12] J. Schlemper, O. Oktay, M. Schaap, M. Heinrich, B. Kainz, B. Glocker, and D. Rueckert, "Attention gated networks: Learning to leverage salient regions in medical images," *Medical Image Analysis*, 2019.

[13] O. Oktay, J. Schlemper, L. L. Folgoc, M. Lee, M. Heinrich, K. Misawa, K. Mori, S. McDonagh, N. Y. Hammerla, B. Kainz, *et al.*, "Attention U-Net: Learning Where to Look for the Pancreas," in *International Conference on Medical Imaging with Deep Learning (MIDL)*, 2018.

[14] X. Wang, Y. Peng, L. Lu, Z. Lu, M. Bagheri, and R. M. Summers, "ChestX-Ray8: Hospital-Scale Chest X-Ray Database and Benchmarks on Weakly-Supervised Classification and Localization of Common Thorax Diseases," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[15] Y. Tang, Y. Tang, J. Xiao, and R. M. Summers, "XLSor: A Robust and Accurate Lung Segmentor on Chest X-Rays Using Criss-Cross Attention and Customized Radiorealistic Abnormalities Generation," in *International Conference on Medical Imaging with Deep Learning (MIDL)*, 2019.