FOCUSNET++: ATTENTIVE AGGREGATED TRANSFORMATIONS FOR EFFICIENT AND ACCURATE MEDICAL IMAGE SEGMENTATION

Chaitanya Kaul[⋆] *Nick Pears*[†] *Hang Dai*[‡] *Roderick Murray-Smith*[⋆] *Suresh Manandhar*[††]

[⋆] School of Computing Science, University of Glasgow, G12 8QQ, United Kingdom
[†] Department of Computer Science, University of York, YO10 5DD, United Kingdom
[‡] Mohamed bin Zayed University of Artificial Intelligence, Abu Dhabi, United Arab Emirates
[††] NAAMII, Katunje, Bhaktapur, Kathmandu, Nepal

ABSTRACT

We propose a new residual block for convolutional neural networks and demonstrate its state-of-the-art performance in medical image segmentation. We combine attention mechanisms with group convolutions to create our group attention mechanism, which forms the fundamental building block of our network, FocusNet++. We employ a hybrid loss based on balanced cross entropy, Tversky loss and the adaptive logarithmic loss to enhance the performance along with fast convergence. Our results show that FocusNet++ achieves state-of-the-art results across various benchmark metrics for the ISIC 2018 melanoma segmentation and the cell nuclei segmentation datasets with fewer parameters and FLOPs.

Index Terms— Group Attention, Medical Image Segmentation, Residual Learning

1. INTRODUCTION

Recently, the use of attention mechanisms in deep learning has been shown to learn better features [1] [2]. Learning better feature extractors is the most important task a network can do, especially for attention-based architectures, as the attention mechanisms are learnt over the extracted features. This has resulted in an general emphasis on optimizing convolutions. The simplest form of attention networks are the Spatial Transformer Networks [3] that learn the regions of interest from images with random clutter or noise. One of the first major visual attention methods was a two-level approach [4], where the images were first passed through an RCNN and selective search algorithms to generate proposals, followed by a gating operation using softmax over the ImageNet classes to remove low probability proposals. The remaining patches were then passed through a SVM classifier. The approach worked well on a subset of the ImageNet dataset, but requires a large amount of computation as well as hyperparameter tuning. SE-Nets [5] proposed to global average pool feature map (channel) information into a single vector creating a global representation. Using 'Squeeze-and-excitation', CNNs levaraged channel wise context to improve accuracy. One of the first works to explicitly show how filter groups leads to learn better representations is *Deep roots* [6], where, a sparse connecting structure resembling a tree root reduces parameters without any significant effect on the network accuracy. The impact of group convolutions was made apparent with ResNeXt [7] which performed impressively on the ILSVRC 2016 tasks. In this research, we extend group convolutions by incorporating attention mechanisms inside filter groups.

To this end, we propose FocusNet++, a deep learning architecture for medical image segmentation that harnesses the power of grouped convolutions, and combines them with a FocusNet-style attention mechanism [1] to get an improved performance compared to FocusNet, with fewer than half the parameters than it's successor. We enhance the decoding using fine-grained information from each decoder scale, which helps improve the network's segmentation ability. We compare with state of the art architectures, namely, Wide UNet and UNet++ [8], R2U-Net [9], Attention U-Net [2], BCDU-Net [10] and FocusNet [1] to show the superiority of our method.

The rest of this paper is organized as follows. We introduce FocusNet++ in Section 2 where we describe our novel group attention block, as well as the loss function used for our experiments. Section 3 summarizes our experiments on the skin cancer and cell nuclei segmentation datasets highlighting our model's state-of-the-art performance with reduced parameters and FLOPs compared to benchmark state-of-the-art medical image segmentation architectures. Our conclusions are provided in Section 4.

2. FOCUSNET++

Figure 1 shows the FocusNet++ architecture that adopts an encoder-decoder structure to learn multi-scale features for medical image segmentation. We carefully designed a group

Chaitanya Kaul and Roderick Murray-Smith acknowledge support from the iCAIRD project, funded by Innovate UK (project number 104690).

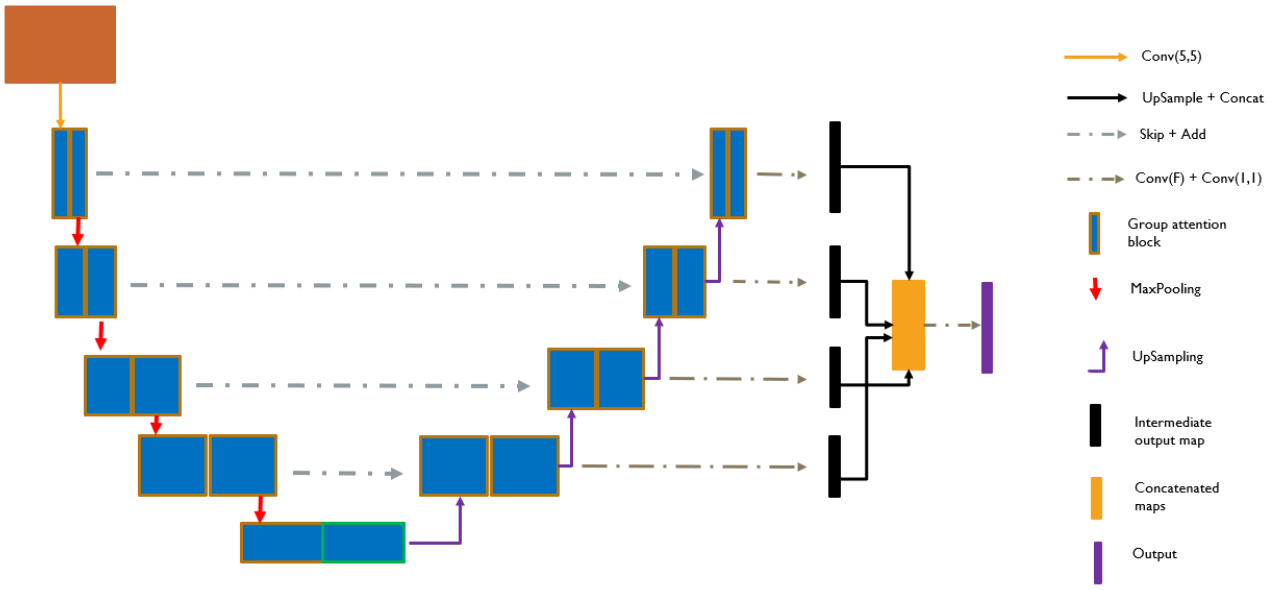

Fig. 1: The figure shows the architecture diagram for FocusNet++. The input image is processed by a series of residual group attention-max pooling blocks into a bottleneck and then decoded into a segmentation mask.

attention block, called the Residual Group Attention Network, that employs our novel attention methodology inside group convolutions for effective feature learning. Our network aims to address the problem of the relatively inferior decoding ability of existing segmentation architectures, combined with better feature extraction via our group attention block. We do this by creating a scheme that combines the output from each decoder scale to the final output, which leads to superior performance. The output from each scale passes through a *Conv-BN-LeakyReLU-Conv-Sigmoid* block to give intermediate outputs that are up-sampled, if required, to the output size, and then concatenated.

Finally, the concatenated volume is passed through a *Conv-BN-LeakyReLU-Conv-sigmoid* block to get the output segmentation map. In the encoder, the feature is down-sampled by a max-pooling operation. We add skip connections from the encoder to the decoder, rather than concatenating them. To up-sample the feature in decoder, we repeat the values in a kernel from a lower scale into an up-sampling scale and letting convolutions learn their values. We use dropout operation in the bottleneck layer to avoid over-fitting.

The receptive field of the first convolution kernel is 5×5. Following that, all convolutions kernels have a receptive field of 3×3 when used for feature extraction, and 1×1 when used to learn attention weights (i.e. preceding the sigmoid gating). The number of filters in each layer are $32 \rightarrow 64 \rightarrow 128 \rightarrow 192 \rightarrow 256 \rightarrow 192 \rightarrow 128 \rightarrow 64 \rightarrow 32$, divided into 4 filter groups in each layer. The general structure of FocusNet++ is similar to the U-Net architecture, apart from the mentioned changes.

2.1. Residual Group Attention Network (ResGANet)

As shown in Figure 2, the ResGANet employs pixel-wise attention inside filter groups, followed by combining the groups via a permutation invariant 1D convolution embedding. The squeeze and excitation block then re-calibrates the feature maps, which is followed by the residual mapping. The input to ResGANet is a feature volume that is processed by a 1×1 convolution operation. The general form of the aggregated transformation mapping is,

$$M = \mathbb{C}_{i=1}^{r} P_i(x)$$

where M is the output of the residual block, $\mathbb{C}$ denotes concatenation, and $P_i(x)$ is some transformation learnt by r separate stackings of trainable neurons transforming some input x. Here, $r = 4$, as we divide this input features into groups of four, to be processed by four separate convolution groups. Each group, alternatively, is responsible for learning the attention weights for the group to it's right, and the next group learns the features that need to be extracted. The attention weights for each attention group are obtained via two *BN-LeakyReLU-Conv* operations followed by a *Conv-Sigmoid* operation to get the per-pixel weights. Each attention group transforms its input in the following way,

$$A_r = \sigma(\mathbf{W_a}, \delta(x_r, \mathbf{W_k}))$$

where $\mathbf{W_k}$ and $\mathbf{W_a}$ are the convolution weights and the attention weights respectively, x_r is the r^{th} group that is input into this block, and δ denotes the *LeakyReLU* activation. The residual block contains two *BN-LeakyReLU-Conv*

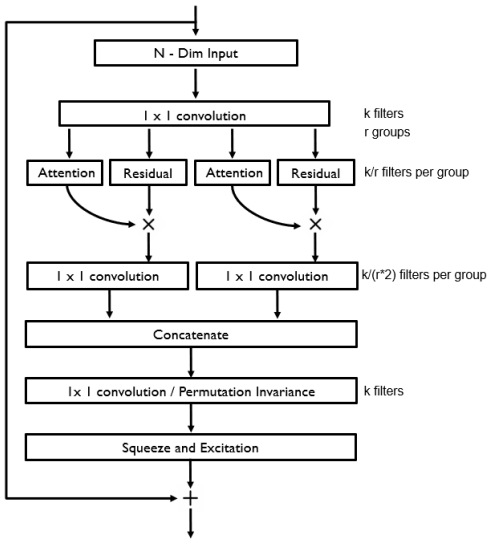

Fig. 2: Our novel residual block that first employs pixel-wise attention inside filter groups, followed by combining the groups via a permutation invariant embedding. The squeeze-and-excitation block then re-calibrates the feature maps, which is followed by the residual mapping.

operations followed by a skip connection that adds the features from the previous step to the residual block features. If the residual mapping is given by $O_r = x_r + F(x_r)$ then the network learns this $F(x_r)$ using some weights $\mathbf{W_k}$ as $F(x_r) = \delta(x_r, \mathbf{W_k})$. The output from the residual block is multiplied point-wise with the output from the attention block as $A = A_r \bigodot O_r$, weighting the pixels with a higher importance more prominently. Here, $\bigodot$ denotes the Hadamard product. The attention-infused output for each group propagates further in the block and is processed with a convolution block with a 1×1 receptive field and twice the number of filters. These intermediate filter maps are then concatenated and passed through a final 1×1 convolution. The feature maps are then re-calibrated using a squeeze-and-excitation operation that, finally, is followed by a residual connection.

2.2. Hybrid adaptive logarithmic loss

In order to have better recall, we adapt the balanced cross entropy loss with the Tversky loss in a novel way to create our hybrid loss function. The loss is defined as,

$$HL = (k)C_b + (1-k)TL \tag{1}$$

where $C_b = \Omega p \log(\hat{p}) + (1-\Omega)(1-p)\log(1-\hat{p})$, $TL = \sum_c (1 - TI)$, the subscript indicates a summation over the number of classes c, and $TI = \frac{|G \cap P|}{|G \cap P| + \alpha |P \backslash G| + \beta |G \backslash P|}$. To create a higher emphasis on the true positives, we select $\Omega = 0.7$. Generally, $\alpha = 0.3$, $\beta = 0.7$ proves to be the optimal setting in TL, adding higher weights to optimize over false positives and false negatives, so we retain those hyperparameter values. We weight the influence of both losses equally by setting $k = 0.5$. In order to optimize the loss further, we use a function whose derivative gives a non-linear response closer to the global minimum leading to a heavy penalty for misclassification. Hence, to mitigate the problem of pixel-class imbalance and poor convergence close to the minimum, we use the adaptive logarithmic loss [11] for our problem. The loss is defined as,

$$ALL\text{-}HL(x) = \begin{cases} \omega \ln\left(1 + \frac{|HL|}{\epsilon}\right) & |HL| < \gamma \\ |HL| - C & \texttt{otherwise} \end{cases} \tag{2}$$

where $C = \gamma - \omega \ln\left(1 + \left(\frac{\gamma}{\epsilon}\right)\right)$. We observe that the default hyperparameters of this loss are optimal for our experiments. Hence, we set $\gamma = 0.1,\ \omega = 10.0$ and $\epsilon = 0.5$.

3. EXPERIMENTS

For all experiments, the train-validate-test data split is fixed and no data augmentation is used. As a pre-processing step, we scale all pixel values to the range [0,1]. We convert the segmentation mask to binary by setting every pixel above the threshold of 0.5 to 1. All our experiments are trained with the hybrid loss (ALL-HL) strategy. The experiments are conducted using Keras [12] using a TensorFlow backend. The batch size for all experiments is kept constant at eight. The networks are trained on Nvidia GTX 1080Ti GPUs using a carefully constructed learning rate schedule, optimized for every architecture. All architectures were trained for a maximum of 50 epochs and the best model weights were saved by monitoring the validation loss. No early stopping was used.

3.1. Skin Cancer Segmentation

Method	Precision	Recall	DI	JI
FCN [13]	0.7176	0.8966	0.7861	0.7013
U-Net [14]	0.7398	0.9043	0.8167	0.7268
Wide UNet [8]	0.7439	0.9167	0.8224	0.7334
R2U-Net [9]	0.7381	0.9122	0.8271	0.7511
BCU-Net [10]	0.7576	0.9272	0.8637	0.7665
UNet++ [8]	0.7516	0.8889	0.8437	0.7435
Attn U-Net [2]	0.7526	0.9286	0.8741	0.7813
FocusNet [1]	0.7805	0.9328	0.8676	0.7751
FocusNet++	**0.8322**	**0.9471**	**0.9014**	**0.8271**

Table 1: Segmentation results on ISIC 2018 dataset.

The ISIC 2018 skin cancer segmentation dataset [15] has become a major benchmark dataset for the evaluation of medical imaging algorithms. We use the 2594 images with corresponding ground truths for our experiments. We divide these

Fig. 3: Results for skin cancer segmentation. From left, original image, ground truth, segmentation results from Attention U-Net [2], FocusNet [1] and FocusNet++.

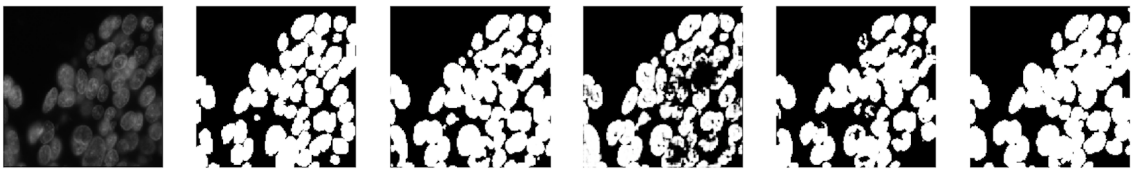

Fig. 4: Results for cell nuclei segmentation. From left, original image, ground truth, segmentation results from BCDU-Net [10], Attention U-Net [2], FocusNet [1] and FocusNet++.

Architecture	Params	FLOPs
UNet [14]	7.94×10^8	16.12×10^8
UNet++ [8]	9.04×10^8	42.44×10^8
BCDU-Net [10]	20.66×10^8	39.76×10^8
Attn U-Net [2]	8.91×10^8	17.82×10^8
FocusNet [1]	19.07×10^8	91.36×10^8
FocusNet++	$\mathbf{7.80}\times10^8$	$\mathbf{15.64}\times10^8$

Table 2: Comparing the model complexity and performance (on ISIC 2018) for FocusNet++ against state of the art segmentation architectures.

images into a training set of 1815 images, a validation set of 259 images, and a test set of 520 images. We resize every images to a smaller 256×256 size, via an anti-aliasing downsampling technique.

Table 1 summarizes our results for the experiments. FocusNet++ significantly outperforms every architecture across all metrics for the ISIC 2018 dataset with considerably fewer parameters and FLOPs. We get a 4.6% higher JI over the next best model. Table 2 summarizes the number of parameters and FLOPs of each architecture.

3.2. Cell Nuclei Segmentation

We now consider the segmentation of smaller regions inside images. For this, we use the cell nuclei segmentation dataset [16], which was a part of the Data Science Bowl 2018. It contains 670 images which we divided into a training set of 540 and a validation set of 130. We resize all images to 256×256. For this task, we evaluate the performance of our architecture by reducing the number of parameters (via reducing the number of filters per layer) for it in a way that it has fewer than one million FLOPs. We also reduced the number of parameters for the other architectures to account for the smaller size of this dataset so that we don't overfit. Our results are summarized in Table 3. FocusNet++ outperforms BCDU-Net with 2.5 times fewer parameters and 10 times fewer FLOPs.

Method	Params	FLOPs	Precision	Recall
U-Net [14]	3.62	1.89	0.8976	0.9052
BCU-Net [10]	5.22	9.98	0.9024	0.9078
Attn U-Net [2]	2.32	1.84	0.8782	0.9019
FocusNet [1]	5.03	22.38	0.9016	0.8981
FocusNet++	**1.84**	**0.98**	**0.9173**	**0.9139**

Table 3: Segmentation results on the cell nuclei segmentation dataset. Params and FLOPs are of the order of $\times10^8$.

4. CONCLUSION

We proposed an extremely efficient and accurate medical image segmentation architecture, FocusNet++, based on our novel residual group attention block that outperforms existing state-of-the-art architectures. We also propose an extremely lightweight variant of this architecture that outperforms architectures that are almost 2.5 times its size. We adapt the Tversky loss and balanced cross entropy loss in the adaptive logarithmic loss setting to boost performance over true positives and true negatives in order to obtain more well-rounded segmentations. Based on our experiments, our architecture requires lesser parameters and FLOPs, while giving better results compared to other architectures.

5. COMPLIANCE WITH ETHICAL STANDARDS

This research study was conducted retrospectively using human subject data made available in open access by [15] and [16]. Ethical approval was not required as confirmed by the license attached with the open access data.

6. ACKNOWLEDGEMENTS

Chaitanya Kaul and Roderick Murray-Smith acknowledge support from the iCAIRD project, funded by Innovate UK (project number 104690. No other conflicts of interest.

7. REFERENCES

[1] C. Kaul, S. Manandhar, and N. Pears, "FocusNet: An attention-based fully convolutional network for medical image segmentation," in *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*, April 2019, pp. 455–458.

[2] Ozan Oktay, Jo Schlemper, Loic Le Folgoc, Matthew Lee, Mattias Heinrich, Kazunari Misawa, Kensaku Mori, Steven McDonagh, Nils Y Hammerla, Bernhard Kainz, Ben Glocker, and Daniel Rueckert, "Attention u-net: Learning where to look for the pancreas," 2018.

[3] Max Jaderberg, Karen Simonyan, Andrew Zisserman, et al., "Spatial transformer networks," in *Advances in neural information processing systems*, 2015, pp. 2017–2025.

[4] Tianjun Xiao, Yichong Xu, Kuiyuan Yang, Jiaxing Zhang, Yuxin Peng, and Zheng Zhang, "The application of two-level attention models in deep convolutional neural network for fine-grained image classification," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 842–850.

[5] 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.

[6] Yani Ioannou, Duncan Robertson, Roberto Cipolla, and Antonio Criminisi, "Deep roots: Improving cnn efficiency with hierarchical filter groups," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1231–1240.

[7] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He, "Aggregated residual transformations for deep neural networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1492–1500.

[8] 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.

[9] Md Zahangir Alom, M. Hasan, C. Yakopcic, T. M. Taha, and V. K. Asari, "Recurrent residual convolutional neural network based on U-Net (R2U-Net) for medical image segmentation," *CoRR*, vol. abs/1802.06955, 2018.

[10] Reza Azad, Maryam Asadi-Aghbolaghi, Mahmood Fathy, and Sergio Escalera, "Bi-directional convlstm u-net with densley connected convolutions," in *Proceedings of the IEEE International Conference on Computer Vision Workshops*, 2019, pp. 0–0.

[11] Chaitanya Kaul, Nick Pears, and Suresh Manandhar, "Penalizing small errors using an adaptive logarithmic loss," *arXiv preprint arXiv:1910.09717*, 2019.

[12] Francois Chollet et al., "Keras," 2015.

[13] Evan Shelhamer, Jonathan Long, and Trevor Darrell, "Fully convolutional networks for semantic segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 4, pp. 640–651, Apr. 2017.

[14] O. Ronneberger, P.Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*. 2015, vol. 9351 of *LNCS*, pp. 234–241, Springer, (available on arXiv:1505.04597 [cs.CV]).

[15] N. Codella, V. Rotemberg, P. Tschandl, M. E. Celebi, S. Dusz, D. Gutman, B. Helba, A. Kalloo, K. Liopyris, M. Marchetti, H. Kittler, and A. Halpern, "Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (ISIC)," *CoRR*, vol. abs/1902.03368, 2019.

[16] "2018 data science bowl," .