Learning To Pay Attention To Mistakes

Mou-Cheng Xu
moucheng.xu.18@ucl.ac.uk
Neil P. Oxtoby
n.oxtoby@ucl.ac.uk
Daniel C. Alexander
d.alexander@ucl.ac.uk
Joseph Jacob
j.jacob@ucl.ac.uk

Centre For Medical Image Computing
Department of Computer Science
Department of Medical Physics &
Biomedical Engineering
University College London
London, UK

Abstract

In convolutional neural network based medical image segmentation, the periphery of foreground regions representing malignant tissues may be disproportionately assigned as belonging to the background class of healthy tissues [18][20][24][12][9]. Misclassification of foreground pixels as the background class can lead to high false negative detection rates. In this paper, we propose a novel attention mechanism to directly address such high false negative rates, called Paying Attention to Mistakes. Our attention mechanism steers the models towards false positive identification, which counters the existing bias towards false negatives. The proposed mechanism has two complementary implementations: (a) "explicit" steering of the model to attend to a larger Effective Receptive Field on the foreground areas; (b) "implicit" steering towards false positives, by attending to a smaller Effective Receptive Field on the background areas. We validated our methods on three tasks: 1) binary dense prediction between vehicles and the background using CityScapes; 2) Enhanced Tumour Core segmentation with multi-modal MRI scans in BRATS2018; 3) segmenting stroke lesions using ultrasound images in ISLES2018. We compared our methods with state-of-the-art attention mechanisms in medical imaging, including self-attention, spatial-attention and spatial-channel mixed attention. Across all of the three different tasks, our models consistently outperform the baseline models in Intersection over Union (IoU) and/or Hausdorff Distance (HD). For instance, in the second task, the "explicit" implementation of our mechanism reduces the HD of the best baseline by more than 26%, whilst improving the IoU by more than 3%. We believe our proposed attention mechanism can benefit a wide range of medical and computer vision tasks, which suffer from over-detection of background.

1 Introduction

Convolutional Neural Networks (CNN) enhanced by attention mechanisms have recently been transferred from computer vision to medical image analysis to tackle segmentation tasks [20, 21]. Attention mechanisms aim to focus learning on salient regions of interest (RoI), i.e. foreground pixels in medical images, to minimise the misclassification of RoI. Attention is implemented through a normalisation step in the latent feature space to focus the

network on the RoI. However, foreground pixels in medical images are often heavily under-represented. The resulting bias towards detection of background areas causes an under-detection of foreground pixels (True Positives: TPs) and an over-detection of background pixels (False Negatives: FNs), especially at the edges of a RoI (see visual results in [4, 12, 18, 21, 24]). It is apparent that reducing FN detection is a key challenge to increasing the utility of CNN based segmentation.

Attention is a natural mechanism by which TP detection can be improved. Yet existing attention mechanisms do not adequately reduce FNs (see section 5). Existing attention mechanisms focus on the labelled foreground areas as the RoI (area within the black circle in **Figure** 1), which we refer to as *Paying Attention to True Positives (TPs)*. Ideally, a well-trained over-parameterized model is able to detect the entirety of the RoI, but the bias towards detecting background pixels leads to the area of focus shrinking (dotted black circle in **Figure** 1, Left). This is exacerbated by the optimisation towards local minimum solutions which results in further shrinkage of the regions identified as TPs (red circle in **Figure** 1, Left).

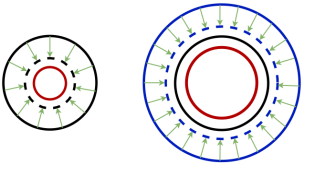

Figure 1: Left: Performance illustration of the existing attention mechanisms. Right: Performance illustration of our proposed solution to reduce FNs.

To overcome the high FN rate inherent to existing attention mechanisms, we propose an alternative strategy, which we refer to as *Paying Attention To Mistakes*. The hypothesis is that FNs can be reduced by encouraging a bias towards False Positives (FPs) detection, focused particularly around the boundary of TPs. This is shown conceptually in **Figure** 1 (Right), where our proposed attention mechanism learns to focus on an expanded RoI (blue solid circle). After shrinkage occurs, as explained in the last paragraph, resulting in more TPs, the red circle in Figure 1, Right is now more closely approximated to the black circle, representing the ground truth RoI. The black circles in Figure 1 (left,right) are identical in size. *Paying Attention To Mistakes* requires neither extra annotations of FPs, nor modifications to the source data distribution, which might risk information loss. Our main **contributions** are three-fold:

- We are the first to use an attention mechanism to ameliorate the pixel-wise classification bias towards false negatives in medical imaging.
- We are the first to develop an attention mechanism based on the Effective Receptive Field (ERF). This was implemented to make our attention mechanism "transparent". Our mechanism has two complimentary implementations to *Pay Attention To Mistakes* which are explicit and implicit respectively.
- We perform extensive experiments including comparisons with state-of-the-art baselines and ablation studies on different configurations on three different data sets.

2 Related works

Attention mechanisms. We review a few representative attention mechanisms that *Pay Attention To TPs* here. According to the focusing target (e.g. what to attend to or where to attend), most existing attention mechanisms can be divided into three groups: channel attention (e.g. importance of each channel of a feature map) [7], spatial attention (e.g. importance of each spatial location of a feature map) [10, 11, 25] and spatial-channel mixed attention (e.g. importance of each spatial location at each channel of a feature map) [23, 27]. Channel

attention is based on weighting each channel according to each channel's most representative feature (e.g. mean [7], maximum or both [27]). In spatial attention, self-attention mechanisms have been used as in [11, 25]; or semantic features from deep layers have been used as "keys" to enhance representation learning in shallow layers as in [10, 23].

Effective receptive field. In a layer of a CNN, the size of the region corresponding to the neuron in the next layer is called the "receptive field" (RF). The centre of the RF has the highest impact on the layer output. The impact of the pixels across the RF has been shown to resemble a Gaussian distribution. Therefore, in a forward pass, the gradient of the signal decays from the RF centre to the periphery in a squared exponential manner [14]. Accordingly, only a fraction of the RF is detected and contributes to the output. This effective area is called the "Effective RF" (ERF). The ERF also reduces [14] as the network goes deeper. More importantly, it has been shown that the size of an ERF is also influenced by neural network topology [14]. By combining the ERF with latent space embedding [28], we can achieve flexible control of the area over which we want the model to focus.

Other related work. Dilated convolutional layers [2] have been proposed to expand the receptive field in supervised image segmentation tasks. Work also has previously been proposed for reversed attention (RA) [3, 8] to improve classification accuracy in confusing regions. Our methods differ from related works in both motivation and implementation. We are motivated to use an attention mechanism to expand or shrink the focus on the RoI, to control detection biases at decision boundaries, whereas previous works are motivated to eliminate the bias. Regarding implementation, our approach is the first attention mechanism built upon the ERF.

3 Methods

Our research hypothesis is that the bias towards FNs detection in medical image segmentation could be mitigated by forcing networks to favour FPs detection using an attention mechanism. It is possible to "explicitly" shift the bias at decision boundaries towards detection of FPs, termed "False Positive Attention". It is also possible to reverse labels for foreground and background pixels and "implicitly" focus on detecting the foreground. The two candidates can result in complimentary implementations. We first present the implementation which "explicitly" shifts the bias towards detection of FPs.

3.1 False Positive Attention (FPA)

To "explicitly" shift the bias towards FPs, we use the strategy as explained in **Figure** 1. We expand the focus of the model beyond the area of labelled TPs, to extend into regions of FPs by applying a convolutional layer [2] to learn a larger smoothed ERF. The larger ERF extends from the original ERF, guided by a dilated convolutional layer at the same depth. The rationale behind the architecture of FPA is based on one empirical result in [14]: at the same depth, the ERF of the dilated convolutional layer is larger than the ERF of the original convolutional layer. To merge the information contained in the regions surrounding the ERF to the ERF, it would be intuitive to use the mean of the the outputs of the convolutional and dilated convolutional layers. However this mean calculation eliminates the bias towards FP detection [26]. We therefore use a Sigmoid function on the output of the dilated convolutional layer to create a smoothed larger ERF, before performing an element-wise multiplication on the output of the corresponding convolutional layer.

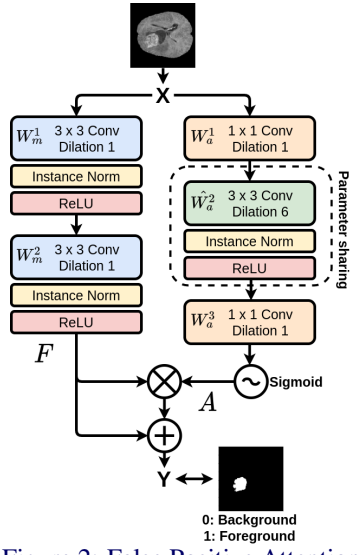

Figure 2: False Positive Attention

Hence, our approach whereby we shift the bias of the model towards FP detection represents an attention mechanism. We now describe the operation of FPA. The FPA module consists of two parallel branches: the main branch processes visual information, whereas the attention branch generates the smoothed enlarged ERF. Given the input feature map $X \in R^{C\times H\times W}$, the output feature map of main branch is $F \in R^{2C\times H/2\times W/2}$ and the output attention weights of the attention branch is $A \in R^{2C\times H/2\times W/2}$. We achieve the output feature map ($Y \in R^{2C\times H/2\times W/2}$) following the equations:

$$\begin{aligned} F &= W_m^2[W_m^1(X)] \\ A &= \sigma(W_a^3\{\hat{W}_a^2[W_a^1(X)]\}) \\ Y &= F \odot A + F \end{aligned} \tag{1}$$

Where W_m^i denotes the i^{th} standard convolutional layer in the main branch; $\hat{W}_a^i$ denotes the i^{th} dilated convolutional layer in the attention branch; W_a^i denotes the i^{th} standard convolutional layer in the attention branch; σ is Sigmoid function; $\odot$ is element-wise multiplication. All of the Non-linear activation and normalisation layers are omitted for simplicity of expression. The detailed architectures can be found in **Figure** 2.

3.2 Reverse False Negative Attention (RFNA)

On the contrary to FPA which directly favours FP detection, we present an alternative approach to reduce over-detection of background by favouring "reverse FN detection". It is obvious that if we directly bias towards FN, we risk in deteriorating the performance which already suffers from high FN rate. Our solution is to "implicitly" *Pay Attention to Mistakes*. In binary segmentation, the labels values normally are: 1 for foreground and 0 for background, and the models would naturally favour FNs detection (section 1) where foreground pixels are classified as label value 0. Our implementation first reverses the labels values whereby the models would naturally favour reverse FP detection, which is classification of foreground with a label value of 1. We now encourage the models to focus less on reverse FP detection, which is equivalent to biasing towards reverse FNs detection. This bias towards reverse FNs in RFNA share the same goal with the bias towards FPs in FPA, which is to reduce false detection of foreground areas as background class, albeit RFNA reduces over-detection of background in an "implicit" way.

As this implementation favours reverse FNs detection, it is termed "Reverse False Negative Attention (RFNA)".The RFNA module has two branches but these are different from the parallel architecture in FPA. The attention branch in RFNA is placed sequentially after the main branch. The rationale behind the architecture of RFNA is based on another empirical observation in [14]: the RF increases linearly with network depth, while the ratio between the ERF and the RF rapidly decreases. This phenomenon leads to a situation where a proximal deeper layer might have a smaller ERF than a shallower layer. For instance, the 40th convolutional layer would have a smaller ERF than the 20th layer as shown in **Figure 1** in

[14]. Additionally, it was found in [14] that residual connections also generate a smaller ERF.

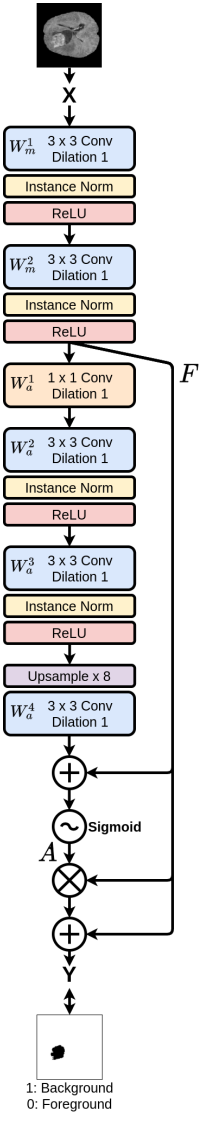

Figure 3: Reverse False Negative Attention

We combine these two characteristics to generate a smaller ERF to guide the networks to shrink the focus on reverse FP. We use the same notations from FPA to describe the operation of RFNA as follows:

$$\begin{aligned} F &= W_m^2[W_m^1(X)] \\ A &= \sigma\{W_a^4[UpSample(W_a^3\{W_a^2[W_a^1(F)]\})] + F\} \\ Y &= F \odot A + F \end{aligned} \tag{2}$$

Where $UpSample$ denotes a bilinear upsampling layer, the default upsampling ratio is 8 (**Figure** 3). We use the output of the main branch as the input for the attention branch to make sure that we use deeper features, leading to a smaller ERF. There is an additional residual connection before the Sigmoid function, which aids generation of a smaller ERF. The architecture of RFNA is in **Figure** 3.

3.3 Implementation

We enhance a 3 down-sampling stage, 2D U-net [19] as our backbone, by replacing the batch normalisation with instance normalisation [22] and replacing deconvolutional layers with bilinear upsampling layers. We replace convolutional blocks in the encoder in U-net with the proposed attention modules. The channel number in the first encoder in the backbone is 32, as we found the original number 64 is redundant for our tasks. In FPA, $\hat{W}_a^2$ is repeated twice via parameter sharing. In RFNA, W_a^1 expands the channel number to 4 times of the channel number of output; W_a^2 and W_a^3 use depth-wise convolution layers for computational efficiency. Ablation studies on different configurations of these components are outlined in a later section 4.3.

4 Experiments

4.1 Materials

CityScape Our first task is a synthetic medical task using high-resolution RGB images from CityScapes [5]. We perform a binary segmentation task to distinguish between vehicles and the background, replicating medical image segmentation tasks which are typically dense binary predictions. Due to computational restrictions, we downsample images to 256x128. We use cities from "train" and "val" for training and testing, respectively. We hold one city for validation in training data.

BRATS The publicly available BRATS 2018 Training data [15] has 210 High Grade Glioma (HGG) cases and 76 Low Grade Glioma (LGG) cases. Since the enhancing tumour

core is the hardest class to discern [9, 17], we focus on this class to illustrate the effectiveness of our methods. We noticed that LGG cases contain almost no enhancing tumour cores [9]. Therefore to focus on assessing the network, we only analysed the HGG cases. Pre-processing steps included: normalisation of each case for each modality; centre cropping to 160x160; concatenation of each modality as 4D input. 5 fold cross-validation with a case-wise split.

ISLES The ISLES2018 [6] training data contains 94 acute stroke CT perfusion scans. We randomly split the cases at the ratio of 0.7:0.1:0.2 for training, validation and testing. We use CBF, MTT, CBV, TMAX modalities, and normalise each case in each modality and centre crop them to 224x224. Our task was to segment the stroke lesions in the CT brain images.

4.2 Baselines

We compare our models with existing attention mechanisms in medical imaging, which *Pay Attention To TPs*. We follow the categories in section 2 to select our baselines. For spatial attention, we use the state-of-the-art "Attention U-net" [18], denoted as "AUnet". For self-attention, we use a state-of-the-art efficient non-local module "GCN" [1], to avoid the high computational burden of the original self-attention module [25] and we integrate them into the backbone as "GCNUnet". The backbone is described in section 3.3.

Table 1: Where the convolutional blocks are replaced with attention modules in baselines and our networks. Type: attention category.

Networks	Type	Encoder	Decoder
AUnet	spatial		✓
CSCUnet	mixed	✓	✓
CBAMUnet	mixed	✓	
GCNUnet	self	✓	
FPA/RFNA	mixed	✓	

For spatial-channel mixed attention, we first use the state-of-the-art mixed attention U-net called "concurrent spatial and channel squeeze-excitation U-net" [20], this is denoted as "CSCUnet". We also include another state-of-the-art mixed attention mechanism, namely "CBAM" [27] and we also implement them into the backbone as "CBAMUnet". No channel attention is included, as "CSC", "CBAM", "GCN" already comprise the state-of-the-art channel attention [7]. In implementation of baselines, the convolutional blocks in the backbone are replaced with different existing attention modules, either following the references or in the encoder as we insert our FPA/RFNA in the encoder, see details in **Table** 1.

4.3 Ablation Studies

Effect of attention branches We study the effect of our attention modules by pruning the branches in FPA/RFNA. As shown in **Figure** 2, without the main branch, we have a U-net with dilated convolutional layers in the encoder. We replace the convolutional layers in the encoder in the backbone with dilated convolutional layers as "D6/9 Unet" (6 or 9 is the dilation rate). Without the attention branch in **Figure** 2, it becomes the backbone Unet. Similarly, we also have Unet without the attention branch in RFNA.

Effect of downsampling ratio in RFNA We prune the convolutoinal layers in RFNA to study how the downsampling ratio in the attention branch influences performance. We first remove W_a^3 and use $UpSample \times 4$ to see the effect of a downsampling ratio of 4. Then we further remove W_a^2 and use $UpSample \times 2$ to check the effect of a downsampling ratio of 2.

Effect of dilation rate in FPA The dilation rate in FPA is an important hyper-parameter, so we compare different dilation ratios at 6, 9 and 12. No smaller dilation is used as we found that a dilation ratio 3 has no obvious effect on segmentation results.

Effect of model capacity To examine whether the effect of our attention modules was a consequence of accumulating more parameters, we double the channel number in the backbone to make a wide U-net, denoted as "WUnet".

Effect of channel numbers We study the impact of model capacity in both FPA and RFNA. For each configuration in FPA with a different dilation rate and each configuration in RFNA with a different downsampling ratio, two variants using depth-wise convolutional layers and no depth-wise convolutional layers in attention branches are implemented. For the variant of using depth-wise convolutional layers, we explore the impact of channel expansion ratio at 2, 4 and 8 of output channel number, in W_a^1. For the variant of using no depth-wise convolutional layers, we explore the impact of channel expansion ratio at 1 and 2, due to computational restriction. In default (section 3.3), FPA uses no depth-wise convolutional layers and RFNA uses depth-wise convolutional layers.

4.4 Training

AdamW optimiser [13] is used for optimisation. Dice Loss [16] is used as objective function. Random horizontal flipping is used for augmentation. Training details are in **Table** 2. No further improvements were observed with more training epochs in our settings. All experiments were run for at least 3 times on a NVIDIA TITAN V GPU. Our demo code is implemented in Pytorch 1.0 and it is available in:https://github.com/moucheng2017/Pay_Attention_To_Mistakes.

Table 2: Training details. lr: learning rate.

Dataset	Epoch	Batch	lr
CityScapes	100	8	2e-4
BRATS	80	50	1e-4
ISLES	60	80	1e-3

5 Results

Table 3: Results on CityScapes.

Networks	IoU (%)	HD
AUnet	54.17 ± 0.15	64.69 ± 0.67
CSCUnet	54.45 ± 0.31	60.15 ± 1.96
RFNA	55.62 ± 0.45	65.18 ± 3.36
FPA	**59.39 ± 0.65**	**49.10 ± 2.51**

Table 4: Results on ISLES2018.

Networks	IoU (%)	HD
Unet	52.41 ± 0.78	25.49 ± 1.38
AUnet	52.35 ± 0.63	23.66 ± 0.87
CSCUnet	52.70 ± 0.69	27.28 ± 0.72
RFNA	52.99 ± 1.63	**18.10 ± 4.75**
FPA	**54.61 ± 0.54**	19.44 ± 1.41

Table 5: Results on BRATS2018. HD: Hausdorff distance at 95% percentile. FP: False Positive Rate. FN: False Negative Rate. Param: Parameters.

Networks	IoU (%)	HD	FP (%)	FN (%)	Param (M)
AUnet	66.48 ± 1.04	15.51 ± 0.82	<1e-2	58.42 ± 2.58	3.16
CSCUnet	66.75 ± 1.00	15.50 ± 1.79	<1e-2	58.11 ± 3.28	3.14
GCNUnet	66.57 ± 1.20	15.11 ± 1.05	<1e-2	58.16 ± 2.19	3.15
CBAMUnet	66.52 ± 1.50	15.19 ± 1.59	<1e-2	58.68 ± 3.11	3.14
RFNA	**71.15 ± 1.78**	21.76 ± 1.08	<1e-2	50.15 ± 2.94	**5.63**
FPA	70.10 ± 2.06	**10.86 ± 1.66**	<1e-2	**49.39 ± 4.56**	4.2

Our *Paying Attention To Mistakes* outperforms all of the baselines across three different multi-modal data sets in either Intersection over Union (IoU) or Hausdorff Distance (HD) or both. In the first task, RFNA improves the best baseline by 1.17% in IoU; whereas FPA improves the IoU by 4.94% and reduces the Hausdorff Distance (HD) of the best baseline by 18%. In the second task, the FPA reduces the HD of the best baseline by 28.12%. While the RFNA achieves the highest IoU score, with a 4.4% margin compared to the best baseline. In the third task, despite the fact that perfusion CT images are challenging to interpret, both FPA and RFNA outperform the baselines in both accuracy metrics. RFNA reduces the HD by 23.49% of the best baseline and FPA has a 2.26% gain in IoU.

FPA consistently improved performance in both accuracy metrics across all three tasks. Although RFNA also consistently improved performance as measured by IoU, it has mixed impacts in HD. HD and IoU are good at measuring different types of mistakes, suggesting that FPA and RFNA actually refine the segmentation in different ways. To evaluate this further we visualised the attention maps in FPA and RFNA of the foreground class in the deepest encoder in **Figure** 4. As demonstrated in **Figure** 4, FPA and RFNA utilise global and local spatial information, respectively. FPA and RFNA also experience different levels of bias in segmentation outcomes due to label reversal. Also, the ratios between background and foreground are different in each data set. Eventually, the bias differences in the data plus the mechanistic differences led to different performances between FPA and RFNA.

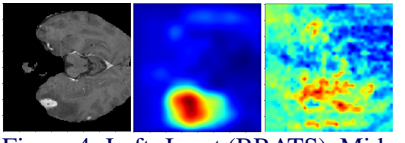

Figure 4: Left: Input (BRATS). Middle: RFNA. Right: FPA.

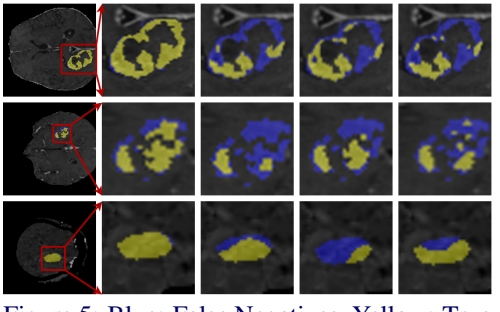

Figure 5: Blue: False Negatives, Yellow: True Positives. Row 1&2: BRATS; row3: ISLES. 1st column: FPA, 2nd column: FPA (Zoomed in). 3rd column: AUNet, 4th column: CS-CUnet, 5th column: UNet.

The inferior performance of the baselines might be a result of their focus on TP regions, as discussed in section 1. By *Paying Attention To Mistakes*, we successively reduce FN detection as qualitatively shown in **FN** column in **Table** 5, where FPA reduces the FN rate of the best baseline by 15%. The reduction of FN detection can also be seen in **Figure** 5 and **Figure** 6.

5.1 Results on ablation studies

Both FPA/RFNA in **Table** 5 performed better than backbone "Unet" in **Table** 6, suggesting that attention branches improve the performance of the backbone. We find that the combination of an attention branch and a main branch in FPA outperforms the use of an attention branch alone, as both "D6Unet" and "D9Unet" in **Table** 6 performed worse than FPA. We also show that the positive effect of our attention mechanisms is not just from a larger number of parameters, as "WUnet" (**Table** 6) achieves a lower performance despite having three times more parameters than FPA.

We found that in RFNA, a higher downsampling ratio consistently lead to better segmentation performance ((e) and (h) in **Figure**7). However, the dilation rate in FPA only resulted in better performance when used with depth-wise convolutional layers ((a) and (d) in **Figure**7). No clear relationships between channel expansion ratio and performance in either FPA or RFNA were observed, which might suggest that network architectures are more influential on performance than the parameters numbers.

Figure 6: Blue: False Negatives, Yellow: True Positives. Visual results on Cityscapes. 1st column: FPA, 2nd column: RFNA, 3rd column: AUnet, 4th column: CSCUNet.

Table 6: Ablation studies of removing branches in FPA/RFNA on BRATS2018.

Networks	IoU (%)	HD	FP (%)	FN (%)	Param (M)
Unet	66.67 ± 1.15	**14.77 ± 1.61**	<1e-2	57.83 ± 2.25	3.13
WUnet	**66.89 ± 1.07**	15.40 ± 0.87	<1e-2	**57.66 ± 3.20**	**12.51**
D6Unet	66.42 ± 1.14	15.174 ± 1.92	<1e-2	58.52 ± 2.68	3.13
D9Unet	66.63 ± 1.41	15.61 ± 1.08	<1e-2	58.07 ± 3.79	3.13

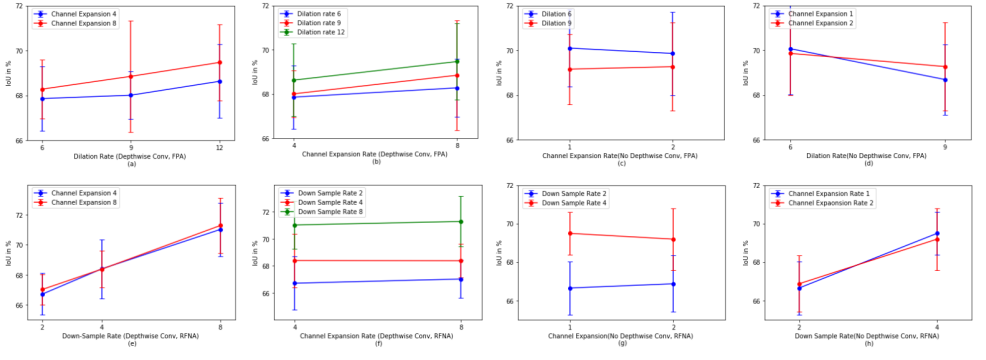

Figure 7: Ablation studies of different configurations of FPA/RFNA on BRATS2018.

6 Conclusion

In this paper, we present *Pay Attention To Mistakes* to tackle the high FNs detection rate in medical image segmentation. Our method effectively reduces the FN rate of backbone and achieves superior performance compared to existing state-of-the-art attention mechanisms in medical image segmentation across three different data sets. Although both the FPA and RFNA implementations are shown to be effective on tasks suffering from over-detection of FNs, the use of RFNA on tasks suffering from over-detection of FPs will require exploration in future work. By flexibly apply FPA and RFNA we could potentially cover most of the situations where over-detection of FNs and over-detection of FPs happen in both medical and computer vision domains.

7 Acknowledgement

Mou-Cheng is supported by GSK funding (BIDS3000034123) via UCL EPSRC CDT in i4health and UCL Engineering Dean's Prize.

Neil is a UKRI Future Leaders Fellow (MR/S03546X/1) supported by the National Institute for Health Research University College London Hospitals Biomedical Research Centre.

We thank NVIDIA for hardware donation. We also thank Fred Wilson from GSK; Eyjolfur Gdmundsson and Yi-Peng Hu from UCL CMIC for their feedback on the draft.

References

[1] Yue Cao, Jiarui Xu, Stephen Lin, Fangyun Wei, and Han Hu. Gcnet: Non-local networks meet squeeze-excitation networks and beyond. *ICCV workshop*, 2019.

[2] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2016.

[3] Shuhan Chen, Xiuli Tan, Ben Wang, and Xuelong Hu. Reverse attention for salient object detection. *ECCV*, 2018.

[4] Xu Chen, Bryan M. Williams, Srinivasa R. Vallabhaneni1, and et al. Learning active contour models for medical image segmentation. *CVPR*, 2019.

[5] Marius Cordts, Mohamed Omran, Sebastian Ramos, and et al. The cityscapes dataset for semantic urban scene understanding. *CVPR*, 2016.

[6] Arsany Hakim, Mauricio Reyes, Roland Wiest, Maarten G Lansberg, Søren Christensen, Greg Zaharchuk, Stefan Winzeck, David Robben, and Christian Lucas. Ischemic stroke lesion segmentation challenge 2018. *http://www.isles-challenge.org/*, 2018.

[7] Jie Hu, Li Shen, Samuel Albanie, Gang Sun, and Enhua Wu. Squeeze-and-excitation networks. *CVPR*, 2018.

[8] Qin Huang, Chunyang Xia, Wuchi Hao, SIyang Li, Ye Wang, Yuhang Song, and C Jay Kuo. Semantic segmentation with reverse attention. *BMVC*, 2017.

[9] Fabian Isensee, Phillpp Kickngereder, Wolfgang Wick, Martin Bendszus, and Klaus H. Maier-Hein. No new net. *International MICCAI Brainlesion Workshop in MICCAI*, 2018.

[10] Saumya Jetley, Nicholas A. Lord, Namhoon Lee, and Philip H.S. Torr. Learn to pay attention. *ICLR*, 2018.

[11] Tsung-Yu Lin, Aruni RoyChowdhury, and Subhransu Maji. Bilinear cnn models for fine-grained visual recognition. *ICCV*, 2015.

[12] Michelle Livne, Jana Rieger, Orhun Utku Aydin, and et al. A u-net deep learning framework for high performance vessel segmentation in patients with cerebrovascular disease. *Frontier Neuroscience*, 2019.

[13] Ilya Loshchilov and Frank Hutter. Decoupled weigth decay regularization. *ICLR*, 2019.

[14] Wenjie Luo, Yujia Li, Raquel Urtasun, and Richard Zemel. Understanding the effective receptive field in deep convolutional neural networks. *NeurIPS*, 2016.

[15] Bjoern H. Menze, Andras Jakab, Stefan Bauer, Jayashree Kalpathy-Cramer, Keyvan Farahani, Justin Kirby, Yuliya Burren, and et al. The multimodal brain tumor image segmentation benchmark (brats). *IEEE Transactions on Medical Imaging*, 2015.

[16] Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation. *3DV*, 2016.

[17] Andriy Myronenko. 3d mri brain tumor segmentation using autoencoder regularization. *International MICCAI Brainlesion Workshop in MICCAI*, 2018.

[18] 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. *MIDL*, 2018.

[19] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. *MICCAI*, 2015.

[20] Abhijit Guha Roy, Nassir Navab, and Christian Wachinger. Concurrent spatial and channel 'squeeze excitation' in fully convolutional networks. *MICCAI*, 2018.

[21] Jo Schlemper, Ozan Oktay, Michiel Schaap, Mattias Heinrich, Bernhard Kainz, Kazunari Misawa, Ben Glocker, and Daniel Rueckert. Attention gated networks: Learning to leverage salient regions in medical images. *Medical Image Analysis*, 2019.

[22] Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. Instance normalization: The missing ingredient for fast stylization. *arXiv:1701.02096*, 2017.

[23] Fei Wang, Mengqing Jiang, Chen Qian, Shuo Yang, Cheng Li, Honggang Zhang, Xiaogang Wang, and Xiaoou Tang. Residual attention network for image classification. *CVPR*, 2017.

[24] Guotai Wang, Wenqi Li, Maria A. Zuluaga, and et al. Interactive medical image segmentation using deep learning with image-specific fine tuning. *IEEE Transactions on Medical Imaging*, 2018.

[25] Xiaolong Wang, Ross B. Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. *CVPR*, 2018.

[26] Yunchao Wei, Huaxin Xiao, Honghui Shi, Zequn Jie, Jiashi Feng, and Thomas S. Huang. Revisiting dilated convolution: A simple approach for weakly- and semisupervised semantic segmentation. *CVPR*, 2018.

[27] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. Cbam: Convolutional block attention module. *ECCV*, 2018.

[28] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. *CVPR*, 2016.