HIGH SCHOOL EDITION
Journal of Student Research

Carotid Intima-Media Thickness Segmentation using Attention Mechanism based Convolutional Neural Network with Domain-Specific Objective Function

Juheon Rhee[1] and Pyae Sone Kway[#]

[1]International School, Taguig, Metro Manila, Philippines
[#]Advisor

ABSTRACT

CIMT (Carotid Intima-Media Thickness) has been proven to be both a significant and reliable marker for the evaluation of the risk of cardiovascular disease. Cardiovascular disease is the leading cause of mortality globally, and yet could be easily treated if detected in its early stages. A prime indicator of Cardiovascular disease, CIMT has previously been measured through manual examination of ultrasound videos for the gap between the Lumen-Intima and the Media-Adventitia interfaces, the two inner layers of the Carotid Artery. However, this method is not only inconvenient, but also time consuming. There has been a significant number of previous deep learning approaches to this issue, which have yielded substantial results. However, as this problem concerns the morality of patients directly, medical professionals have been hesitant to be dependent on these approaches, as the current accuracy of the state-of-the-art model still falls short to human observations. Furthermore, high performing models come at high computational costs. CIMT can actually be determined by a miniscule region of the Carotid Ultrasonic image, which many past researches have not taken into consideration. This paper proposes to use an attention mechanism to determine the region of interest and an encoder-decoder system which significantly reduces computational trade off while maintaining comparable accuracy. We also propose a novel connection loss to solve the disconnection problem in the prediction. The proposed model yields an unprecedented accuracy in terms of IoU and ACC of 0.78 and 0.99 respectively, substantially higher than previous state-of-the-art models by 18% and 8.8% on average.

Introduction

The CIMT (Carotid Intima-Media Thickness) test measures the thickness between the inner and middle layers of the carotid artery, or the Lumen-Intima and the Media-Adventitia interfaces. CIMT is directly correlated with age, and increases approximately 0.005 to 0.01mm per year. For example, CIMT of healthy middle-aged adults has been measured to be between 0.6mm and 0.7mm while the same measurement has been reported to be approximately 0.4mm for young children and adolescents [1].

CIMT is a substantial indicator for atherosclerosis, which is the thickening and hardening of arteries due to the buildup of plaque in its inner lining. Furthermore, CIMT has been recorded to have a high correlation with cardiovascular disease, and the CIMT test has been acclaimed to be one of the rare, noninvasive methods for determining the risk of cardiovascular disease. Cardiovascular disease is currently the number one cause of mortality worldwide. Henceforth, CIMT testing has become an esteemed toll in both clinical trials and in the medical field in general. However, its limitation lies in that there is no optimal or set protocol in measuring the

CIMT. In the recent past, CIMT has been measured manually by medical professionals, which is not only prone to errors, but is also very time consuming.

Due to the significance of this measurement, researchers and scientists alike have proposed various deep-learning approaches to the CIMT measurement, many which have proved themselves to be both effective and efficient [2, 3]. These methods show that it is feasible to apply deep-learning techniques to measure CIMT. However, they came with a trade off: high computational costs. This is because while deeper networks generally show better performance than comparatively shallower networks, they also need much more computation. The performance of previous state-of-the-art models have been highly reliant on the depth of the convolutional neural network, which means that these models require a lot of computation to yield their top accuracies. To solve this problem, the model that the paper proposes utilizes an attention mechanism, which allows the model to focus more specifically to a significant region of the ultrasound input, and yield comparable results with a significantly lower computational cost.

Thus, in order to address this issue, we propose a novel attention-based CIMT segmentation system. The proposed system consists of two modules: the AttentionNet and the UNet module which consists of an encoder and a decoder. Given a carotid ultrasound input image, the encoder extracts the hierarchical features of the input image. The proposed AttentionNet also takes the sample input carotid image and produces the attention map which determines the region of interest that the decoder should focus on in order to yield better results. Finally, pixel-wise multiplication is applied on the two outputs of the AttentionNet and the encoder which is then fed to the decoder, which pixel-wise segments the Lumen-Intima and Media-Adventitia interfaces.

We also proposed a novel connection loss function. The connection loss function penalizes the model when the Lumen-Intima or Media-Adventitia Interface is not connected. It solves the disconnected Lumen-Intima and Media-Adventitia problem as it enforces the trained model to output and display a continuous line.

The proposed method yields an IoU metric of 0.78 and an ACC of 0.99, surpassing previous results by 18% and 8.8% on average. The main contributions of this paper is as follows: The proposed connection loss function and the utilization of an attention mechanism in a UNet (Encoder Decoder System). The employment of these two components allowed the model to display a continuous line to signify the two interfaces and also increased accuracy levels.

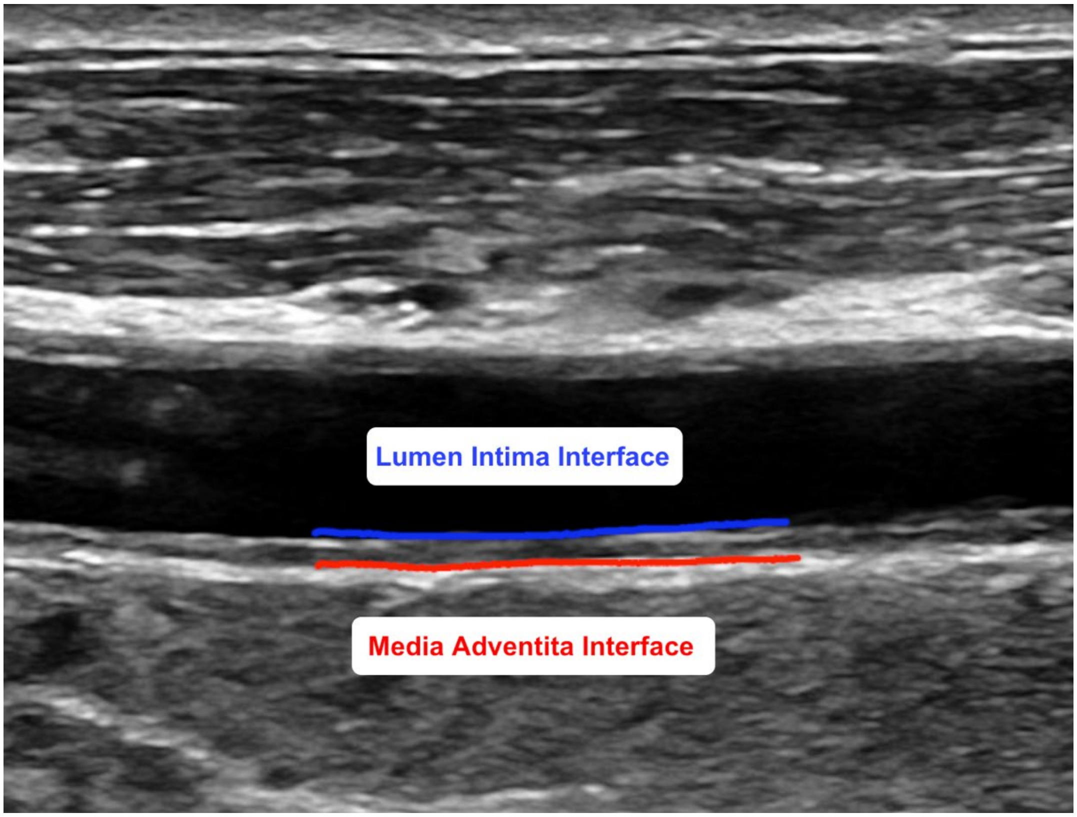


Figure 1. Example of Lumen-Intima and Media-Adventitia Interface in Carotid Ultrasound Image (Red line denotes the Media Adventitia Interface and Blue line denotes the Lumen Intima Interface)

Methods

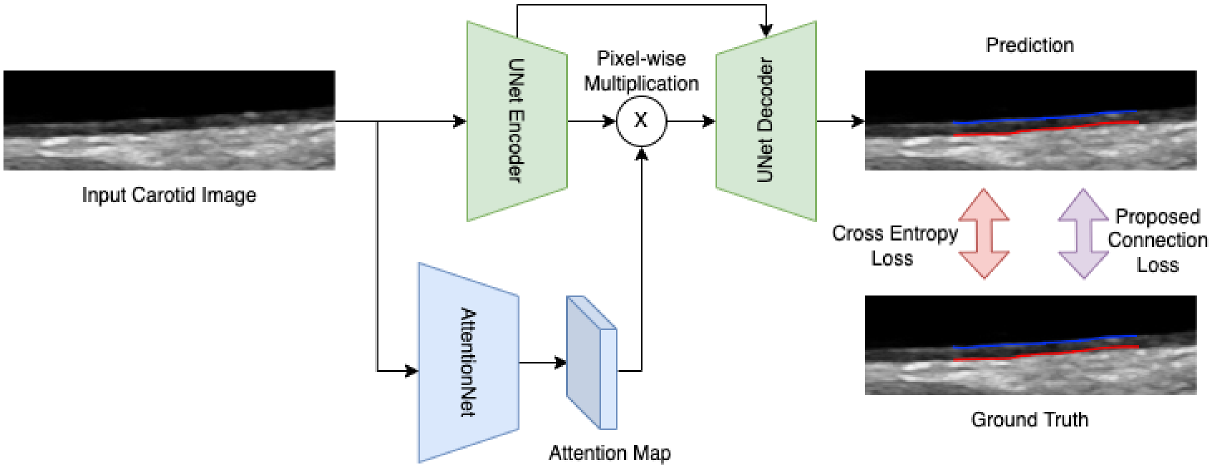


Figure 2. Overall architecture of the proposed system

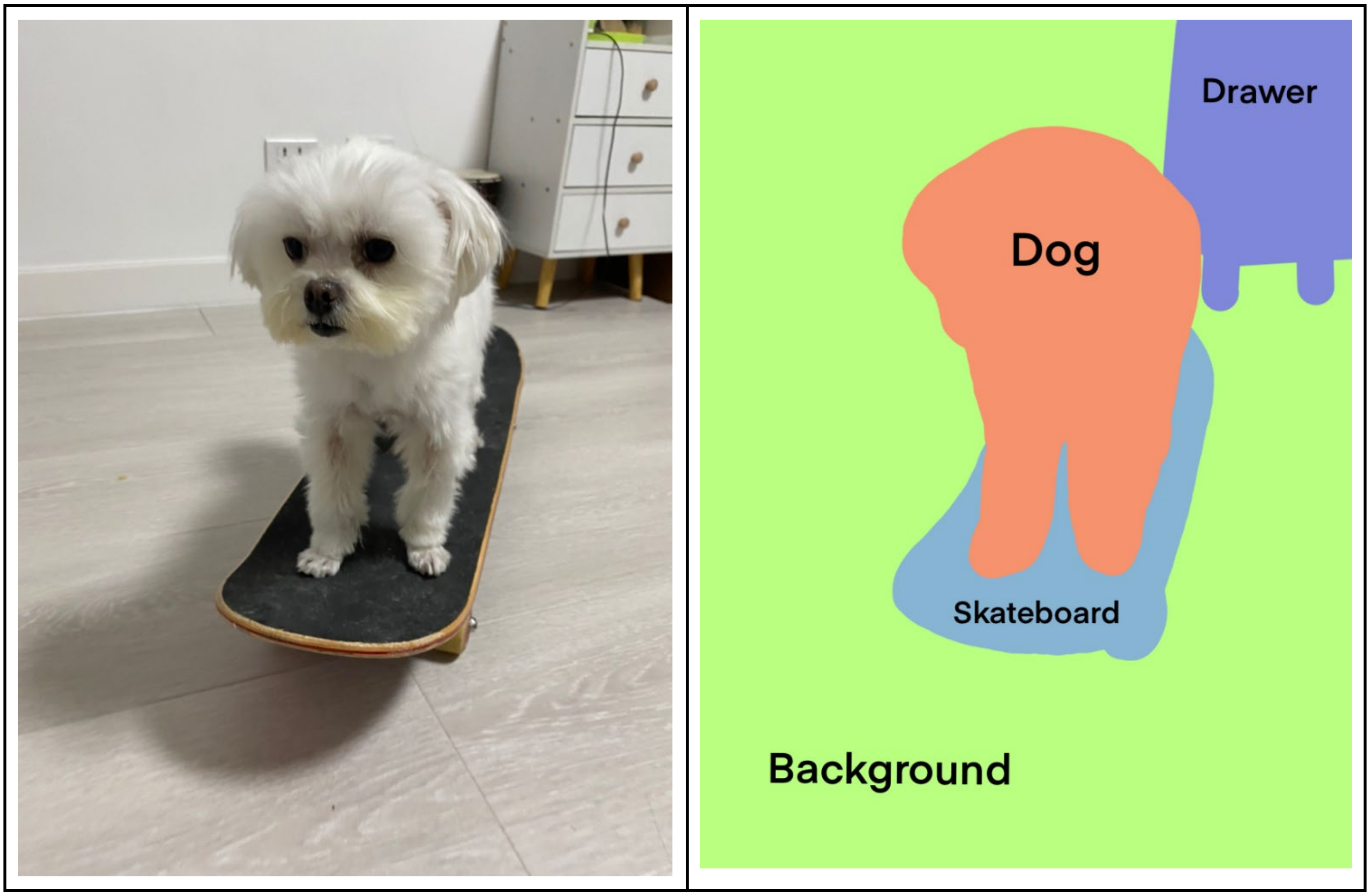
Figure 2 displays the overall architecture of the proposed system. The input image is processed through both the UNet Encoder and the proposed AttentionNet. Pixel-wise multiplication is applied on the two outputs of both modules and its result is then fed to the UNet Decoder to produce the final segmentation prediction. Two different types of loss function are used to train the proposed system.

UNet

In this paper, we exploit the UNet [4] to extract the hierarchical features from the input carotid image. The UNet consists of an encoder and a decoder, which are responsible for downsampling and upsampling the input image to extract important information including patterns and locations. It is often used in segmentation tasks due to its proficiency in determining the area of prominence [5, 6].

Given a carotid ultrasound input image, the encoder produces a hierarchical feature which contains the useful information of the input image for predicting accurate segmentation results. The same input image is also fed to the proposed AttentionNet, yielding an attention map which determines the region of interest that the UNet Decoder should focus on in order to yield better results. This process will be explained in detail in the AttentionNet section. The pixel-wise multiplication is applied between the feature map, results of the encoder, and the attention map and its product then fed to the Unet Decoder, which predicts the final segmentation results as shown in Figure 2.

In this paper, we consider this segmentation as a pixel-wise classification as shown in Figure 3. This final output consists of the percentage values on whether or not each pixel of the input image falls into the background or the designated classes such as Lumen-Intima or Media-Adventitia category.



(a)	(b)

Figure 3. Example of pixel-wise classification: (a) Input image and (b) Prediction map of pixel-wise classification

Table 1. Architecture of UNet used in this paper

UNET ARCHITECTURE TABLE	
ENCODER/DECODER	OUTPUT SHAPE [Channel, Height, Width]
INPUT	[3, 128, 128]
Encoder DOWN1	[64, 128, 128]
Encoder DOWN2	[128, 64, 64]
Encoder DOWN3	[256, 32, 32]
Encoder DOWN4	[512, 16, 16]
Decoder UP1	[512, 8, 8]
Decoder UP2	[256, 16, 16]
Decoder UP3	[64, 64, 64])
Decoder UP4	[32, 128, 128]
OUTCONV	[2, 128, 128]

Table 1. shows the UNet architecture used in the proposed system. Aforementioned, UNet is composed of two modules, an Encoder and a Decoder. The Encoder consists of max pooling and convolution layers, while the decoding consists of up sampling and convolution layers.

The Encoder extracts hierarchical features through downsizing input image, while the Decoder constructs the segmentation map which gives information about the patterns present in the input image and where the lumen intima interface and the media adventitia interface is located.

AttentionNet

As aforementioned the proposed attention module receives a carotid ultrasound image as its input and outputs an attention map. Afterwards, pixel-wise multiplication is performed with the attention map and the feature map which is output of the Encoder. This output is then processed through the UNet Decoder to produce the final segmentation map. This attention map provides information on the region of interest that the UNet decoder should focus on.

As shown in Figure 1, the region containing the Lumen-Intima and the Media-Adventitia interfaces make up a very small region of the carotid ultrasound image. Thus, the implementation of the attention module prevents the decoder from being distracted from insignificant background regions on the Carotid ultrasound image input.

To develop the proposed attention module, we exploit the ResNet18 [7] which is commonly used in various classification tasks due to its proficient performance and efficiency. The effectiveness of the proposed AttentionNet is studied in the ablation study section.

Data Augmentation

Data augmentation is a technique used in machine learning to create more trainable data by slightly changing an aspect of the existing data. This technique allows the model to see a greater variation of input samples which allow the trained model to perform better on unseen data during the test phase. Furthermore, it can combat problems regarding overfitting and other similar ones that often arise due to lack of data and small sample sizes. In this paper, we use random translation and rotation augmentation, which are commonly applied in segmentation studies [5, 6].

Sometimes due to the flaw of the carotid ultrasound devices, input carotid images can contain noise. This has a detrimental impact on the performance of the model as the trained model previously has not encountered said noise-containing input. To solve this problem, we implemented the gaussian noise augmentation in order to enforce the trained model to yield better performance in such cases.

Loss Function

A loss function quantitatively measures the efficiency of the given model by calculating the difference between the prediction of the model and the corresponding ground truth.
The training process aims to minimize the predefined loss function which yields the characteristic of the trained model.

In this paper, we use two different types of loss functions to train the proposed model. The total loss L is represented in Equation 1.

Equation 1. Total Loss Function

$$L = \alpha L_{ce} + \beta L_{connection}$$

Where L_{ce} denotes the cross-entropy loss function and the $L_{connection}$ denotes the proposed connection loss function. The weight of the two loss functions are denoted by and , for the cross-entropy loss function and the proposed connection loss function, respectively.

Cross-Entropy Loss

The cross-entropy loss function calculates the logarithmic loss, through comparing the ground truth and the prediction. Cross-entropy loss function is a very commonly used in classification tasks [8-10]. The equation for the cross entropy function is the following.

Equation 2. Cross-Entropy Loss Function

$$L_{ce} = -\sum_{i=1}^{C} y_i \times log\,\widehat{y_i}$$

Where C is the number of classes, y_i is the ground truth, and $\widehat{y_i}$ is the prediction. For example, if the ground truth is 1, which is a positive sample, but the prediction is 0, then loss should ideally be infinite. Another example is when the ground truth and the prediction are equal to each other, in which the loss should ideally be 0.

Connection Loss Function

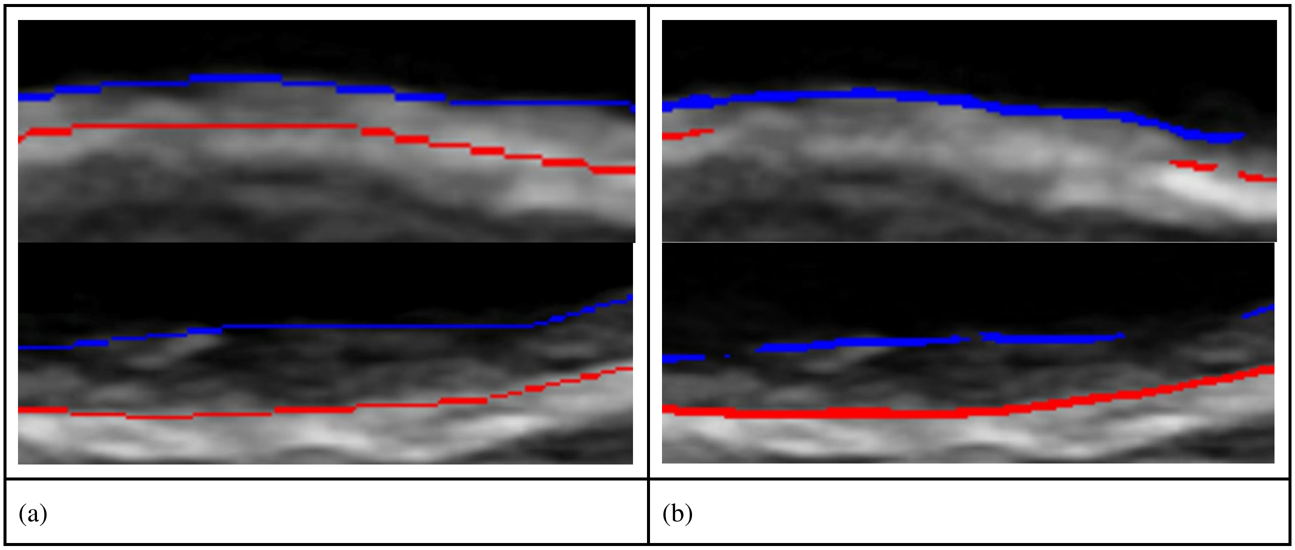
Even while using the cross-entropy loss function, there are many instances where the output presents a disconnected Lumen-Intima or Media-Adventitia Interface, which is not only anatomically impossible, but also has an adverse impact on the accuracy of the model. In this paper, we propose a connection loss function which penalizes the model when the Lumen-Intima or Media-Adventitia Interface is not connected, hence its name.

This thus solves the disconnected Lumen-Intima or Media-Adventitia problem as the proposed connection loss function enforces the model to output and display a continuous line. The equation for the proposed connection loss function is the following.

Equation 3. Connection Loss Function

$$L_{connection} = \left|\left|f(x)\right|\right|_1$$

Where, x is the input image and f() is the overall system.



(a)

(b)

Figure 4. Effectiveness of the proposed connection loss where (a) Result of the trained network with the proposed connection loss and (b) without connection loss.

Figure 4. shows the effectiveness of the proposed domain-specific connection loss. The proposed loss enforces the model to yield a continuous line instead of a segmented line. The connection loss enforces the model to output and display continuous lines that denote the lumen-intima interface and the media-adventitia interface. The detailed effectiveness of the proposed connection loss function is further examined in the ablation study section.

Results

Dataset

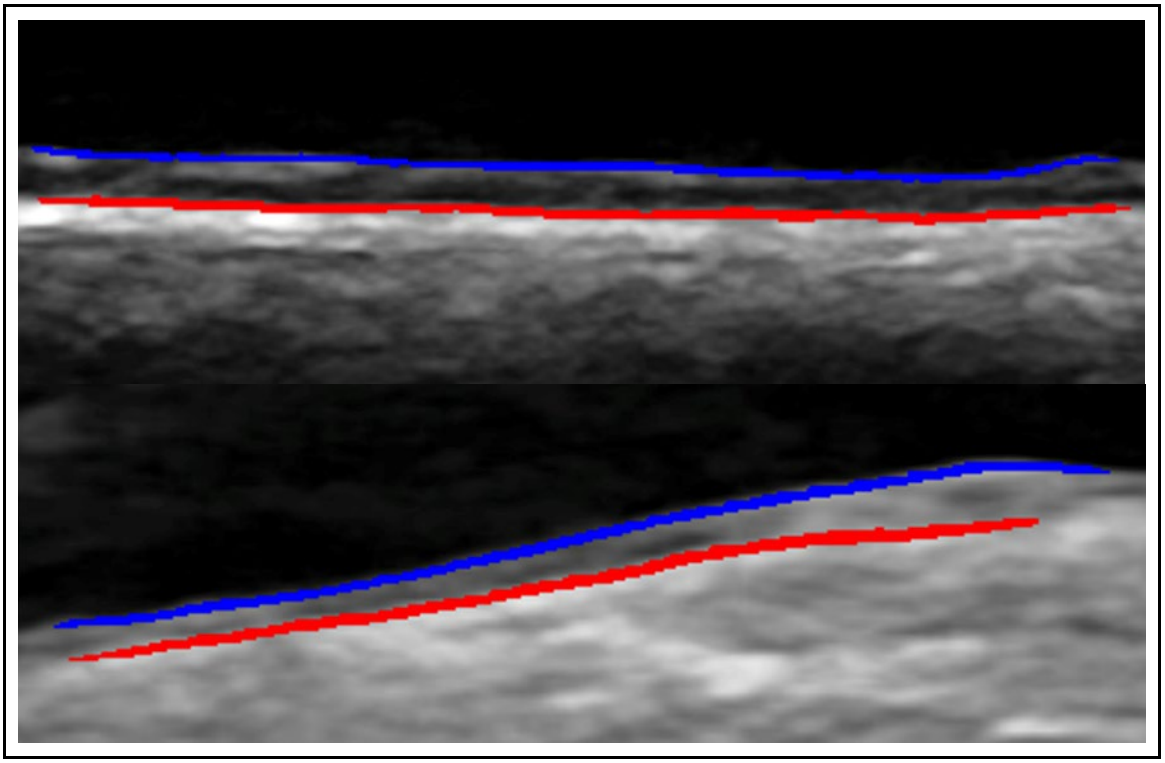


Figure 5. Snapshot of the samples used to train the proposed model. (The blue line denotes the Lumen-Intima interface and the red line denotes the Media-Adventitia interface.)

Figure 5 showcases the snapshot of the samples used in this paper. The blue line denotes the Interface of the lumen intima and the red line denotes the interface of the media adventitia interface respectively. The dataset consists of 4,129 samples. The Lumen intima interface and the media adventitia interface was annotated by experienced specialists. As different individuals have different carotid arteries of various shapes and sizes, the dataset is varied. This variety caused the segmentation task to be more challenging than initially expected. All 4,129 samples in the dataset are used. Among which, 90% is used to train the proposed model and the remaining 10% is used to test the quality and effectiveness of the model.

Protocol

In this paper, we utilize two quantitative evaluation metrics, IoU (Intersection over Union), and ACC (Accuracy). These two metrics are used to determine the general effectiveness of various state-of-the-art methods and compare them numerically.

Equation 4. IoU Calculation

$$IoU(X,Y) = \frac{|X \cap Y|}{|X \cup Y|}$$

Equation 4 shows how the IoU is calculated. Where X denotes the prediction while Y denotes the ground truth. The numerator is the true positive, while the denominator is the sum of the true positive, false positive, and false negative.

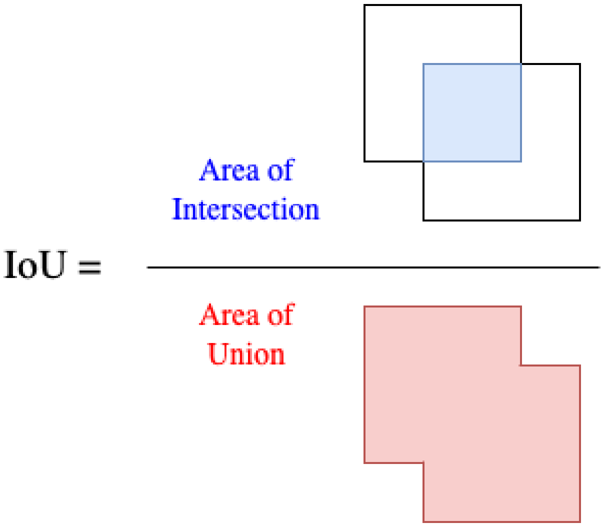


Figure 6. An illustration of IoU calculation. (IoU is an evaluation metric which can be calculated by dividing the area of intersection by the area of union.)

IoU is a metric value between 0 and 1 inclusive. As can be seen from Figure 6, the metric is used to evaluate the similarity between prediction and the ground truth, or the area of intersection and area of union respectively. Used most frequently in image segmentation tasks, IoU has proven itself to be a significant indicator of the quality and reliability of results [13, 14]. For instance, in which the prediction and the ground truth do not overlap in any sections, the IoU value is 0, and if it coincides completely, the IoU value is 1.

ACC is a metric value between 0 and 1 inclusive, and it is most commonly used to evaluate classification models. As we are considering Carotid Intima-Media Thickness segmentation as a pixel-wise classification task, it is important to also calculate the ACC value. ACC is calculated by dividing the total number of predictions by the number of correct predictions. In which a model has predicted all of its values correctly, the model has an accuracy value of 1. In binary classification, ACC is calculated in boolean terms, in which the metric is equivalent to the sum of true positives and true negatives over the sum of true positives, true negatives, false positives, and false negatives.

Comparison

Table 2. Quantitative result Comparison with the state-of-the-art methods.

	IoU	ACC
VGGNet [15]	0.63	0.88
Resnet [7]	0.69	0.94
Ours	0.78	0.99

Table 2 compares the IoU and ACC values of the proposed model to the existing state-of-the-art methods [7, 15] that have shown comparable performance in many computer vision problems. The proposed model has an IoU value of 0.78 and an ACC value of 0.99.

Compared to the VGGNet [15], the IoU value of the proposed model is greater by 0.15, and the ACC value is greater by 0.11. Thus, the proposed model demonstrates a 23.8% increase in performance in terms of IoU and 12.5% increase in performance in terms of ACC.

The IoU value of the proposed model is also greater than that of the Resnet [7] by 0.09, while the ACC value is greater by 0.05. Thus, the proposed model demonstrates a 13% and 5.3% increase in terms of IoU and ACC value, respectively.

We can attribute the superiority of the proposed model to the previous state-of-the-art methods to the inclusion of the attention module. Attention module allows the proposed model to focus on areas most likely to contain the carotid lumen intima and carotid media adventitia interfaces, which allows the aforementioned model to perform with greater skill. In addition, the proposed model also utilizes the proposed connection loss function, which contains a domain-specific penalty condition, further improving model performance. This is further elaborated in the ablation study section.

Ablation Study

An ablation study is conducted to verify the effectiveness of the proposed methods. We compare the full model with three control models that each individually lack a distinct component of the proposed model; attention module, pretrained weight, and the proposed connection loss function. Through this, we are able to showcase the overall impact each component had on the effectiveness of the model.

Table 3. Effectiveness of the proposed methods.

Model	IoU
W/o pretrained weight	0.75
w/o connection loss	0.73
w/o attention module	0.69
Full model	0.78

The first control model is training without the attention module, the second is without the pretrained weight, and the third is without the proposed connection loss function.

As can seen from table 3, compared to the proposed system, the IoU value of the first control model is less by 0.09, with the first control model yielding an IoU value of 0.69. We attribute the performance drop to the lack of attention module. Thus, it can be assumed that the implementation of the attention module contributes to increasing the performance levels in terms of IoU value by 13%.

The second control model yields an IoU value of 0.75 which is lower than that of the proposed model by 0.03. Hence, we can assume that employing a pretrained weight is correlated with an overall increase in performance quality by 4%, with regards to IoU value.

Finally, in regard to those of the proposed system, the third control model yields an IoU value that is lower by 0.05. This decrease in value reveals that the implementation of the proposed connection loss function contributed to the overall increase in performance and effectiveness of the model, by 6.8% in terms of IoU measurement.

The proposed connection loss function penalizes when the predicted interface is disconnected, which is anatomically impossible. The lumen intima interface and the media adventitia interface are always connected under any circumstances, so a noticeable gap connotes that the incorrect interface is predicted. As this is indubitable, enforcing a penalty otherwise can only improve the performance level of the trained model.

Conclusion

This paper is aimed to propose a deep learning based carotid intima-media thickness segmentation system by considering the segmentation problem as a pixel-wise classification. The proposed system consists of UNet and AttentionNet. The UNet is composed of an encoder and a decoder. First, the input carotid image is processed through the encoder and the proposed AttentionNet. The outputs of the AttentionNet and the feature map (output of the encoder) are multiplied pixel-wisely, and afterwards fed into the decoder where it is upsampled to produce the final segmentation prediction.

This model yielded accuracy in both the IoU and ACC metric at 0.78 and 0.99 respectively. These results are significantly higher than previous state-of-the-art methods, by approximately 18% and 8.8% on average.

An ablation study was adopted to examine how each component increased the performance of the trained model. It was confirmed that all the pre-trained model, the AttentionNet, and the proposed connect loss function contributed to the increased accuracy with the attention map having the most significant impact. This result aligns with the original hypothesis, as the attention map allows the model to focus on the region of interest of the input image, and prevents it from getting distracted by background components. The implementation of the proposed connection loss function allows the model to take basic human anatomical factors into consideration when training, including the fact that arteries of a living person can be not segmented.

We also propose the gaussian noise for data augmentation in order to make the trained model predict more precise results for seldom inputs that contain noise. In the future, we also plan on creating a lighter version of this model that can easily be applied to edge devices.

Declaration of Competing Interest

The author of this study declare that she has no conflict of competing interest

Funding Statement

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Acknowledgments

I would like to thank my advisor for the valuable insight provided to me on this topic.

References

[1] O'Leary, D. H., & Bots, M. L. (2010). Imaging of atherosclerosis: carotid intima-media thickness. *European heart journal*, *31*(14), 1682–1689. https://doi.org/10.1093/eurheartj/ehq185

[2] Shin, Jae, et al. "Automating carotid intima-media thickness video interpretation with convolutional neural networks." *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2016.

[3] Al-Mohannadi, A., Al-Maadeed, S., Elharrouss, O., & Sadasivuni, K. K. (2021). Encoder-Decoder Architecture for Ultrasound IMC Segmentation and cIMT Measurement. *Sensors (Basel, Switzerland)*, *21*(20), 6839. https://doi.org/10.3390/s21206839

[4] Ronneberger, Olaf, Philipp Fischer, and Thomas Brox. "U-net: Convolutional networks for biomedical image segmentation." *International Conference on Medical image computing and computer-assisted intervention*. Springer, Cham, 2015. https://doi.org/10.48550/arXiv.1505.04597

[5] Badrinarayanan, Vijay, Alex Kendall, and Roberto Cipolla. "Segnet: A deep convolutional encoder-decoder architecture for image segmentation." *IEEE transactions on pattern analysis and machine intelligence* 39.12 (2017): 2481-2495. https://doi.org/10.48550/arXiv.1511.00561

[6] Chen, Liang-Chieh, et al. "Encoder-decoder with atrous separable convolution for semantic image segmentation." *Proceedings of the European conference on computer vision (ECCV)*. 2018. https://doi.org/10.48550/arXiv.1802.02611

[7] He, Kaiming, et al. "Deep residual learning for image recognition." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016. https://doi.org/10.48550/arXiv.1512.03385

[8] Howard, Andrew G., et al. "Mobilenets: Efficient convolutional neural networks for mobile vision applications." *arXiv preprint arXiv:1704.04861* (2017). https://doi.org/10.48550/arXiv.1704.04861

[9] Huang, Gao, et al. "Densely connected convolutional networks." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017. https://doi.org/10.48550/arXiv.1608.06993

[10] Szegedy, Christian, et al. "Going deeper with convolutions." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015. https://doi.org/10.48550/arXiv.1409.4842

[11] Kingma, Diederik P., and Jimmy Ba. "Adam: A method for stochastic optimization." *arXiv preprint arXiv:1412.6980* (2014). https://doi.org/10.48550/arXiv.1412.6980

[12] J. Deng, W. Dong, R. Socher, L. -J. Li, Kai Li and Li Fei-Fei, "ImageNet: A large-scale hierarchical image database," 2009 IEEE Conference on Computer Vision and Pattern Recognition, 2009, pp. 248-255, doi: 10.1109/CVPR.2009.5206848.

[13] Jain, Jitesh, et al. "SeMask: Semantically Masked Transformers for Semantic Segmentation." *arXiv preprint arXiv:2112.12782* (2021). https://doi.org/10.48550/arXiv.2112.12782

[14] Cheng, Bowen, et al. "Masked-attention mask transformer for universal image segmentation." *arXiv preprint arXiv:2112.01527* (2021). https://doi.org/10.48550/arXiv.2112.01527

[15] Simonyan, Karen, and Andrew Zisserman. "Very deep convolutional networks for large-scale image recognition." *arXiv preprint arXiv:1409.1556* (2014). https://doi.org/10.48550/arXiv.1409.1556