FOCUSNET: AN ATTENTION-BASED FULLY CONVOLUTIONAL NETWORK FOR MEDICAL IMAGE SEGMENTATION

Chaitanya Kaul, Suresh Manandhar, Nick Pears

Department of Computer Science, University of York, Heslington, York, United Kingdom, YO10 5DD

ABSTRACT

We propose a novel technique to incorporate attention within convolutional neural networks using feature maps generated by a separate convolutional autoencoder. Our attention architecture is well suited for incorporation with deep convolutional networks. We evaluate our model on benchmark segmentation datasets in skin cancer segmentation and lung lesion segmentation. Results show highly competitive performance when compared with U-Net and it's residual variant.

Index Terms— Semantic segmentation, attention in CNNs, medical imaging, U-Net, residual connections

1. INTRODUCTION

Convolutional Neural Networks (CNNs) have had great success in in various computer vision tasks [1] [2] [3]. These architectures eradicate the need for hand crafted features, leading to training end-to-end systems that are able to learn important features and simultaneously perform the required task leading to state-of-the-art accuracy. As a consequence, CNNs have quickly become the baseline for most computer vision tasks. Techniques such as dropout have been used as a means of regularization to increase the generalizability of deep neural networks. Batch normalization improves the vanishing gradient problem in deep networks. Despite these techniques, the accuracy tends to stagnate with very deep CNNs. Architectures such as ResNet [4] address this issue by proposing residual learning that permits very deep CNNs without sacrificing accuracy. Identity mappings [5] has been proposed as an enhancement to the residual blocks used within ResNet.

More recently attention based methods have been developed that allow a network to focus on the most relevant parts of the data [6] [7]. These techniques of attention have been applied to tasks such as visual question answering [6] and lip reading in the wild [7], but their application to refine prediction in medical imaging is still to be tested. Self attention blocks such as those implemented within SE nets [8] re-calibrate the response of network filters using *squeeze* and *excitation* blocks.

In this paper, we propose a general architecture for combining a separate attention mechanism into a ResNet + SE network hybrid architecture. We demonstrate that our attention mechanism improves on an already state-of-the-art ResNet + SE architecture. For medical image processing tasks, existing works [9] [10] [11] demonstrate that pre- and post-processing such as, histogram equalization [9], image denoising [10] , contrast equalization [11], are crucial in achieving high accuracy. We demonstrate that our proposed architecture requires very minimal processing and performs well on images with varying conditions.

The paper is organized as follows: The next section describes our architecture. Section 3 describes the two datasets used, along with our evaluations. Section 4 presents our results, and a final section is used for conclusions.

Accepted as an ISBI 2019 conference paper

2. ATTENTION ARCHITECTURE

A conventional autoencoder first creates a low dimensional representation of the input and then upsamples from that representation to recreate the input. We exploit this encoder-decoder architecture to hierarchically extract latent attention maps leading to more accurate decoding. We propose the FocusNet architecture (see Fig.1), that employs two parallel branches of information flow with one branch solely devoted to attention. The attention branch employs an encoder-decoder structure with skip connections from the encoder to the decoder to facilitate better gradient flow. Our architecture provides a strong bias for the two networks to specialise and learn different representations.

Given an image, $x_i \in X$ where X is the mini batch, each layer of an encoder learns a mapping G, given by,

$$E_l = G_l(x, \mathbf{W}_l) \tag{1}$$

The decoder corresponding to this layer, decodes this representation in the following form:

$$D_l = [G_l(x, \mathbf{W}_l)\ ;\ H_l(H_{l-1}(x, \mathbf{W}_{l-\mathbf{1}}))] \tag{2}$$

where $[\ ;\ ]$ denotes concatenation via skip connection and H_{l-1} is the output from the previous decoder layer.

In the encoder in the second branch, the output can be represented as:

$$A_l = F_l(x, \mathbf{W}_l) \cdot \sigma(D_l) \tag{3}$$

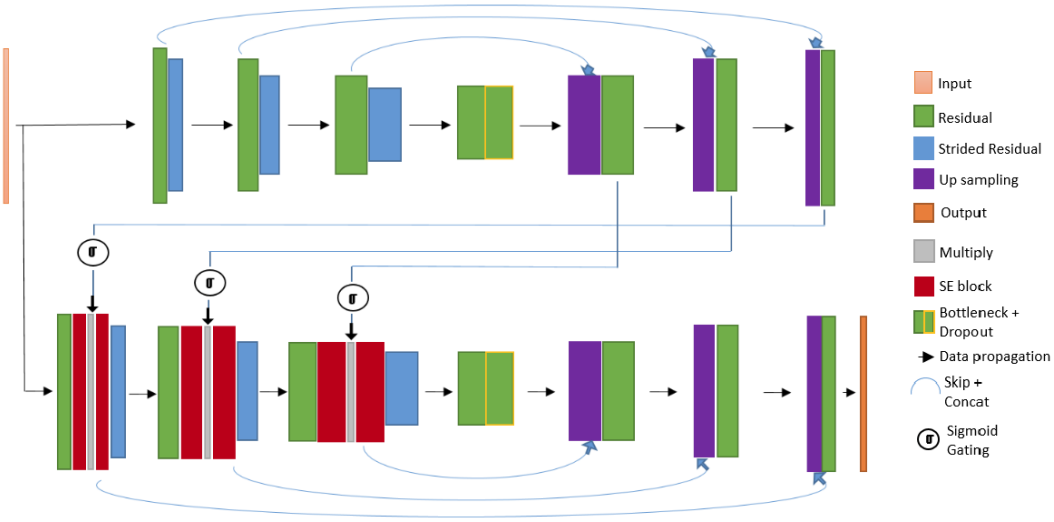

Fig. 1. Our network architecture uses attention to give better per pixel predictions, leading to better segmentation. The two branches are comprised of encoder-decoder structures where the per-layer decoded output is passed through a sigmoid gating function and multiplied with the output of the first SE block. The arrows show the direction of information flow in the network.

where A_l is the output of the l^{th} layer of the second encoder after *gating-and-multiplication*.
We use no bottleneck full pre-activation residual blocks [5] in the second branch (see bottom of Fig.1). Downsampling is done using strided convolutions rather than max pooling. The output is a 1x1 convolution with sigmoid activation that outputs per-pixel predictions. The rest of the convolutions have a 3x3 receptive field. The filter bank volumes used are $32 \rightarrow 64 \rightarrow 128 \rightarrow 256 \rightarrow 512 \rightarrow 256 \rightarrow 128 \rightarrow 64$.
Skip connections throughout the architecture facilitate better gradient flow, leading to easier training of a deeper network. Additionally, SE blocks are used extensively to recalibrate the weighting of the output feature maps at intermediate steps. For hyperparameter settings, we use dropout with a fixed rate of 0.2 throughout.

3. EVALUATION

3.1. Datasets

To demonstrate the performance of our architecture, we evaluate it on two different medical imaging datasets: skin cancer lesion segmentation and lung segmentation, as follows.

3.1.1. Skin Cancer Segmentation

This dataset [12] is a part of the ISIC skin cancer segmentation challenge on finding regions of melanoma inside skin images. The dataset contains 2000 RGB images in its training set, 150 images in its validation set and 600 test images. The images are high resolution and of varying sizes. The masks are 8-bit grayscale images with intensity value zero representing the background, and intensity value 255 representing the cancer region.

3.1.2. Lung Segmentation

The lung segmentation dataset [13] contains 2D images in .tif format with provided ground truth segmentation maps. The images are single channel with the size of 512x512 pixels. The dataset is fairly small, containing a total of 267 images.

3.2. Experiment Details

All experiments were done in python, using Keras [14] with a Tensorflow backend. Two Nvidia GTX 1080ti GPUs were used for all experiments. The batch size for training and testing was kept at 8. We trained all segmentation experiments using the dice coefficient loss. We subtracted the value of the loss from 1 to get a value between [0,1] for mathematical convenience, such that the loss could converge reducing towards zero. The Adam optimizer was used with the default learning rate. We reduced the learning rate on a plateau by half, if the validation loss didn't improve by an epsilon of 0.001 over 5 consecutive epochs. The validation loss was also monitored at every epoch such that the best performing model on the validation set could be saved. All experiments were run for a maximum of 80 epochs.
Images from both the datasets were resized to 256x256 pixels. No pre-processing was used other than subtracting the mean pixel value for the respective datasets and dividing by the standard deviation to normalize it. This was necessary so that the network wouldn't run into the exploding gradient

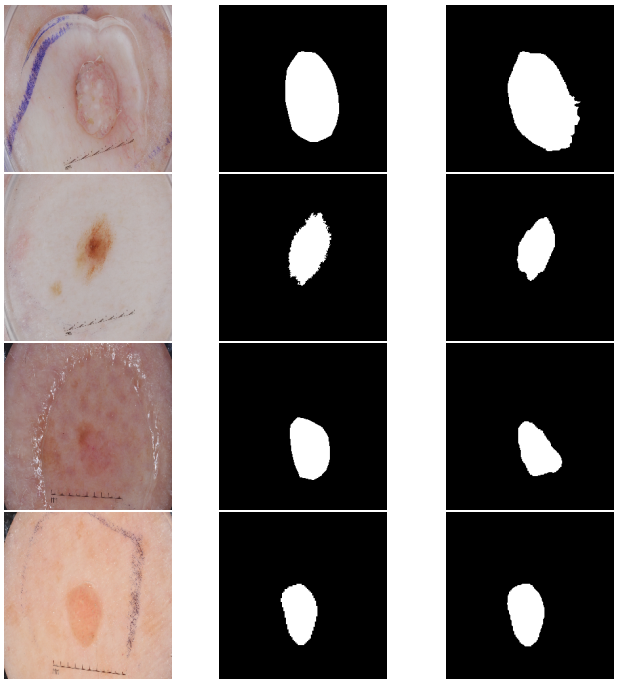

Fig. 2. Experimental segmentation results on the melenoma dataset. Column 1 is the input image, column 2 is the ground truth, and column 3 is the segmentation. The algorithm achieves good results given the problem setup, but the data clearly requires some amount of pre-processing to get better accuracy.

problem. The lung segmentation data was divided into a 80-20 training-validation split. The skin cancer segmentation data came with its own validation set and was used as-is.
We used data augmentation to increase the size of our datasets. For the lung cancer dataset, we used random zooms and flipping to increase the training images size to 1700. For the skin cancer dataset, we also added a very small number of random channel shifts to generate an augmented dataset of 6000 training images.

3.3. Evaluation Metrics

To evaluate the performance of our network, we used the following metrics. We measured the Dice (DI) and Jaccard index (JI) values for all networks as these are standard metrics in evaluating segmentation results. Along with them, we also measured the accuracy (AC), sensitivity (SE) i.e. the true positive rate and specificity (SP) i.e. true negative rate values for our results.

4. RESULTS

Table 1 provides the results on the lung segmentation dataset. We compared our results with the recurrent version of the U-Net, and a recurrent U-Net architecture with residual mappings instead of convolutions. Compared to the R2U-Net [15] and the corresponding recurrent class of architectures presented in that paper, our network outperforms in every metric except the sensitivity value.
Table 2 gives the results on the test set of the skin cancer dataset for detecting melanoma. It can be seen that our results are comparable with the recent results on this dataset. Compared to the LIN architecture [16], it can be observed that even without any pre- and post-processing, our network outperforms the architecture (in terms of the Jaccard Index, which was the evaluation metric of the competition) due to how our network incorporates a bias in the second branch of the encoder, forcing it to learn more robust features from the data. The LIN architecture, on the other hand, relied heavy image pre-processing.

Method	SE	SP	AC	JI
U-Net [15]	0.9696	0.9872	0.9828	0.9858
Res-U-Net [15]	0.9555	0.9945	0.9849	0.9850
RU-Net [15]	0.9734	0.9866	0.9836	0.9836
R2U-Net [15]	0.9826	0.9918	0.9897	0.9897
R2U-Net [15]	**0.9832**	0.9944	0.9918	0.9918
FocusNet (ours)	0.9757	**0.9981**	**0.9932**	**0.9965**

Table 1. Segmentation results on the validation set for lung segmentation dataset. We extend the table presented by [15] with our results on the dataset.

5. CONCLUSIONS

In this paper, we presented our novel attention-based deep neural network architecture, FocusNet. The architecture learns a better encoding of the data in its encoder layers to predict robust segmentation maps, making it a useful architecture to employ in image segmentation tasks. The residual nature of the convolutions used, facilitates training deeper neural networks that don't overfit. The skip connections used throughout help in gradient flow through the entire network. A drawback of the network is its lack of responsiveness to sensitivity metric of the data. We believe it is partly due to the fact that we trained the architecture without applying any pre- or post-processing on the datasets. Our method of incorporating attention in CNNs can be easily generalized to other domains of computer vision in addition to segmentation.

6. REFERENCES

[1] Karen Simonyan and Andrew Zisserman, "Very deep convolutional networks for large-scale image recognition," *CoRR*, vol. abs/1409.1556, 2014.

Method	SE	SP	AC	JI	DI
FCN-8s [16]	0.806	0.954	0.933	0.696	0.783
U-Net [2]	0.853	0.957	0.920	0.651	0.768
II-FCN [17]	0.841	0.984	0.929	0.699	0.794
Auto-ED [18]	0.836	0.966	**0.936**	0.738	0.824
Thao *et al.* [19]	0.6513	0.9421	0.8772	0.5065	0.6317
LIN [16]	**0.855**	0.974	0.934	0.753	**0.839**
FocusNet (ours)	0.7673	**0.9896**	0.9214	**0.7562**	0.8315

Table 2. Segmentation results on the test set for skin cancer detection. We extend the table presented by [16] with a few more results [18], [19], including ours. The results on the FCN and U-Net are reported from [16] and have been trained on data pre-processed using their strategy.

[2] Olaf Ronneberger, Philipp Fischer, and Thomas Brox, "U-net: Convolutional networks for biomedical image segmentation," *CoRR*, vol. abs/1505.04597, 2015.

[3] Joseph Redmon, Santosh Kumar Divvala, Ross B. Girshick, and Ali Farhadi, "You only look once: Unified, real-time object detection," *CoRR*, vol. abs/1506.02640, 2015.

[4] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," *CoRR*, vol. abs/1512.03385, 2015.

[5] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Identity mappings in deep residual networks," in *European conference on computer vision*. Springer, 2016, pp. 630–645.

[6] Kan Chen, Jiang Wang, Liang-Chieh Chen, Haoyuan Gao, Wei Xu, and Ram Nevatia, "ABC-CNN: an attention based convolutional neural network for visual question answering," *CoRR*, vol. abs/1511.05960, 2015.

[7] Joon Son Chung, Andrew W. Senior, Oriol Vinyals, and Andrew Zisserman, "Lip reading sentences in the wild," *CoRR*, vol. abs/1611.05358, 2016.

[8] Jie Hu, Li Shen, and Gang Sun, "Squeeze-and-excitation networks," *arXiv preprint arXiv:1709.01507*, vol. 7, 2017.

[9] Stephen M. Pizer, E. Philip Amburn, John D. Austin, Robert Cromartie, Ari Geselowitz, Trey Greer, Bart Ter Haar Romeny, and John B. Zimmerman, "Adaptive histogram equalization and its variations," *Comput. Vision Graph. Image Process.*, vol. 39, no. 3, pp. 355–368, Sept. 1987.

[10] H. Chen, Y. Zhang, W. Zhang, P. Liao, K. Li, J. Zhou, and G. Wang, "Low-dose ct denoising with convolutional neural network," in *2017 IEEE 14th International Symposium on Biomedical Imaging (ISBI 2017)*, April 2017, pp. 143–146.

[11] Karel Zuiderveld, "Graphics gems iv," chapter Contrast Limited Adaptive Histogram Equalization, pp. 474–485. Academic Press Professional, Inc., San Diego, CA, USA, 1994.

[12] Noel C. F. Codella, David Gutman, M. Emre Celebi, Brian Helba, Michael A. Marchetti, Stephen W. Dusza, Aadi Kalloo, Konstantinos Liopyris, Nabin K. Mishra, Harald Kittler, and Allan Halpern, "Skin lesion analysis toward melanoma detection: A challenge at the 2017 international symposium on biomedical imaging (isbi), hosted by the international skin imaging collaboration (ISIC)," *CoRR*, vol. abs/1710.05006, 2017.

[13] Kevin Mader, "Finding and measuring lungs in ct data," https://www.kaggle.com/kmader/finding-lungs-in-ct-data/home.

[14] François Chollet et al., "Keras," https://keras.io, 2015.

[15] Md. Zahangir Alom, Mahmudul Hasan, Chris Yakopcic, Tarek M. Taha, and Vijayan K. Asari, "Recurrent residual convolutional neural network based on u-net (r2u-net) for medical image segmentation," *CoRR*, vol. abs/1802.06955, 2018.

[16] Yuexiang Li and Linlin Shen, "Skin lesion analysis towards melanoma detection using deep learning network," *CoRR*, vol. abs/1703.00577, 2017.

[17] Hongdiao Wen, "II-FCN for skin lesion analysis towards melanoma detection," *CoRR*, vol. abs/1702.08699, 2017.

[18] Mohamed Attia, Mustafa Hossny, Saeid Nahavandi, and Anousha Yazdabadi, "Spatially aware melanoma segmentation using hybrid deep learning techniques," *CoRR*, vol. abs/1702.07963, 2017.

[19] L. T. Thao and N. H. Quang, "Automatic skin lesion analysis towards melanoma detection," in *2017 21st Asia Pacific Symposium on Intelligent and Evolutionary Systems (IES)*, Nov 2017, pp. 106–111.