Attention U-net: Mitochondria Segmentation

Yijie Cao (yc3544), Yiran Wang (yw3201), Chaoqun Zhou (cz2514)

***Abstract*—The mitochondria play an important role in neural function. However, the work to manually label a huge amount of medical images is tedious and the accuracy cannot be sustained when people get tired. Hence, automated medical image segmentation to label the position of mitochondria in cerebral cortex section image has been explored here. Due to the limited number of images available, addictive attention gate is performed based on a standard U-net neural network with augmentation techniques. With this attention u-net, training for a small number of data sets is achieved by flipping, rotating, zooming, shifting and elastically distorting available images to create more data sets. The attention gate used allow u-net accurately predict position of mitochondria without excessive and redundant used of computational resources and model parameters.**

***Index Terms*—Medical image segmentation, Attention U-net, Augmentation, Mitochondria**

I. Introduction

Since the work to manually label a huge amount of medical images is tedious and the accuracy cannot be sustained when people get tired, automated medical image segmentation attracts more and more people to explore in this field. In order to increase clinical work efficiency, help decision making, and relive valuable clinic employees from this tedious work, high accuracy and reliable solutions that can extract useful information automatically and fast are desired.

Mitochondria are important for supplying cellular energy and many essential cellular tasks such as signaling, cellular differentiation, and cell death, as well as maintaining control of the cell cycle and cell growth. [1] Tons of studies have shown that localization and morphology of mitochondria are closely related to neural functionality. For example, pre- and post-synaptic presence of mitochondria is proven to have an crucial character in synaptic function. Also, it is highly possible that a link between mitochondrial defects and neuro-degenerative diseases exists. However, electron microscopy (EM), with higher resolution, provides plenty of other information the same time as it provides structure of mitochondria. To analysis such an image manually can require months of tedious labelling process. Hence, a reliable automated image segmentation is explored here to make labeling of mitochondria fast and accurately.

Convolutional neural networks (CNNs), with high representation power, fast inference, and filter sharing properties, is the standard method for image segmentation. It has already been applied in a lot of automated medical image analysis tasks including cardiac MR segmentation, cancerous lung nodule detection, CT pancreas segmentation and so on.Since CT pancreas segmentation share same properties with the task here, it is used as the reference. Fully convolutional networks (FCNs) and the U-Net are two commonly used architectures in CNN. Here, U-net is chosen as it is designed for biomedical image segmentation with fewer training images but yielding more precise segmentation. One challenge exists in our EM image: it is hard to distinguish the cell nuclear and the mitochondria. In order to learn a higher weight for the mitochondria's mask and provide a smoother border, attention gates (AGs) is added into the original U-net. AGs automatically learn to focus on target structures, and it avoids excessive and redundant use of computational resources and model parameters. At test time, these gates generate soft region proposals implicitly on-the-fly and highlight salient features useful for a specific task without significant computational overhead and a large number of model parameters. In return, AGs improve model sensitivity and accuracy for dense label predictions by suppressing feature activations in irrelevant regions.[2]

II. Materials and Methods

A. Datasets

The data used is from Lucci laboratory. The data-set contains 330 raw images as shown in Figure 1. They are separated into half for train and test purposes. Each image is a cerebral cortex section acquired by using Electron Microscopy and it has plenty of different organelles including the mitochondria and so on. Those raw images also have manually labeled images respectively which segment where the mitochondria is and highlight the mitochondria with white color. Figure 2 is an example of the labeled image.

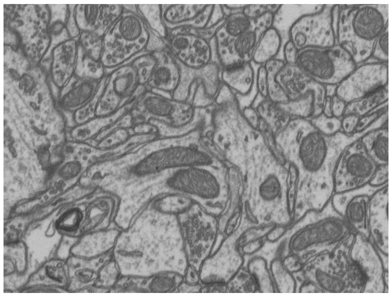

Fig. 1: Example of raw image

B. Data Pre-processing

The raw images first go though a gamma correction process to control the overall brightness of those images. It is defined by the following power-law expression:

$$V_{out} = AV_{in}^{\gamma} \tag{1}$$

where the non-negative real input value V_{in} is raised to the power γ, which is chosen to be 0.5 in the experiment and multiplied by the constant A. Then, the contrast of those images are increased with histogram equalization which adjust

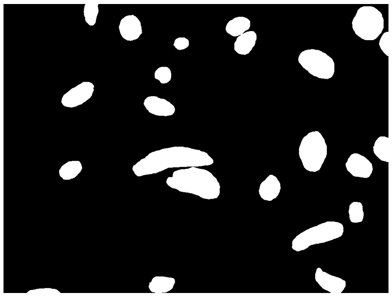

Fig. 2: Example of labeled image

image intensities. An example of the image after gamma correction and histogram equalization is shown in Figure 3.

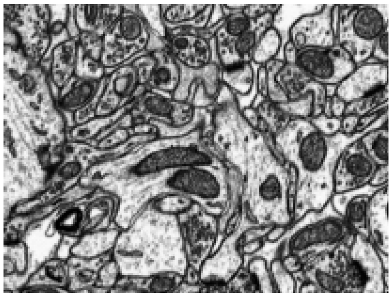

Fig. 3: Example of image after gamma correction and histogram equalization

C. Data Augmentation

Since the number of raw data sets is limited to 330 and having a large data set is crucial for a good performance, augmentation is introduced here in order to get more data. With the assumption that neural network would consider minor changed images as distinct images, flip, rotation, zoom, shift and elastic distortion methods are applied on the pre-processed images. Besides the benefit of increased number of data, the convolutional neural network's accuracy can be improved when facing variety of conditions in target application with this augmentation technique. An example of the image after this technique is shown in Figure 4.

D. U-net

The U-net architecture is illustrated in Figure 5. It consists of a contracting path on left side and an expansive path on right side. The contracting path follows the typical architecture of a convolutional neural network. The original image input size is 768*1024 pixel and it is compressed into 768*768 pixel for U-net input. The U-net consists of the repeated application of two 3x3 convolutions (unpadded convolutions), each followed by a rectified linear unit (ReLU) and a 2x2 max pooling operation with stride 2 for downsampling. At each downsampling step, the number of feature channels is doubled. Every step in the expansive path consists of an upsampling of the feature map, which propagates context information to higher resolution

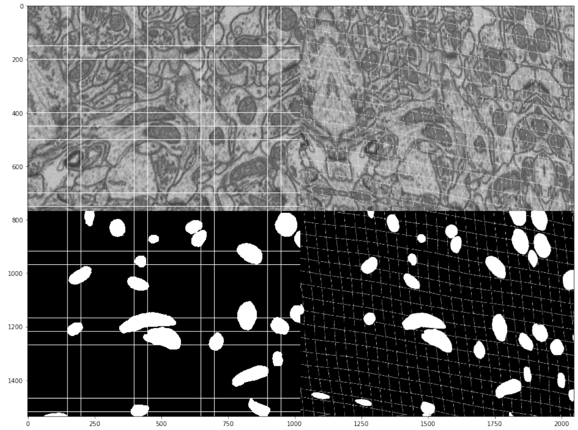

Fig. 4: Example of augmented image and its corresponding labeled image

layers, followed by a 2x2 convolution (up-convolution) that halves the number of feature channels, a concatenation with the correspondingly cropped feature map from the contracting path, and two 3x3 convolutions, each followed by a ReLU. The cropping is necessary due to the loss of border pixels in every convolution. At the final layer a 1x1 convolution is used to map each 64 component feature vector to the desired number of classes. In total the network has 23 convolutional layers. The output of predicted mask size (768*768 pixel) is the same as input dimension.

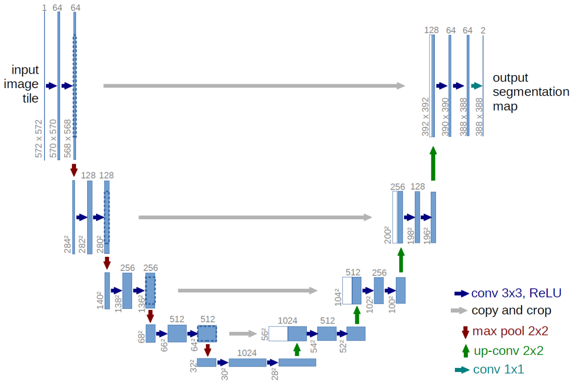

Fig. 5: Schematic of U-net architecture

The network is trained with stochastic gradient descent implementation of keras built-in Adam Optimizer. Since the output image is a constant border width smaller than the input caused by the unpadded convolutions, batch is reduced to a single image. Also, a high momentum (0.99)[3]

The first objective function chosen for the model is Binary Categorical Loss, which is defined as:

$$\text{Binary Loss} = -\frac{TP + TN}{TP + TN + FP + FN} \tag{2}$$

where TP is true positive, TN is true negative, FP is false positive and FN is false negative. Since the most of pixels in mask label are black (0), and around 89 percent accuracy is obtained even if all black is predicted, TP value rather than TN is of more interest here. Hence, The dice loss is used as shown below:

$$\text{Dice Loss} = -\text{Dice Score} = -\frac{2TP}{2TP + FP + FN} \tag{3}$$

E. Attention

The general architecture for attention u-net is shown in Figure 6. Attention gates filter the features propagated through the skip connections as illustrated in Figure 7.

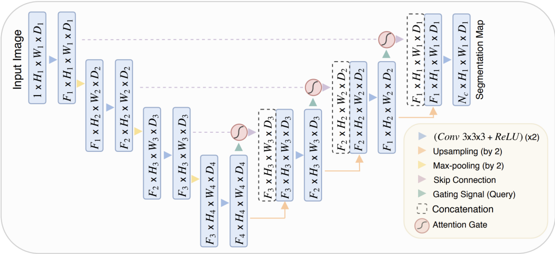
Input Image
$1 \times H_1 \times W_1 \times D_1$
$F_1 \times H_1 \times W_1 \times D_1$
$F_1 \times H_2 \times W_2 \times D_2$
$F_2 \times H_2 \times W_2 \times D_2$
$F_2 \times H_3 \times W_3 \times D_3$
$F_3 \times H_3 \times W_3 \times D_3$
$F_3 \times H_4 \times W_4 \times D_4$
$F_4 \times H_4 \times W_4 \times D_4$
$F_3 \times H_3 \times W_3 \times D_3$
$F_3 \times H_3 \times W_3 \times D_3$
$F_2 \times H_3 \times W_3 \times D_3$
$F_2 \times H_2 \times W_2 \times D_2$
$F_2 \times H_2 \times W_2 \times D_2$
$F_1 \times H_2 \times W_2 \times D_2$
$F_1 \times H_1 \times W_1 \times D_1$
$F_1 \times H_1 \times W_1 \times D_1$
$N_c \times H_1 \times W_1 \times D_1$
Segmentation Map
(Conv 3x3x3 + ReLU) (x2)
Upsampling (by 2)
Max-pooling (by 2)
Skip Connection
Gating Signal (Query)
Concatenation
Attention Gate

Fig. 6: Attention U-net architecture

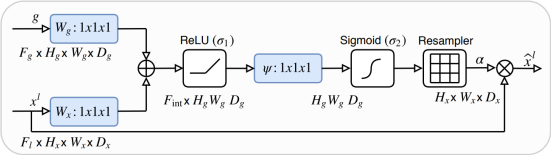
g
$W_g : 1x1x1$
$F_g \times H_g \times W_g \times D_g$
x^l
$W_x : 1x1x1$
$F_l \times H_x \times W_x \times D_x$
ReLU (σ_1)
$F_{\text{int}} \times H_g W_g\ D_g$
$\psi : 1x1x1$
Sigmoid (σ_2)
$H_g W_g\ D_g$
Resampler
$H_x \times W_x \times D_x$
α
$\hat{x}^l$

Fig. 7: Schematic of additive attention gate (AG)

Information extracted from coarse scale is used in gating to disambiguate irrelevant and noisy responses in skip connections. This is performed right before the concatenation operation to merge only relevant activations. Additionally, AGs filter the neuron activations during the forward pass and the backward pass. Gradients originating from background regions are down weighted during the backward pass. This allows model parameters in shallower layers to be updated mostly based on spatial regions that are relevant to a given task. The update rule for convolution parameters in layer l-1 can be calculated as:

$$\frac{\partial(\hat{x_i^l})}{\partial(\Phi^{l-1})} = \frac{\partial(\alpha_i^l f(x_i^{l-1}; \Phi^{l-1}))}{\partial(\Phi^{l-1})} \tag{4}$$

where α_i^l is the attention coefficient corresponding to the preservation of only relevant activations. [2]

III. Results

A. Train and Validation Loss

The train and validation loss plots for u-net and u-net with attention gate are shown in Figure 8 and Figure 9 respectively. The train loss of attention u-net model converges around 10 epoch which is more quickly than the standard u-net (around 20 epoch). However, the validation loss of standard u-net is smoother as illustrated in the plots.

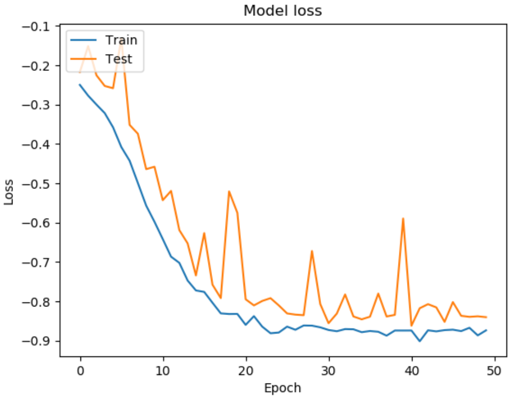

Fig. 8: Standard U-net train and validation loss

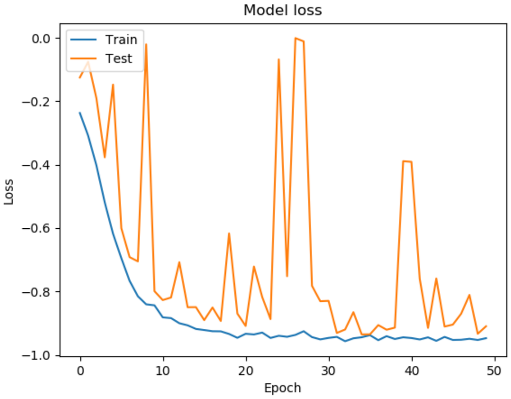

Fig. 9: Attention U-net train and validation loss

B. Quantitative Result

To quantitatively evaluate the U-net and attention U-net model developed, 80 images are randomly chosen as validation data set. The train and test deice scores are calculated according to Eqn 3. Black is considered as 0 and white is considered as 1. It is used to compare the similarity between the labeled mask and predicted mask. Also, test accuracy is calculated according to binary classification accuracy, which compares if each pixel of the predicted mask matches with the labeled mask.

TABLE I: Evaluation of U-net and Attention U-net model

	Train Dice Score	Test Accuracy	Test Dice Score
U-net	0.8820	0.9213	0.4043
Attention U-net	0.9474	0.9225	0.4206

As illustrated by table 1, attention u-net has an overall higher accuracy than the standard u-net model. The test accuracy of this attention u-net is 92.25%, which indicates that the predicted result matches the labeled mask pretty well.

C. Qualitative Result

The masks predicted by u-net model and attention u-net model are shown in Figure 10 and 11. The corresponding labeled mask is shown in Figure 12.

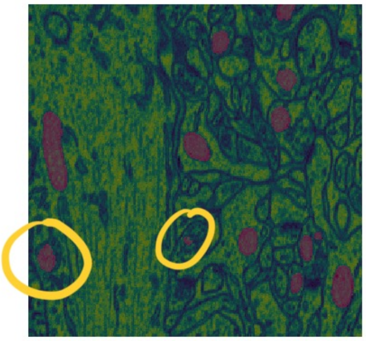

Fig. 10: U-net mask prediction

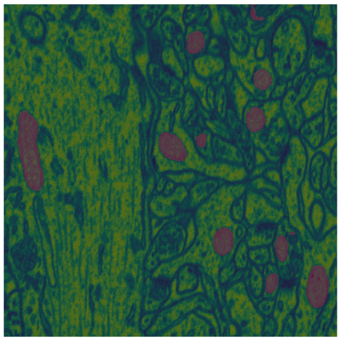

Fig. 11: Attention U-net mask prediction

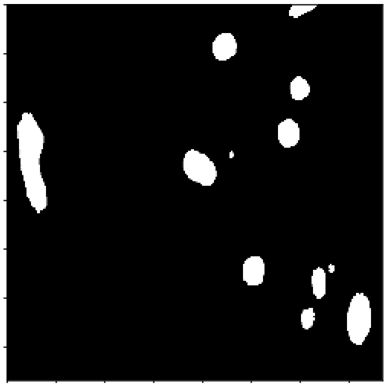

Fig. 12: U-net mask prediction

U-net model mistakenly considers two cell nuclear, circled in Figure 10, as mitochondria and a small mitochondria is missed in the segmentation. The attention u-net avoids highlighting irrelevant regions successfully and it finds all the mitochondria except a tiny region on the upper tight corner. Hence, attention u-net segments the image correctly and has a better performance than the standard u-net.

IV. Conclusion

Pre-process techniques are used to increase accuracy of prediction and augmentation is applied to increase number of available data set. Attention u-net model created correctly predicts position of mitochondria with a high accuracy and dice score. It suppresses irrelevant region and avoids considering other organelle as mitochondria correctly. With addictive attention gates, the model performs better than standard u-net model with a faster convergence in train loss, a higher train and test dice score and also a better test accuracy. However, the standard u-net has a smoother validation loss.

However, there still exits deficiencies in our prediction where some pixels are lost (marked) black in some mitochondrias. A further improvement would be utilizing superpixel image pre-process method by simple linear iterative clustering (SLIC) method, which produces smooth regular-sized superpixels in the smooth regions and highly irregular superpixels in the textured regions.

References

[1] N. M, “Mitochondria: more than just a powerhouse,” *Current Biology*, 2006.

[2] O. Oktay, J. Schlemper, L. Le Folgoc, M. Lee, M. Heinrich, K. Misawa, K. Mori, S. McDonagh, N. Y Hammerla, B. Kainz, B. Glocker, and D. Rueckert, “Attention U-Net: Learning Where to Look for the Pancreas,” *arXiv e-prints*, p. arXiv:1804.03999, Apr 2018.

[3] 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*, N. Navab, J. Hornegger, W. M. Wells, and A. F. Frangi, Eds. Cham: Springer International Publishing, 2015, pp. 234–241.